УДК 004.5:004.75

А.М. Пелещишин, Н.Б. Шаховська
Національний університет «Львівська політехніка», Україна

Використання апарату нечітких множин для опису аудиторії веб-сайту

Описано проблеми моделювання аудиторії сайту. Уведено модель користувача WWW та методи кількісного визначення аудиторії сайту.

Наукова новизна статті. У статті пропонується нова модель аудиторії сайта як нечітка множина, яка описує міру приналежності користувачів WWW до аудиторії сайта. Пропонуються нові підходи до визначення такої міри. На основі отриманих результатів сформульовано нові підходи до оцінки розмірів аудиторії сайта.

Постановка проблеми та її актуальність

Проблема моделювання аудиторії World Wide Web є задачею, яка зумовлена рядом факторів, серед яких найважливішими є:

1) високий користувацький попит на сервіси пошуку, класифікації та аналізу інформаційних ресурсів WWW;

2) потреба власників сайтів у точному відображенні тематики сайта в сервіси пошуку, класифікації та аналізу інформаційних ресурсів WWW.

Задача визначення моделювання аудиторії сайта неодноразово розглядалася як з теоретичної точки зору, так і зі спробами реального впровадження. Проте дослідження у даній сфері носять односторонній характер – це автоматизоване (частково чи повністю) визначення підмножини користувачів Інтернету, які відвідали сайт протягом контрольного проміжку часу.

Аналіз останніх досліджень

Традиційно структурування аудиторії сайта здійснюється шляхом виділення спеціальних груп користувачів, які і складають аудиторію сайта. Таке групування базується на певних групових ознаках.

Найпоширенішими груповими ознаками є ознаки одного з наступних класів:

1) технічні ознаки – ознаки, що характеризують обладнання користувача, програмне забезпечення, методи мережного підключення тощо;

2) демографічні ознаки;

3) поведінкові ознаки – ознаки, що характеризують різні аспекти поведінки користувача в глобальному інформаційному середовищі.

Виділення невирішених раніше частин загальної проблеми

Існуючі моделі (з чітким фіксуванням приналежності користувача WWW до аудиторії сайта) не в достатній мірі задовольняють наступним вимогам.

1. Модель аудиторії сайта повинна відображати різний ступінь приналежності користувача до аудиторії сайта.
2. Модель аудиторії сайта повинна враховувати методи отримання прибутку власниками сайта від факту приналежності користувача до аудиторії.
3. Модель аудиторії сайта повинна забезпечувати можливість порівняння реальної ємкості аудиторій різних сайтів.

Формулювання цілей статті

Головними цілями цієї статті є:

1) дослідження та формалізація де-факто існуючих методів опису аудиторії сайта;

2) побудова формальної моделі аудиторії сайта та моделі користувача WWW.

Модель користувача WWW

Користувачем WWW вважатимемо програмних агентів та інших людей, які вирішують певні задачі, здійснюючи доступ до сайтів через мережу Інтернет.

Формально користувач описується наступним відношенням

(*основні характеристики, цілі користувача, історія користувача*),

де *основні характеристики* – базова інформація, що ідентифікує користувача;

цілі користувача – набір цілей, досягнення яких прагне користувач при роботі в WWW;

історія користувача – історія змістовних дій користувача в WWW.

Цілі користувача в системі WWW поділятимемо на наступні класи:

– *інформаційні цілі* – отримання необхідної інформації зі системи WWW;
– *операційні цілі* – здійснення в системі WWW певних операцій (купівля чи продаж товару, відсилання листа електронної пошти і т.п.);
– *комунікативні цілі* – спілкування в системі WWW з її іншими користувачами.

Важливим підкласом інформаційних цілей користувача є *навігаційні (пошукові) цілі* – знаходження необхідного ресурсу в системі WWW.

Історія користувача (UH) в базовому варіанті може подаватися як історія навігації користувача по WWW (UNH) – історія доступу користувачем до інформаційних ресурсів WWW. У такому разі

$$UNH = \{Q_i\}_{i=1}^{N_{UNH}}, \tag{1}$$

$$Q_i = (t_Q, Host, URI, Ref, RM, QS, SA, Page), \tag{2}$$

де t_Q – момент запиту на отримання ресурсу; $Host$ – унікальна адреса клієнта; URI –

унікальний ідентифікатор ресурсу; Ref – вказівник на попередній запис в історії навігації; RM – метод запиту; QS – параметри запиту; SA – відповідь сервера; $Page$ – отримана інформація.

Дана модель історії користувача фактично є відображенням журналу доступу до ресурсів WWW. Основним недоліком цієї моделі є недостатній ступінь відображення семантики дій користувача в системі WWW та занадто високий ступінь деталізації моделі. У даній моделі поведінка користувача описується послідовністю залежних записів, таким чином при проведенні різних видів аналізу необхідно враховувати не лише записи в історії, а й взаємозв'язки між ними.

Іншим способом опису історії користувача WWW є використання історії транзакцій користувача у системі WWW.

$$UTH = \{Tr_i\}_{i=1}^{N_{UTN}}, \tag{3}$$

$$Tr = \left(t_Q, \bar{t}_Q, Id_{Tr}, UNH\left(Id_{Tr}\right)\right), \tag{4}$$

де t_Q – момент початку трансакції; $\bar{t}_Q$ – момент завершення трансакції; Id_{Tr} – унікальний ідентифікатор трансакції; $UNH\left(Id_{Tr}\right)$ – записи в історії навігації користувача, що входять до даної трансакції.

Транзакція – послідовність взаємозв'язаних запитів користувача до сервісів WWW та результатів їхнього опрацювання. Одна транзакція може складатися з багатьох запитів користувача у WWW. Проте один запит може одночасно входити до кількох транзакцій.

Основними відмінностями транзакційної моделі історії користувача від навігаційної є:

1) відображення завершеної взаємодії користувача та глобального середовища;
2) незалежність кожної транзакції;
3) наявність у транзакції двох часових міток (початку та кінця транзакції);
4) наявність у транзакції окремих запитів користувачів до різних сайтів, які логічно пов'язані.

Транзакційна модель має ряд принципових переваг над навігаційною моделлю користувача. Це зокрема:

1) можливість поглибленого аналізу поведінки користувача та його реальних потреб;
2) можливість точнішої оцінки досягнутих результатів взаємодії користувача та WWW.

Модель аудиторії сайта

Вимоги до моделі аудиторії сайта

Існуючі підходи до моделювання аудиторії сайта базуються на визначенні певної характеристики груп користувачів, які відвідують, чи можуть відвідати, сайт.

Така модель передбачає опис аудиторії сайта як підмножини користувачів WWW.

$$Aud = \{U_i\}_{i=1}^{N_U}, \tag{5}$$

де U_i – користувач WWW, який може бути віднесеним до аудиторії сайта.

Більш конструктивною, хоча і не такою точною, є модель, яка передбачає опис аудиторії сайта як сукупності підмножин користувачів WWW, які згруповані на основі певної ознаки.

$$\begin{aligned} Aud &= Aud_1 \cup ... Aud_{N_{Aud}} \\ Aud_i &= \{U_i\}_{i=1}^{N_U^{(i)}}, \end{aligned} \tag{6}$$

де Aud_i – підмножина користувачів WWW, яка може бути віднесена до аудиторії сайта за однією спільною ознакою.

Модель сайта на основі нечітких множин

Використання множин для опису аудиторії сайта та її структури в недостатній мірі задовольняє наведеним вище вимогам до моделі сайта. Можливим альтернативним підходом до моделювання сайта є опис сайта як нечіткої множини

$$Aud(Site) = \{(U_i, B(U_i, Site))\}_{i=1}^{N_U}, \tag{7}$$

де $Aud(Site)$ – аудиторія сайту; U_i – користувач WWW; $B(U_i, Site)$ – міра приналежності користувача U_i до аудиторії сайта *Site*.

Таким чином, кожному користувачу WWW ставиться у відповідність величина, яка характеризує міру приналежності користувача до аудиторії сайта. Природними обмеженнями на величину приналежності є:

$$0 \le B(U_i, Site) \le 1. \tag{8}$$

Для користувачів, які взагалі не мають жодного відношення до сайта, приналежність рівна нулю.

Визначення реального змісту функції приналежності – це задача, яка носить суб'єктивний для власників сайта характер. Проте доцільним є побудова даної функції у відповідності до цілей, що ставляться власниками щодо свого сайта.

У такому разі реальний зміст функції *B* визначається природою міри цінності користувача для сайта. У структурі функції відображається основний механізм отримання прибутку від відвідування сайта користувачами WWW.

Так, для тих мір цінності, які пропорційно залежать від числа відвідувачів сайта (наприклад, прибуток від пропаганди ідей, розміщених на сайті), функція приналежності може визначатися як імовірність відвідування користувачем WWW сайта протягом контрольного періоду:

$$B(U_i, Site) = \Pr(T, U_i, Site).$$

У деяких випадках може бути важливим не тільки факт відвідування користувачем сайта, а й частота відвідування. У такому разі можливе визначення функції приналежності, наприклад, наступним чином:

$$B(U_i, Site) = \sum_{C=1}^{N_C} C * \Pr(T, U_i, C, Site), \quad (9)$$

де C – число заходів користувача на сайт протягом контрольного періоду T; N_C – максимальне цінне число заходів користувача протягом контрольного періоду T

Таким чином, використання апарату нечітких множин дозволяє будувати модель аудиторії сайта, яка задовольняє наведеним нижче вимогам.

1. Вірогідність попадання користувача WWW в аудиторію сайта описується за допомогою функції приналежності $B(U_i, Site)$.
2. Суб'єктивні інтереси власника сайта враховуються у структурі функції приналежності $B(U_i, Site)$.
3. Порівняння аудиторій сайтів стає можливим завдяки наявності числових оцінок, поданих як функція приналежності $B(U_i, Site)$. Детальніше проблему порівняння обсягів аудиторії сайтів розглянемо пізніше.

На практиці окреме визначення міри приналежності для кожного користувача WWW є неможливим у силу великої розмірності множини та відсутності аналітичного подання функції приналежності.

Проте достатньо ефективним методом опису аудиторії сайта як нечіткої множини є використання додаткових характеристик чи ознак користувача як аргументів функції приналежності.

У такому разі для кожного користувача U_i виділяються набір характеристик $\left(Ch_1^{(i)} \ldots Ch_{N_{Ch}}^{(i)}\right)$, які служать базовими для визначення міри приналежності до аудиторії сайта.

Тоді

$$B(U_i, Site) = B\left(Ch_1^{(i)} \ldots Ch_{N_{Ch}}^{(i)}, Site\right) \quad (10)$$

Такий підхід дозволяє побудувати достатньо компактний та ефективний опис аудиторії сайту як нечіткої множини на основі базових характеристик користувача WWW.

У якості базових характеристик можуть використовуватися наведені вище ознаки користувачів – технічні, демографічні, поведінкові.

Функція приналежності може задаватися як в аналітичному, так і в табличному виді (при невеликій скінченній множині можливих значень базових ознак).

Простим прикладом опису аудиторії сайта як нечіткої множини може бути нечітка множина, де функція приналежності залежить від віку користувача:

$$B(U_i, Site) = \Pr(T, Age(U_i), Site), \quad (11)$$

де $Age(U_i)$ – вік користувача.

Складнішим прикладом опису аудиторії сайта можуть бути визначення функції приналежності до аудиторії сайта як функції від географічних координат користувача WWW:

$$B(U_i, Site) = \Pr(T, Geo(U_i), Site), \quad (12)$$

де $Geo(U_i)$ – координати користувача WWW.

Окремим важливим випадком опису аудиторії сайта є визначення міри приналежності користувача до аудиторії сайта як функції від шляху від історії транзакцій користувача WWW:

$$B(U_i, Site) = \Pr(T, UTH(U_i), Site). \quad (13)$$

На практиці замість наведеного часто використовується дуже спрощений варіант функції:

$$B(U_i, Site) = \Pr(T, Ref(U_i), Site),$$

де Ref – WWW-ресурс, з якого користувач потрапив на сайт.

Підхід до моделювання аудиторії сайта, де міра приналежності визначається транзакційною історією користувача (або її навігаційним замінником), є одним з найважливіших при розв'язанні практичних задач оптимізації сайта та просування сайта в Інтернет. Тому він вимагає окремого детального розгляду.

Групування користувачів сайта

На практиці рідко є можливість побудувати функцію приналежності користувача WWW до аудиторії сайта в аналітичному вигляді. У такому разі опис функції здійснюється в табличному вигляді. Для зменшення розмірності такого опису користувачів WWW доцільно групувати.

Групою користувачів сайта (*j*-тою групою користувачів) вважатимемо множину користувачів WWW, об'єднаних однаковим значенням ознаки *Ch*:

$$Aud_j = \{U_i\}, \forall U_i : Ch(U_i) = Ch_j. \quad (14)$$

З точки зору структурування аудиторії сайта групування має сенс, якщо всі користувачі, що віднесені до однієї групи за спільною ознакою, мають однакову міру приналежності до аудиторії сайта. *Тобто у якості групової ознаки повинні вибиратися такі характеристики користувача, які визначають міру його приналежності до аудиторії сайта.*

Вважатимемо, що для всіх користувачів певної групи міра приналежності до аудиторії сайта є константою.

$$B(U_i, Site) = B_j(Site) \quad \forall U_i \in Aud_j, \; j = 0 \ldots N_{Aud} \quad (15)$$

Аудиторія сайта є об'єднанням усіх груп користувачів:

$$Aud(Site) = \bigcup_{j=0}^{N_{Aud}} Aud_j(Site). \quad (16)$$

У випадку, якщо для групування по окремих групах використовується єдина групова ознака *Ch*, то користувач може належати лише до однієї групи.

Відмітимо, що для кожного сайта має місце спеціальна група Aud_0 користувачів WWW, для яких міра приналежності до аудиторії сайта дорівнює нулю:

$$B_0(Site) < \varepsilon\,, \tag{17}$$

де ε – мала контрольна величина.

До даної групи відносяться користувачі WWW, які практично ніколи не скористаються послугами сайта. Можна говорити, що імовірність потрапити на сайт для них дорівнює нулю у сенсі геометричної імовірності.

При практичному описі аудиторії сайта група Aud_0 з розгляду, як правило, усувається. Крім того, слід відмітити, що часто під аудиторією сайта розуміють множину користувачів WWW, які не потрапляють до групи Aud_0.

Дана група для більшості сайтів (за винятком кількох найпопулярніших сайтів WWW) значно переважає по кількості членів інші групи.

Кількісне порівняння та оцінка обсягів аудиторій сайтів

Для сайтів близької чи суміжної тематики часто має місце задача порівняння обсягів аудиторії. Зокрема, дана задача має місце у плануванні рекламних заходів в Інтернет, оцінюванні ефективності сайтів, визначенні актуальності та якості матеріалів, що розміщені на сайтах.

На сьогодні традиційним і найпоширенішим методом порівняння обсягів аудиторії конкуруючих сайтів є порівняння простих технічних показників – кількості відвідувачів протягом певного періоду часу («хостів») та кількості поданих ними запитів до сайта («хітів»). Даний підхід є основою різноманітних Інтернет-рейтингів сайтів. Відповідно, на даному підході базуються основні підходи до оцінки ефективності функціонування сайта, доцільності розміщення на ньому реклами тощо.

Проте цей підхід не дає достатньо якісної картини щодо реального співвідношення аудиторій різних сайтів та активно критикується власниками сайтів та дослідниками WWW.

Основними недоліками такого підходу є ігнорування структури сайтів та ігнорування різноманітних форм взаємодії сайтів із користувачами. Об'єднуючим фактором у цих недоліках є те, що не враховуються реальні результати взаємодії користувачів із сайтами.

Точнішою оцінкою обсягів аудиторії може служити величина, базована на мірі приналежності користувача до аудиторії:

$$\|Aud(Site)\| = \sum_{i=1}^{N_U} B(U_i, Site), \tag{18}$$

де $\|Aud(Site)\|$ – обсяг аудиторії сайта.

У випадку, коли аудиторія сайта структурована по групах:

$$\|Aud(Site)\| = \sum_{j=1}^{N_{Aud}} B_j(Site)\|Aud_j(Site)\|. \tag{19}$$

Висновок

Моделювання аудиторії веб-сайту є ключовим етапом для багатьох важливих задач проектування та управління сайта. Серед них – проектування структури та наповнення сайта, визначення та оптимізація тематики сайта, організація та проведення рекламних кампаній в Інтернеті.

Існуючі на сьогодні моделі аудиторії веб-сайтів передбачають жорсткий опис приналежності користувача WWW до множини користувачів сайта. Аудиторія сайта визначається як підмножина користувачів World Wide Web. Проте такі моделі не в достатній мірі:

1) відображають різний ступінь приналежності користувача до аудиторії сайта;
2) враховують методи отримання прибутку власниками сайта від факту приналежності користувача до аудиторії;
3) забезпечують можливість порівняння реальної ємкості аудиторій різних сайтів.

Практична цінність. Наукові результати, отримані в даній статті, дозволяють провадити подальші практичні дослідження з моделювання та оптимізації веб-сайтів та побудови формальних методів проектування та організації успішної діяльності сайтів у глобальному середовищі. Зокрема, можливе точніше позиціонування тематики сайта щодо потреб користувачів та проведення робіт по покращенню позиціонування сайта в пошукових системах.

Література

1. Dai H., Luo T., Nakagawa M., Sun Y., and Wiltshire J. Discovery of aggregate usage profiles for Web personalizationMobasher // Proc. of the WebKDD 2000 Workshop at the ACM SIGKKD 2000. – Boston, 2000.
2. Flake G., Lawrence S., Giles C. Efficient identification of web communities // Proc. of the Sixth International Conf. on Knowledge Discovery and Data Mining (ACM SIGKDD-2000). – Boston, MA: ACM Press, 2000.
3. Flake G., Lawrence S., Giles C., Coetzee F. Self-Organization of the Web and Identification of Communities // IEEE Computer. – 2002. – № 35(3). – С. 66-71 // http://webselforganization.com
4. Srivastava J., Cooley R., Deshpande M., Tan P-T. Web usage mining: discovery and applications of usage patterns from Web data // SIGKDD Explorations. – 2000. – № (1) 2.
5. Пелещишин А.М. Методи та алгоритми моделювання Web-систем // Вісник ДУ «Львівська Політехніка». – 2000. – № 406. – С. 199-211.

А.М. Пелещишин, Н.Б. Шаховская

Использование аппарата нечетких множеств для описания аудитории веб-сайта

Описаны проблемы моделирования аудитории сайта. Дана модель пользователя WWW и методы количественного оценивания аудитории сайта.

A.M. Peleschin, N.B. Shaxovska

Using the Apparatus of Fuzzy Set for Discribtion of Web Site Auditory

The main problems of site audience modeling are described. The WWW-user model and the methods quantitative assessment are described.

Статья поступила в редакцию 01.07.2005.