УДК 004.056.5+681.518.5

СПОСОБ АПРИОРНОЙ ОЦЕНКИ ВОЗМОЖНОСТИ ИДЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЕЙ ВЕБ-РЕСУРСОВ НА ОСНОВЕ ЭНТРОПИЙНОГО ПОДХОДА

Захаров И.В., Забузов В.С., Фомин С.И., Эсаулов К.А.

ФГКВОУ ВПО «Военно-космическая академия имени А.Ф. Можайского» Министерства обороны Российской Федерации, Санкт-Петербург, Россия (197198, Санкт-Петербург, ул. Ждановская, 13)

Приведен анализ основных направлений выявления возможности идентификации пользователей веб-ресурсов и кратко обоснована необходимость ее оценивания. Предложен способ количественной оценки возможности идентификации пользователей веб-ресурсов, который базируется на учете явной или неявной передачи веб-ресурсу определенного набора признаков, позволяющих идентифицировать пользователя. Признаками являются параметры, характеризующие программно-аппаратную среду пользователя и его информационную деятельность в Интернете. Под информативностью признака понимается его двоичная энтропия. Показателем возможности идентификации пользователя, обладающего конкретным набором признаков (профилем), служит вероятность однозначности профиля. Получены выражения для оценки вероятности однозначной идентификации, исходя из суммарной информативности признаков пользователя. Представлен табличный способ учета взаимной корреляции признаков, а также вероятности их получения веб-ресурсом. Приведена оценка суммарной информативности признаков, необходимой для идентификации пользователя с заданной вероятностью. Намечены пути апробации и развития предложенного способа.

Ключевые слова: веб-ресурс, идентификация, информативность, признак пользователя, энтропия.

THE WAY OF APRIORISTIC ESTIMATION THE POSSIBILITY TO IDENTIFY USERS OF WEB-RESOURCES BASED ON ENTROPY APPROACH

Zakharov I.V., Zabuzov V.S., Fomin S.I., Esaulov K.A.

Military Space academy n.a. A.F.Mozhaisky, Saint-Petersburg, Russia (197198, Saint-Petersburg, street Zhdanovskaya, 13)

Analysis of basic directions of the observing possibilities to identify users of web-resources is presented and the necessity its estimation is proved briefly. Proposed way of quantitative evaluation of the possibility to identify web-users, which is based on considering the explicit or implicit transfer to web-resource a certain set of signs, allowing to detect the user's identity. Signs are the parameters, characterizing hardware and software user's environment and his information activities on the Internet. The informativeness of sign means its binary entropy. An indicator of the user's identification possibility, which has a specific set of characteristics (profile), is the probability of uniqueness profile. Expressions for an assessment of probability of unambiguous identification proceeding from total informational content of signs of the user. Presented tabular way to account for cross-correlation of signs and also probability of them getting by a web-resource. The estimation of total informativeness, necessary to identify the user with a given probability, has been obtained. Ways of testing and the development of the proposed method are planned.

Keywords: entropy, identification, informativeness, sign of the user, web-resource.

Вопросы, связанные с обеспечением безопасности работы в сети Интернет, с каждым днем встают все более остро. С одной стороны, необходимо обеспечить требуемую меру анонимности пользователя. С другой стороны, целесообразность идентификации пользователей назревает, например, в аспекте защиты информационных систем от различного рода злоумышленников. И в том и в другом случае возникает задача оценивания возможности идентификации пользователей веб-ресурсов.

Для решения данной задачи могут использоваться как качественные, так и количественные показатели. Качественные показатели оценивают, как правило, с помощью метода

экспертных оценок [2]. Они могут характеризовать, например, географическое положение, национальную и ведомственную принадлежность веб-ресурса, объем запрашиваемых регистрационных сведений, необходимость включения *cookie*, *Java*-скриптов и т.п. для работы с веб-ресурсом, перенаправления данных на иные веб-ресурсы, характер скрытого сбора данных о пользователе и т.д. Показатели, характеризующие программно-аппаратную среду пользователя, и показатели, характеризующие информационную деятельность пользователя в Интернете, должны отражать, прежде всего, степень влияния передаваемой пользователем веб-ресурсу информации на возможность его идентификации.

К основным направлениям оценивания показателей возможности идентификации пользователей веб-ресурсов относятся:

- выявление веб-ресурсов, способных идентифицировать своих посетителей, и оценивание достоверности идентификации;
- анализ объема информации о пользователе, доступной с его рабочего места, используемого для выхода в Интернет;
- анализ защищенности рабочего места пользователя;
- анализ технических требований веб-ресурса для работы с ним;
- анализ информативности запросов пользователя с точки зрения оценки возможности выявления его информационного интереса;
- анализ индексов популярности веб-ресурсов;
- анализ трафика, передаваемого веб-ресурсу;
- анализ объема и характера передачи данных веб-ресурсом на сайты разработчиков и сторонние, или «темные», сайты;
- накопление и анализ статистики о характере сбора данных веб-ресурсом и т.д.

Для количественной оценки возможности идентификации пользователей веб-ресурсов требуется ввести показатель, учитывающий возможность явной (поисковый запрос, регистрация и т.п.) или неявной (посредством технологий скрытого сбора данных) передачи веб-ресурсу определенного набора признаков, позволяющих идентифицировать пользователя. Следует отметить, что поскольку для сбора данных веб-ресурс использует достаточно широкий спектр технологий, перечень признаков весьма разнообразен. В качестве примера можно указать, что при помощи объекта *Screen* доступно получение разрешения экрана (браузера) и глубины цвета, с использованием средств *AJAX* на веб-ресурс могут передаваться параметры ввода информации в поля с целью сбора статистики о скорости ввода и типовых ошибках пользователя [5], а с помощью *Navigator.plugins* можно получить список всех установленных в браузере плагинов, кроме *IE*.

В целях решения поставленной задачи могут быть использованы различные подходы. К примеру, заслуживает внимания способ идентификации пользователя [1], в основе которого лежит мера соответствия полученных признаков действительному пользователю в условиях их возможной подмены. Однако в условиях наличия значительной массы пользователей, работающих с популярными веб-ресурсами, следует подойти, напротив, с точки зрения различаемости пользователей при возможной недостаточности набора признаков для однозначной идентификации. При этом наиболее целесообразным представляется использование вероятностных подходов.

При работе в Интернете пользователь и его рабочее место как человеко-машинная система характеризуются совокупностью m признаков, которые в том или ином случае могут оказаться доступными веб-ресурсу. Пусть N – число пользователей (субъектов), взаимодействующих с идентифицирующим их веб-ресурсом. Каждый признак x_i имеет a_i исходов с вероятностями p_{ij}, j=1,…a_i. Тогда M – число всех возможных различных наборов признаков, или профилей:

$$M = \prod_{i=1}^{m} a_i \ .$$

Вероятность того, что случайный субъект имеет профиль $Y=<y_1,\ldots y_m>$, в предположении, что реализации признаков независимы, составляет

$$P(Y) = \prod_{i=1}^{m} p_{iy_i} .$$

Тогда вероятность того, что субъект имеет уникальный профиль (то есть все иные субъекты имеют профили, не совпадающие с Y), определяется как

$$R(Y) = (1 - \prod_{i=1}^{m} p_{iy_i})^{N-1} .$$

Величина $R(Y)$, являясь вероятностью однозначности профиля, служит мерой возможности идентификации субъекта, обладающего конкретным набором признаков (профилем). Однако для веб-ресурса, идентифицирующего пользователя, интересна оценка R для априори неизвестного Y. Предположим вначале, что реализации признаков равновероятны: $p_{ij}=1/a_i$. Тогда

$$R = (1 - \prod_{i=1}^{m} a_i^{-1})^{N-1} .$$

Обозначая $h_i=\log_2 a_i$, получаем

$$R = (1 - 2^{-\sum_{i=1}^{m} h_i})^{N-1} = (1 - 2^{-H})^{N-1}, \quad H = \sum_{i=1}^{m} h_i \ . \qquad (1)$$

Для рассмотренного случая справедливо

$$h_i = -\sum_{j=1}^{a_i} p_{ij} \log_2 p_{ij} \ . \qquad (2)$$

Полученная величина h_i, которую будем называть информативностью признака, соответствует двоичной энтропии [3], что дает возможность распространить полученные для H и R выражения на общий случай не равновероятных реализаций признаков. Смысл информативности признака состоит в том, что ее можно рассматривать как количество информации о пользователе, причем знание каждого признака уменьшает исходную энтропию как меру неопределенности профиля пользователя.

Так, например, по результатам эксперимента, проведенного организацией *Electronic Frontier Foundation* в рамках проекта *panopticlick.eff.org* [6], были получены некоторые оценки энтропии компонентов отпечатка браузера (табл. 1).

Таблица 1. Оценки энтропии компонентов отпечатка браузера по результатам эксперимента *panopticlick.eff.org*

Компоненты отпечатка браузера	*Энтропия компонента (бит)*
Заголовок *User Agent*	10,0
Список установленных плагинов	15,4
Список установленных шрифтов	13,9
Установки видеоподсистемы	4,83
Поддержка *supercookies*	2,12
Заголовок *http accept*	6,09
Временная зона	3,04
Включенность *cookies*	0,353

Однако на практике неизбежна ситуация, когда по имеющимся признакам можно в определенной степени судить о возможных значениях других, то есть имеет место корреляция признаков, снижающая их суммарную информативность. Кроме того, зачастую заведомо неизвестно, какие из признаков будут доступны веб-ресурсу. Следовательно, требуется учет возможности взаимной зависимости (корреляции) признаков, а также вероятностного характера их добывания веб-ресурсом.

Суммарную информативность последовательного анализа m признаков $x_1,\ldots, x_m$ можно рассчитать из выражения [4]

$$H' = H'(x_m | x_{m-1}, x_{m-2}, \ldots, x_1) =$$

$$= -\sum_{i_1=1}^{a_1} \ldots \sum_{i_m=1}^{a_m} p\left((y_m = i_m) | (y_{m-1} = i_{m-1}) \wedge (y_{m-2} = i_{m-2}) \wedge \ldots \wedge (y_1 = i_1)\right) \cdot$$

$$\cdot \log_2 p\left((y_m = i_m) | (y_{m-1} = i_{m-1}) \wedge (y_{m-2} = i_{m-2}) \wedge \ldots \wedge (y_1 = i_1)\right).$$

Пусть задано распределение вероятностей j-го признака при условии k-й реализации i-го признака, которые обозначим $p_{jl}^{(i,k)}$, $l = \overline{1, a_j}$. Тогда в случае y_i=k информативность признака x_j вычисляется как

$$h_j^{(i,k)} = -\sum_{l=1}^{a_i} p_{jl}^{(i,k)} \log_2 p_{jl}^{(i,k)},$$

а с учетом всех реализаций признака X_i

$$\tilde{h}_j(i) = -\sum_{k=1}^{a_i} p_{ik} h_j^{(i,k)} = -\sum_{k=1}^{a_i} p_{ik} \sum_{l=1}^{a_j} p_{jl}^{(i,k)} \log_2 p_{jl}^{(i,k)}.$$

Соответствующие расчеты результирующей информативности признаков для большого их числа достаточно сложны, поэтому предлагается использовать следующий подход. Пусть γ_i – вероятность добывания веб-ресурсом признака x_i, i=1,…,m. Определим величину $\Delta h_{ji} = h_j - \tilde{h}_j(i)$, имеющую смысл снижения информативности j-го признака за счет его корреляции с i-м признаком, и составим таблицу (табл. 2).

Таблица 2. Схема таблицы попарных снижений информативности признаков

------	X_1	X_2	X_3	X_4	…	X_{m-1}	X_m	γ
X_1	------	Δh_{12}	Δh_{13}	Δh_{14}	…	$\Delta h_{1(m-1)}$	Δh_{1m}	γ_1
X_2	Δh_{21}	------	Δh_{23}	Δh_{24}	…	$\Delta h_{2(m-1)}$	Δh_{2m}	γ_2
…	…	…	…	…	…	…	…	…
X_{m-1}	$\Delta h_{(m-1)1}$	$\Delta h_{(m-1)2}$	$\Delta h_{(m-1)3}$	$\Delta h_{(m-1)4}$	…	------	$\Delta h_{(m-1)m}$	γ_{m-1}
X_m	Δh_{m1}	Δh_{m2}	Δh_{m3}	Δh_{m4}	…	$\Delta h_{m(m-1)}$	------	γ_m

Будем исходить из того, что признаки следует ранжировать и последовательно выбирать максимальное Δh каждого признака, причем соответствующие Δh выбираемые пары признаков не должны повторяться. Для этого на первом шаге выберем строку, содержащую максимальный элемент таблицы:

$$\alpha_1 = \max_{i,j=1,m} \{\Delta h_{ij}\}; \; \Delta H^{(0)} = 0.$$

Следующий максимальный элемент будем выбирать из строки, номер которой равен номеру столбца выбранного элемента:

$$\alpha_{t+1} : \Delta h_{\alpha_t \alpha_{t+1}} = \max\{\Delta h^{(t)}_{\alpha_t \alpha_{t+1}}\};$$

$$\Delta H^{(t+1)} = \Delta H^{(t)} + \gamma_{\alpha_t} \Delta h_{\alpha_t \alpha_{t+1}};$$

$$t = \overline{1, m-1}; \Delta\tilde{H} = \Delta H^{(m-1)},$$

где $\Delta h^{(t)}$ – элементы таблицы на t-м шаге с учетом вычеркивания выбранных строк и столбцов.

В результате после (m-1) шагов получим

$$\tilde{H} = H - \Delta\tilde{H},$$

где H рассчитывается согласно (1) и (2). В том случае, если признаки являются независимыми, расчет упрощается:

$$\tilde{H} = \sum_{i=1}^{m} \gamma_i h_i = -\sum_{j=1}^{a_i} \gamma_i p_{ij} \log_2 p_{ij}.$$

Таким образом, оценка уровня возможности идентификации с учетом взаимной зависимости признаков субъекта может быть получена в соответствии с (1):

$$\tilde{R} = (1 - 2^{-\tilde{H}})^{N-1},$$

а при N>>1

$$\tilde{R} \approx 1 - N \cdot 2^{-\tilde{H}}.$$

Для определения суммарной информативности признаков, необходимой для идентификации субъекта с заданной вероятностью Q при N>>1, следует воспользоваться выражением

$$H^* = -\log_2(1 - Q^{1/(N-1)}) \approx \log_2 \frac{N}{1-Q}.$$

Таким образом, предложенный способ позволяет на основе имеющихся статистических данных о распределении признаков пользователей оперативно оценивать степень возможности их идентификации веб-ресурсом. В настоящее время в локальной вычислительной сети учебной лаборатории развернут тестовый сервер, осуществляющий сбор и накопление статистики обращений в целях проработки технологий, связанных с рассматриваемыми вопросами, и анализа их эффективности. Полученные на данный момент результаты позволяют положительно оценивать перспективы использования предложенного способа анализа возможности идентификации пользователей веб-ресурсов на основе энтропийного подхода.

Список литературы

1. Бессонова Е.Е. Способ идентификации пользователя в сети Интернет // Научно-технический вестник информационных технологий, механики и оптики. – 2012. – Вып.3. – С. 133–137.
2. Войцеховский С.В., Хомоненко А.Д. Согласование экспертных оценок при нечетком выводе в системе обнаружения вторжений // Проблемы информационной безопасности. Компьютерные системы. – 2009. – № 4. – С. 42–50.
3. Волькенштейн М.В. Энтропия и информация. – М.: Наука, 1986. – 192 с.
4. Идентификация и техническая диагностика: учебник для вузов / А.К. Дмитриев, Р.М. Юсупов. – МО СССР, 1987. – 521 с.
5. Пауэрс Ш. Добавляем Ajax: пер. с англ. – СПб.: БХВ-Петербург, 2009. – 448 с.
6. Сайт проекта Panopticlick [Электронный ресурс]. – Режим доступа: https://panopticlick.eff.org (дата обращения: 15.10.13).

Рецензенты:

Хомоненко А.Д., д.т.н., профессор, профессор кафедры математического и программного обеспечения ФГКВОУ ВПО «Военно-космическая академия имени А.Ф. Можайского» Министерства обороны Российской Федерации, г. Санкт-Петербург.

Басыров А.Г., д.т.н., доцент, начальник кафедры информационно-вычислительных систем и сетей, ФГКВОУ ВПО «Военно-космическая академия имени А.Ф. Можайского» Министерства обороны Российской Федерации, г. Санкт-Петербург.