КОМП'ЮТЕРНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

УДК 004.77

Т. О. Говорущенко, Д. М. Медзатий, В. С. Семенюк
Хмельницький національний університет

МЕТОД ФОРМУВАННЯ ІНФОРМАЦІЙНОГО ПОРТРЕТА КОРИСТУВАЧА В МЕРЕЖІ ІНТЕРНЕТ

© Говорущенко Т. О., Медзатий Д. М., Семенюк В. С., 2018

Розроблено метод формування інформаційного портрета користувача в мережі інтернет, який дає можливість визначити характеристики, що формують інформаційний портрет користувача, орієнтований на вдосконалення таргетингу інформації або на формування іміджу веб-особистості у мережі інтернет.

Ключові слова: ідентифікація (відстежування, трекінг, фінгерпринтинг) користувача в мережі інтернет, інформаційний портрет користувача, веб-особистість, таргетинг інформації.

The research is devoted to the development of the method for the formation of the user's information portrait on the Internet. The developed method provides the determination of the characteristics, which form the user's information portrait, that is aimed at perfecting the targeting of information or at the formation of the image of the web personality on the Internet.

Key words: identification (tracking, fingerprinting) of the user on the Internet, user's information portrait, web personality, targeting of information.

Вступ

Ідентифікація (відстежування, веб-трекінг, фінгерпринтинг) користувача в мережі інтернет – це виконувані в доброчесних цілях розрахунок та встановлення унікального ідентифікатора для кожного браузера, який відвідує певний сайт. Така ідентифікація має низку переваг: можливість відрізняти звичайних користувачів від ботів, зберігати переваги користувачів та застосовувати їх під час наступного візиту, персоналізація послуг кінцевих користувачів, виявлення шахрайства в інтернеті та запуску цілеспрямованих атак, виявлення вторгнення у конфіденційність користувача, можливість формування контекстної реклами, таргетинг інформації (виокремлення цільової аудиторії, концентрування інформації саме на цільовій аудиторії, подання контекстозалежної інформації – за інтересами, за соціально-демографічними характеристиками, за часом, за психологічними якостями) [1, 2].

Одним з найпопулярніших способів ідентифікації користувачів є cookies. Крім цього, ідентифікувати користувачів можна за допомогою явних ідентифікаторів, характеристик комп'ютера (адреса, тип використовуваної операційної системи, час, які можна отримати з HTTP-заголовків відправлених запитів), поведінкового аналізу та звичок користувача (рух курсора, улюблені розділи та сайти) [1].

Ідентифікація користувача в мережі інтернет дає змогу сформувати його інформаційний портрет. Поняття "інформаційний портрет користувача" окреслює основну необхідну інформацію для верифікації даних певного користувача і є дотичним до поняття "веб-особистість", введеного з метою формалізації засобу персоніфікації контенту; це множина всіх даних користувача та результати його комунікативної діяльності [2]. Веб-особистість – це множина даних, які стосуються конкретної особи й можуть стосуватись будь-якої категорії або будь-якої комбінації категорій даних, доступних у мережі інтернет [2]. Отже, складовими інформаційного портрета користувача є веб-контент, який

створив користувач у мережі інтернет, а також його персональні дані [2]. Інформаційний портрет користувача може бути предметом дослідження багатьох наук – соціології, психології, політології, культурології, етнології, менеджменту, маркетингу, криміналістики тощо.

Відтак *актуальним завданням* є ідентифікація користувача в мережі інтернет та формування його інформаційного портрета з метою вдосконалення таргетингу інформації та формування іміджу веб-особистості.

Аналіз відомих методів та рішень показав, що вже розроблено: методи та засоби комп'ютерно-лінгвістичного аналізу достовірності соціально-демографічних характеристик учасників віртуальних спільнот [3]; методи реєстрації, верифікації та валідації персональних даних користувачів веб-спільнот [4]; моделі віртуальної спільноти та інформаційного середовища віртуальної спільноти [5]; етапи пошуку у WWW різноманітної інформації [6]. Але зазначені методи і засоби не розв'язують задачу формування інформаційного портрета користувача мережі інтернет. Тому *метою цього дослідження* є розроблення методу формування інформаційного портрета користувача в мережі інтернет, який має на меті вдосконалення таргетингу інформації та формування іміджу веб-особистості у мережі інтернет (це важливо, наприклад, для політиків, учасників виборчих кампаній).

Способи ідентифікації користувача в мережі інтернет

Концепція фінгерпринтингу ґрунтується на припущенні, що кожен електронний пристрій має унікальний набір фізичних та/або логічних функцій, які можуть використовуватись для його ідентифікації на основі інформації, переданої браузерами. Залежно від використовуваних методів, користувача можна відстежувати за допомогою функцій браузера (фінгерпринтинг браузера) або на основі системних налаштувань (крос-браузерний фінгепринтинг), який дає змогу ідентифікувати пристрій та користувача, навіть коли використовується декілька браузерів [7].

Найпростіший спосіб трекінгу користувача в мережі інтернет – це побудова ідентифікаторів об'єднанням набору параметрів, доступних у середовищі браузера, кожен з яких сам по собі не становить жодного інтересу, але разом вони утворюють унікальне для кожного комп'ютера значення: User Agent (версія браузера, версія операційної системи, деякі аддони), годинник (відхилення між реальним та системним часом), інформація про CPU та GPU, роздільна здатність екрана та розмір вікна браузера, список встановлених у системі шрифтів, список усіх встановлених плагінів, ActiveX-контролів, Browser Helper Object'ів та їх версій, інформація про встановлені розширення та інше програмне забезпечення [1, 7]. Автори [8] запропонували незалежний від браузера метод ідентифікації користувача – показали, що частини IP-адреси, певного набору шрифтів, часового поясу та роздільної здатності екрана достатньо, щоб однозначно ідентифікувати більшість користувачів п'яти найпопулярніших браузерів. Ці параметри користувацького агента доволі ефективні, що підтвердила перевірка такого методу на наборі даних з майже тисячі записів, зібраних через загальнодоступний тестовий веб-сайт [8].

Низка ознак для ідентифікації користувача міститься в архітектурі локальної мережі та налаштуванні мережевих протоколів. Такі ознаки характерні для всіх браузерів, встановлених на клієнтському комп'ютері, їх не можна приховати за допомогою налаштувань приватності або якихось утиліт. Це: зовнішня IP-адреса, номери портів для вихідних TCP/IP-з'єднань, локальна IP-адреса для користувачів за NAT'ом або HTTP-проксі, інформація про проксі-сервери (отримана з HTTP-заголовка), які використовує клієнт [1, 7].

Ще одним варіантом ідентифікації користувача є аналіз характеристик кінцевого користувача: вибрана мова, кодування за замовчуванням, часовий пояс; дані в кеші клієнта та історія його переглядів; рухи мишею, частота та тривалість натискання клавіш, дані з акселерометра; будь-які зміни стандартних шрифтів сайта та їх розмірів, масштаб, використання спеціальних можливостей перегляду; стан певних функцій браузера, таких як блокування сторонніх cookies, DNS prefetching, блокування спливання вікон, налаштування безпеки [1, 7].

У роботі [9] подано спосіб відстеження користувача на основі профілю його системи з метою збирання даних за очищених або відключених користувачем cookies.

У роботі [10] проаналізовано властивості браузерів, які відправляють на сервер, дозволяючи створювати унікальний “відбиток” цих браузерів. Браузери, які підтримують Flash або Java, в середньому містять принаймні 18,8 біта ідентифікаційної інформації [11].

Вибір характеристик, які формують інформаційний портрет користувача в мережі інтернет

Для розроблення методу формування інформаційного портрета користувача в мережі інтернет визначимо спочатку характеристики, які формують інформаційний портрет користувача в мережі інтернет, орієнтований на вдосконалення таргетингу інформації або на формування іміджу веб-особистості у мережі інтернет.

Подамо інформаційний портрет користувача (множину характеристик, що формують інформаційний портрет) у такому формалізованому вигляді:

$$UIP = IC \cap WP, \tag{1}$$

де IC – інформаційний контент (наповнення) мережі інтернет, WP – множина характеристик веб-особистості.

Інформаційний контент мережі інтернет можна подати у формалізованому вигляді:

$$IC = \{ic_1, ..., ic_n\}, \tag{2}$$

де ic_i – i-й елемент інформаційного контенту мережі інтернет ($i = \overline{1..n}$), $n \to \infty$ – неможливо оцінити кількість елементів інформаційного контенту мережі інтернет, оскільки статистика [12] показує, що наприкінці 2017 р. загальний обсяг даних інтернету становив 17,2 зетабайта ($17{,}2 \cdot 2^{70}$ байтів), але цей обсяг невпинно зростає і до 2020 р., за прогнозами, сягне 44 зетабайтів.

Своєю чергою, множину характеристик веб-особистості можна подати у формалізованому вигляді:

$$WP = \{wp_1, ..., wp_m\}, \tag{3}$$

де wp_j – j-та характеристика веб-особистості ($j = \overline{1..m}$). До характеристик веб-особистості належать всі дані, які стосуються конкретної особи та її діяльності в мережі інтернет, а саме: відкриті реєстраційні дані користувача – прізвище, ім’я та по батькові користувача (wp_1), адреса його електронної пошти (wp_2), номер телефону (wp_3); відкриті персональні дані, які надав користувач, – вік (wp_4), стать (wp_5), сімейний стан (wp_6), кількість дітей (wp_7), рівень освіти (wp_8), спеціальність (wp_9), посада та місце роботи (wp_{10}), релігія (wp_{11}), якими іноземними мовами володіє (wp_{12}), політичні погляди (wp_{13}), хобі (wp_{14}), інтереси (wp_{15}); тематика новин, які переглядає користувач (wp_{16}); тематика блогів, які він переглядає (wp_{17}); тематика блогів, у яких бере участь користувач (wp_{18}); тематика відкритих для всіх коментарів користувача у блогах (wp_{19}); тематика форумів, які переглядає користувач (wp_{20}); тематика форумів, в яких бере участь користувач (wp_{21}); тематика відкритих для всіх дописів користувача на форумах (wp_{22}); тематика публікацій, які переглядає користувач (wp_{23}); тематика відкритих для всіх публікацій користувача (wp_{24}); пошукові сервери, які використовує користувач (wp_{25}); тематика пошукових запитів користувача (wp_{26}); в яких соціальних мережах зареєстрований користувач (wp_{27}); тематика публікацій рядків соціальних мереж, які переглядає користувач (wp_{28}); тематика відкритих для всіх публікацій користувача у соціальних мережах (wp_{29}); інтернет-магазини, які відвідує користувач (wp_{30}); категорії товарів, які переглядає користувач в інтернет-магазинах (wp_{31}); категорії товарів, які вибирає (кладе у віртуальний “кошик”) користувач в інтернет-магазинах (wp_{32}); категорії товарів, які купує користувач в інтернеті (wp_{33}); тематика фотографій, які переглядає або скачує користувач (wp_{34}); тематика фотографій, які викладає користувач у вільний

доступ у мережі інтернет (wp_{35}); тематика відео, які переглядає або скачує користувач (wp_{36}); тематика відео, які викладає користувач у вільний доступ в мережі інтернет (wp_{37}); категорія (ї) музики, яку слухає або скачує користувач (wp_{38}); категорія (ї) музики, яку викладає користувач у вільний доступ у мережі інтернет (wp_{39}); який сервер електронної пошти використовує користувач (wp_{40}); які сайти з прогнозами погоди переглядає користувач (wp_{41}); яке програмне забезпечення скачує користувач (wp_{42}); які онлайн-інструменти використовує користувач (wp_{43}); які месенджери використовує користувач для віртуального спілкування (wp_{44}); на яких сайтах знайомств зареєстрований користувач (wp_{45}); інформація про користувача (за прізвищем, ім'ям та по батькові) з порталів новин (wp_{46}); інформація про користувача з соціальних мереж, яку опублікував не користувач (wp_{47}); інформація про користувача з Вікіпедії (wp_{48}); інформація про користувача за хештегами з його прізвищем (wp_{49}); інформація про користувача з публікацій в інтернеті (wp_{50}), тоді $m = 50$. Зрозуміло, що кожен елемент wp_j може бути також множиною, яка складається з декількох записів. Очевидно, що всі використовувані характеристики веб-особистості є неконфіденційними даними, тобто під час формування інформаційного портрета користувача не порушуються таємниця листування, банківська таємниця, таємниця фінансової інформації.

З урахуванням викладеного вище множина характеристик, що формують інформаційний портрет користувача, має вигляд:

$$UIP = WP = \{wp_1, ..., wp_m\}. \quad (4)$$

Аналіз характеристик інформаційного портрета користувача дав можливість сформувати множину характеристик, що формують інформаційний портрет користувача, який використовуватимемо для вдосконалення таргетингу інформації, має вигляд:

$$UIP_t = \{wp_1, wp_2, wp_3, wp_4, wp_5, wp_6, wp_7, wp_8, wp_9, wp_{10}, wp_{11}, wp_{12}, wp_{13}, wp_{14}, wp_{15}, wp_{16}, wp_{17}, wp_{18}, wp_{19}, wp_{20}, wp_{21}, wp_{22}, wp_{23}, wp_{24}, wp_{25}, wp_{26}, wp_{27}, wp_{28}, wp_{29}, wp_{30}, wp_{31}, wp_{32}, wp_{33}, wp_{34}, wp_{35}, wp_{36}, wp_{37}, wp_{38}, wp_{39}, wp_{40}, wp_{41}, wp_{42}, wp_{43}, wp_{44}, wp_{45}\} \quad (5)$$

Отже, інформаційний портрет користувача, який використовуватиметься з метою вдосконалення таргетингу інформації, формують 45 характеристик.

Аналіз характеристик інформаційного портрета користувача дав можливість сформувати множину характеристик, що формують інформаційний портрет користувача, який використаємо з метою формування іміджу веб-особистості у мережі інтернет, у вигляді:

$$UIP_i = \{wp_1, wp_4, wp_5, wp_6, wp_7, wp_8, wp_9, wp_{10}, wp_{11}, wp_{12}, wp_{13}, wp_{14}, wp_{15}, wp_{19}, wp_{22}, wp_{24}, wp_{29}, wp_{35}, wp_{37}, wp_{39}, wp_{46}, wp_{47}, wp_{48}, wp_{49}, wp_{50}\} \quad (6)$$

Отже, інформаційний портрет користувача, який використовуватиметься з метою формування іміджу веб-особистості у мережі інтернет, визначають 25 характеристик.

Метод формування інформаційного портрета користувача в мережі інтернет

Оскільки для характеристик, що формують інформаційний портрет користувача, який використаємо з метою формування іміджу веб-особистості у мережі інтернет, важливим є ступінь довіри до веб-ресурсу, з якого взято ту чи іншу характеристику, то *метод формування інформаційного портрета користувача в мережі інтернет* складатиметься з таких етапів:

1. Використовуючи описані вище способи ідентифікації користувача в мережі інтернет, зібрати характеристики, що формують інформаційний портрет користувача, який використовуватиметься з метою вдосконалення таргетингу інформації, та наповнити зібраними значеннями множину UIP_t (кожен елемент wp_j множини UIP_t може бути множиною, яка складається із декількох записів).

2. Використовуючи викладені вище способи ідентифікації користувача в мережі інтернет, зібрати характеристики, що формують інформаційний портрет користувача, який використаємо з метою

формування іміджу веб-особистості у мережі інтернет, та наповнити зібраними значеннями перший рядок матриці $UIP_{i\_web\_trust}$ – елементи $uipi_k$ $(k=\overline{1..25})$ матриці $UIP_{i\_web\_trust}$ (кожен елемент $uipi_k$ $(k=\overline{1..25})$ матриці $UIP_{i\_web\_trust}$ може бути множиною, яка складається з декількох записів):

$$UIP_{i\_web\_trust}=\begin{vmatrix} uipi_1...uipi_{25} \\ uipw_1...uipw_{25} \\ uipt_1...uipt_{25} \end{vmatrix} \quad (7)$$

3. Для кожної характеристики, що формує інформаційний портрет користувача, який використовуватиметься з метою формування іміджу веб-особистості у мережі інтернет (для кожного елемента $uipi_k$ $(k=\overline{1..25})$ матриці $UIP_{i\_web\_trust}$), з'ясувати, з якого веб-ресуру взято цю характеристику, та наповнити зібраними значеннями другий рядок матриці $UIP_{i\_web\_trust}$ – елементи $uipw_k$ $(k=\overline{1..25})$ матриці $UIP_{i\_web\_trust}$; оскільки кожен елемент $uipi_k$ $(k=\overline{1..25})$ матриці $UIP_{i\_web\_trust}$ може бути множиною, яка складається з декількох записів, то, відповідно, кожен елемент $uipw_k$ $(k=\overline{1..25})$ матриці $UIP_{i\_web\_trust}$ також може бути множиною, яка складається з декількох записів.

4. Проаналізувати (із залученням експертів) веб-ресурси, з яких взято характеристики, що формують інформаційний портрет користувача для формування іміджу веб-особистості у мережі інтернет, тобто проаналізувати елементи другого рядка матриці $UIP_{i\_web\_trust}$ (елементи $uipw_k$ $(k=\overline{1..25})$ матриці $UIP_{i\_web\_trust}$).

5. Встановити міри довіри (із залученням експертів) до веб-ресурсів, з яких взято характеристики, що формують інформаційний портрет користувача для формування іміджу веб-особистості у мережі інтернет, тобто встановити міри довіри до елементів $uipw_k$ $(k=\overline{1..25})$ матриці $UIP_{i\_web\_trust}$ (мірами довіри вважатимемо числа з діапазону [0;1], де 0 – це веб-ресурс, якому взагалі не можна довіряти, 1 – веб-ресурс, якому можна довіряти на 100 %); наповнити значеннями мір довіри третій рядок матриці $UIP_{i\_web\_trust}$ – елементи $uipt_k$ $(k=\overline{1..25})$ матриці $UIP_{i\_web\_trust}$; оскільки кожен елемент $uipw_k$ $(k=\overline{1..25})$ матриці $UIP_{i\_web\_trust}$ може бути множиною, яка складається з декількох записів, то, відповідно, кожен елемент $uipt_k$ $(k=\overline{1..25})$ матриці $UIP_{i\_web\_trust}$ також може бути множиною, яка складається з декількох записів.

6. Якщо значення елемента $0.5 \le uipt_k \le 1$ $(k=\overline{1..25})$, тобто міра довіри до веб-ресурсу, з якого взято k-ту характеристику, становить не менше ніж 50 %, то внести елемент матриці $uipi_k$ (k-ту характеристику інформаційного портрета користувача) у множину UIP_i як її k-й елемент.

Висновки

У статті проаналізовано відомі методи та засоби і з'ясовано, що вони не забезпечують формування інформаційного портрета користувача мережі інтернет, який має на меті вдосконалення таргетингу інформації та формування іміджу веб-особистості у мережі інтернет.

Дослідження способів ідентифікації користувача в мережі інтернет показало, що існує велика кількість різних способів для трекінгу користувача, причому деякі з них неможливо нейтралізувати без повної зміни принципів роботи комп'ютерних мереж, веб-додатків, браузерів. Деяким способам, звісно, можна протидіяти, але інші працюють непомітно для користувача і захиститись від них навряд чи вдасться. Проведене дослідження довело, що, навіть якщо користувач використовує приватний режим перегляду, працюючи з мережею інтернет, все одно всі його пересування можна відстежити.

Вибрано характеристики, які формують інформаційний портрет користувача в мережі інтернет, орієнтований на вдосконалення таргетингу інформації або на формування іміджу веб-особистості у мережі інтернет, що дало змогу розробити метод формування інформаційного портрета користувача в мережі інтернет.

Вперше розроблено метод формування інформаційного портрета користувача в мережі інтернет, що дає можливість визначити характеристики, які формують інформаційний портрет користувача, з метою вдосконалення таргетингу інформації або для формування іміджу веб-особистості у мережі інтернет. Важливим аспектом розробленого методу є те, що під час визначення характеристик, що формують інформаційний портрет користувача для створення іміджу веб-особистості у мережі інтернет, враховується ступінь довіри до веб-ресурсу, з якого взято ту чи іншу характеристику (до інформаційного портрета користувача можуть входити тільки іміджеві характеристики, взяті з веб-ресурсів з мірою довіри не менше ніж 50 %).

1. Жуков А. Фингерпринтинг браузера. Как отслеживают пользователей в сети // [Электронный ресурс] / А. Жуков. – Электронные данные. – Режим доступа: https://xakep.ru/2015/01/30/user-web-tracking-howto/ (дата обращения 13.02.2018) – Название с экрана. 2. Федушко С. Інформаційний слід веб-учасника як основа комп'ютерно-лінгвістичного аналізу веб-контенту / С. Федушко // Інформація, комунікація, суспільство: міжнародна наукова конференція, 19–21 травня 2016 р.: тези доповідей. – Львів, 2016. – С. 72–73. 3. Федушко С. С. Методи та засоби комп'ютерно-лінгвістичного аналізу достовірності соціально-демографічних характеристик учасників віртуальних спільнот: дис. ... канд. техн. наук: 10.02.21 / Федушко Соломія Степанівна. – Львів: Нац. ун-т "Львівська політехніка", 2015. – 205 с. 4. Пелещишин А. М. Методи верифікації персональних даних на основі гендерного аналізу мови користувачів Веб-спільнот / А. М. Пелещишин, С. С. Федушко // Східно-Європейський журнал передових технологій. – 2010. – № 3/4. – С. 37–39. 5. Пелещишин А. М. Аналіз існуючих типів віртуальних спільнот у мережі інтернет та побудова моделі віртуальної спільноти на основі веб-форуму / А. М. Пелещишин, Р. Б. Кравець, О. Ю. Сєров // Вісник Нац. ун-ту "Львівська політехніка". Серія: Інформаційні системи та мережі. – 2011. – № 699. – С. 212–221. 6. Канюк Н. В. Етапи пошуку в WWW інформації, призначеної для аналізу політичних явищ / Н. В. Канюк, А. М. Пелещишин // Східно-Європейський журнал передових технологій. – 2013. – № 4. – С. 57–60. 7. Flood E. Browser Fingerprinting / E. Flood, J. Karlsson. – Chalmers University of Technology, University of Gothenburg, 2012. – 99 p. 8. Boda K. User Tracking on the Web via Cross-Browser Fingerprinting / K. Boda, A. M. Foldes, G. G. Gulyas, S. Imre // 16th Nordic Conference on Secure IT-Systems, October 26–28, 2011: Proceedings. – Tallinn (Estonia), 2011. – P. 31–46. 9. Ali M. User Profiling Through Browser Finger Printing / M. Ali, Z. A. Shaikh, M. K. Khan, T. Tariq // International Conference on Recent Advances in Computer Systems, November 30 – December 01, 2015: Proceedings. – Hail (Saudi Arabia), 2015. – P. 135–140. 10. Kaur N. Browser Fingerprinting as User Tracking Technology / N. Kaur, S. Azam, K. Kannoorpatti, K. C. Yeo, B. Shanmugam // 11th International Conference on Intelligent Systems and Control, January 05–06, 2017: Proceedings. – Coimbatore (India), 2017. – P. 103–111. 11. Eckersley P. How Unique Is Your Web Browser? / P. Eckersley // 10th International Symposium on Privacy Enhancing Technologies, July 21–23, 2010: Proceedings. – Berlin (Germany), 2010. – P. 1–18. 12. Глущенко Н. Слишком большие данные: сколько информации хранится в Интернете? // [Электронный ресурс] / Н. Глущенко. – Электронные данные. – Режим доступа: https://ain.ua/special/skolko-vesit-internet/ (дата обращения 14.02.2018). – Название с экрана.