2000. – 47 p. 16. Kosala R., Blockeel H. Web Mining Research: A Survey. Department of Computer Science, Katholieke Universiteit Leuven SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM, 2000. – 15 p. http://w. 17. Lacroix Z., Shahuguet A., Chandrasekar R. Information Extraction & Database Techniques: a user-oriented approach to querying the Web. IRCS & CIS, University of Pennsylvania. – 1998. – 16 p. 18. Mendelzon A., Mihaila G., Milo T. Querying the World Wide Web. Department of Computer Science and CSRI, University of Toronto, Toronto, Canadca, 1996. – 14 p. 19. Pankowski T. XML-SQL: An XML Query Language Based on SQL and Path Tables. Chair of Control, Robotics and Computer Science, Poznan University of Technology. – 2002. – 15 p. 20. Patamarjarnkul A. A Customized Web Search Engine Using a Tiny WebSQL Query Language. Project Report. Auburn University, Alabama May, 2000. – 73 p. 21. Sarawagi S. Nagaralu S. Data mining models as services on the internet. Indian Institute of Technology. Bombay 2000. – 24 p. 22. Sengupta A., Dalkilic M. DSQL – An SQL for Structured Documents. Extended Abstract, Department of A&IS, Kelley School of Business, Indiana University, 2002. – 4 p. 23. Spertus E., Stein L. Squeal: A Structured Query Language for the Web. MIT Artificial Intelligence Lab. Ninth International World-Wide Web Conference, May 2000. – 12p. 24. Stein L., Spertus E., College M. Squeal: A Structured Query Language for the Web. MIT Artificial Intelligence Lab, 2000. – 12 p. 25. Wang Y. Hu J. A Machine Learning Based Approach for Table Detection on The Web. Dept. of Electrical Engineering, Univ. of Washington 2001. – 13 p. 26. Wang Y. Hu J. Detecting Tables in HTML Documents. Dept. of Electrical Engineering, Univ. of Washington. – 2001. 12 p. 27. Witvoet O., Rauber A., Aschenbrenner A., Bruckner R., Putting the World Wide Web into a Data Warehouse:A DWH-based Approach toWeb Analysis. Department of Software Technology and Interactive Systems, Vienna University of Technology,Vienna, Austria. – 2002.

УДК 683.1

А.М. Пелещишин, Т.Б. Гулка
Національний університет “Львівська політехніка”,
кафедра “Інформаційні системи та мережі”

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ПОВЕДІНКИ ВІДВІДУВАЧІВ ВЕБ-САЙТІВ (ФОРМАЛЬНА МОДЕЛЬ)

Web-site visitors behaviour intelligent analysis main concepts are formally defined. Examples of some tasks formalizations are described based on given definitions.

Зроблено формальне визначення основних понять задачі інтелектуального аналізу поведінки відвідувачів веб-сайтів, наведено приклади формалізації ряду задач на основі поданих визначень.

ПОСТАНОВКА ПРОБЛЕМИ

Однією із основних задач інтелектуального аналізу поведінки користувача сайту (зокрема на основі журналів доступу до сайтів) є задача визначення невідомих характе-

ристик користувачів, сесій та запитів на основі відомих, що може використовуватися у прийнятті рішень щодо організації просування сайту в Інтернет, рекламних акцій, спрямування сайту на цільову аудиторію та інші маркетингові та рекламні задачі. Іншою задачею аналізу поведінки відвідувача сайту є визначення слабких місць сайту та можливих шляхів його вдосконалення. Для формального розв'язання цієї задачі необхідно попередньо формалізувати основні поняття.

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ ТА ПУБЛІКАЦІЙ

Комплексні завдання, такі як Web-дизайн, Web-сервер дизайн, побудова шляхів навігації, проведення рекламних компаній в Інтернеті потребують науково обґрунтованого аналізу. Мета таких досліджень очевидна: знаючи вподобання та характеристики відвідувача, можна починати цілеспрямовану рекламну кампанію, модернізувати чи модифікувати Web-вузол, оптимізувати фізичну структуру Web-системи. Для такого аналізу користуються декількома взаємодоповнюючими методами. Один з них застосовується компаніями, які при наданні своїх послуг в Інтернеті вимагають заповнити анкети відвідувача. В анкетах потрібно вказати свої дані, такі як вік, стать, професія, хобі, місце проживання тощо. Недоліки такого методу очевидні:

1. Більшість сайтів не пропонують ніяких послуг в режимі он-лайні і тому не можуть вимагати заповнення анкети;
2. Перевірити, чи правдиву інформацію про себе надав відвідувач, неможливо.

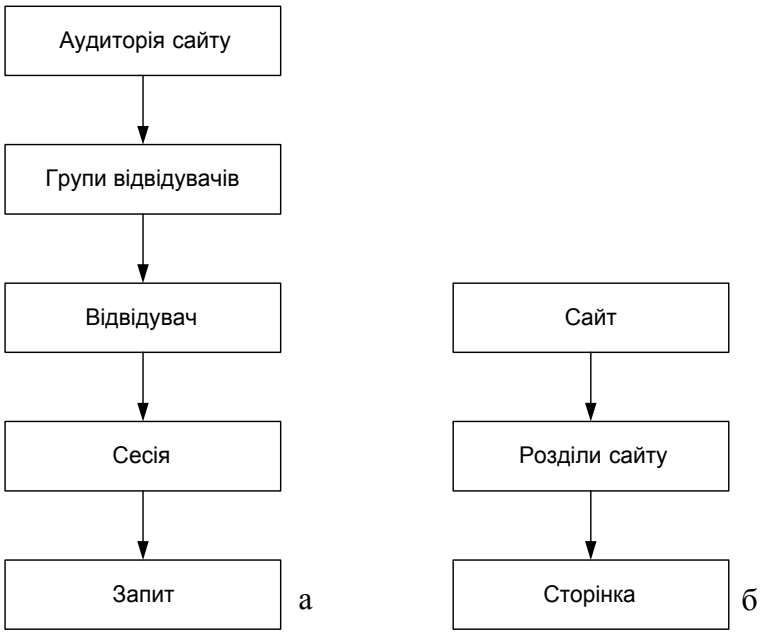
Проте є і переваги: якщо з певною ймовірністю вважати певний відсоток інформації достовірним, то можна отримати деякі дані, які іншим способом отримати неможливо (наприклад, вік, стать, професія).

Інший метод базується на аналізі журналів доступу до сайту. Такі журнали ведуть всі популярні Web-сервери. Крім того, більшість адміністраторів сайтів ведуть свої журнали. Журнали формуються шляхом читання значень змінних середовища CGI за допомогою CGI-скриптів. При запуску на сервері CGI-скрипта сервер формує середовище оточення, в якому скрипт може знайти всю доступну інформацію про HTTP з'єднання і запит. Тобто можна отримати дані про вузол, на який здійснювався запит, URI запиту, рядок запиту, метод передачі даних, IP-адресу користувача, технічні характеристики комп'ютера користувача (операційна система, броузер), з якого ресурсу було здійснено запит, час здійснення запиту тощо. Популярний Web-сервер Apache на додаток до даних, які можна отримати CGI-скриптами, дозволяє відслідкувати обсяг переданих даних, код помилки у разі її появи.

Сьогодні існує ряд комерційних програмних продуктів, які згідно із їх специфікацією аналізують статистику відвідування сайту. Проте ці системи мають принципові недоліки, що значно обмежують їхнє використання. Це, зокрема: обмежені типи аналізу статистичної інформації; закритість систем і, відповідно, неможливість розширення аналітиком напрямків аналізу, формування нерегламентованих запитів, вибірки та агрегування даних; безпосередня робота з журналами доступу; невдалі технічні рішення (зокрема, створення інтефейсу системи через динамічні HTML-сторінки).

Згідно з "клієнт-серверною" моделлю Веб-сайту ([1]), при аналізі його функціонування виділяються активна керуюча компонента, яка формує запити, та компонента, яка обслуговує запити, що надходять до сайту.

Обидві компоненти при аналізі функціонування Веб-вузла доцільно розглядати з різним ступенем деталізації. Рівні деталізації активної та серверної компонент наведено на рисунку.



Рівні деталізації активної (а) та серверної (б) компоненти веб-сайту

Весь комплекс задач аналізу поведінки відвідувача сайту отримується з наведеної схеми деталізації активної та серверної компоненти. У таких задачах використовується певний рівень деталізації серверної компоненти (наприклад, розділ сайту), активної компоненти (наприклад, відвідувач) та додаткові правила агрегування, фільтрування та перетворення інформації (наприклад, агрегація відвідувачів за регіонами або агрегація запитів за часом доби). У простіших задачах потрібно знаходити множини, які характеризують поведінку користувача на сайті і мають просту структуру, яка характеризує співвідношення однієї компоненти відносно іншої (наприклад, розподіл імовірності). Проте у деяких випадках результат аналізу вимагає введення додаткових понять та об'єктів (наприклад, при визначенні типових шляхів навігації по сайту).

ФОРМУЛЮВАННЯ ЦІЛЕЙ СТАТТІ

Основною задачею даного наукового дослідження є побудова формальної математичної моделі поведінки відвідувача Веб-сайту. Ця модель повинна стати основою для подальших досліджень щодо побудови методів та алгоритмів дослідження поведінки користувача сайту, які, в свою чергу, є важливим елементом загальних методів адаптації сайту під вимоги активного середовища, в якому він функціонує.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ ДОСЛІДЖЕННЯ

Веб-сайт (в подальшому сайт) являє собою множину статичних або динамічно згенерованих сторінок.

$St = \{Pt_i\}$, де Pt – розділи сайту, частковим випадком є сайт з одним розділом.

$Pt = \{Pg_i\}$, де Pg_i – веб-сторінка.

Веб-сторінка (в подальшому сторінка) містить певну текстову та/або графічну інформацію та навігаційні елементи переходу на інші сторінки. Частковим випадком є сторінки, які не мають навігаційних елементів. Кожна сторінка сайту має свою семантику, яка задається за допомогою ключових слів, які дозволяють знайти цю сторінку в Інтернеті.

$$Pg = (Inf, Ln, Kw),$$

де Inf – інформація розміщена на веб-сторінці, Ln – навігаційні елементи веб-сторінки, Kw – ключові слова сторінки.

Важливим поняттям при здійсненні аналізу функціонування Веб-сайту є шлях навігації. Шлях навігації по сторінках – це множина

$$W_{pg}^{pg} = \{(Pg_j^r, Pg_j)\},$$

де Pg_j^r – сторінка, з якої зайшли (referrer), Pg_j – сторінка, на яку зайшли.

Традиційно шлях навігації зображається у вигляді орієнтованого графа.

Аналогічно визначається шлях навігації типу сторінка – розділ W_{pt}^{pg}, розділ – сторінка W_{pt}^{pg}, розділ – розділ W_{pt}^{pt}.

$$W_{pt}^{pg} = \{(Pg_j^r, Pt_j)\},\ W_{pg}^{pt} = \{(Pt_j^r, Pg_j)\},\ W_{pt}^{pt} = \{(Pt_j^r, Pt_j)\},$$

де Pt_j^r – розділ сайту, з якого зайшли, Pt_j – розділ сайту, на який зайшли.

Побудова шляху навігації передбачає проведення аналізу використання користувачем методів навігації та елементів навігації на кожній сторінці.

Для розуміння поняття “відвідувач” та характеристик, пов’язаних з ним, дамо формальне визначення цих понять.

Аудиторія відвідувачів $A = \{A_i\}$, де A_i групи користувачів, об’єднані за певними ознаками. $A_i = \{C_j\}$

Відвідувач визначається як

$$C = (Id_c, G_c, T_c, P_c, Sm_c),$$

де Id_c – індивідуальні характеристики (атрибути) відвідувача. Визначаються як

$$Id_c = (S, Ag, M, J, Ad),$$

де S – стать відвідувача, Ag – вік, M – сімейний стан, J – рід занять, Ad – адреса. Частковим випадком відвідувача є інтелектуальні активні пошукові агенти, для яких ці параметри не характерні.

- G_c – регіональні характеристики, які можна визначити так:

$$G_c = (C_G, Ct_G, T_G),$$

де C_G – країна, Ct_G – місто, T_G – часова зона.

- T_c – технічні характеристики. Визначаються так:

$$T_c = (O_t, B_t, L_t, M_t),$$

де O_t – операційна система, що встановлена на комп’ютері відвідувача, B_t – броузер, яким користується відвідувач, L_t – мова операційної системи, M_t – розподільча здатність монітора.

- S – сесії, проведені відвідувачем на сайті $S=\{S_i\}$

$$S_i = \left(Id^s, P_i^s, P_o^s, C_p^s, T_i^s, T_o^s, Tm^s, D_{Tm}^s, Th^s, G^s, Sm^s\right),$$

де Id^s – індивідуальні характеристики відвідувача, P_i^s – сторінка, з якої відвідувач вперше зайшов на сайт в межах сесії, P_o^s – сторінка, з якої відвідувач полишив сайт, C_p – кількість переглянутих сторінок за сесію, T_i^s – час першого заходу в межах сесії, T_o^s – час виходу із сайту, Th^s – технічні характеристики, Tm^s – середня тривалість перебування відвідувача на одній сторінці протягом сесії, D_{Tm}^s – дисперсія величини Tm^s, G^s – регіональні характеристики, Sm^s – семантичні характеристики.

Технічні та регіональні характеристики необхідно визначати для кожної сесії у зв'язку з можливістю заходу відвідувачем на сайт з різних комп'ютерів (з дому, з роботи, з Інтернет-клубу), з різного географічного розташування (з різних міст, країн).

Ряд характеристик сесії та користувача є достатньо близькими та схожими. Відмінністю поняття користувача від сесії є вищий рівень агрегації. Фактично у випадку, якщо відвідувач за весь час існування сайту відвідував його один раз, то ці поняття рівнозначні.

Окрім того, для ідентифікації користувача можуть подаватись додаткові характеристики, які можуть визначатись лише шляхом аналізу кількох його сесій (наприклад, кількість повторних заходів на сайт).

- Sm_c – семантичні характеристики відвідувача, які визначаються так:

$$Sm_c = \left(I_{s,} P_s, R_s\right),$$

де I_s – метод та шлях першого заходу на сайт, P_s – множина переглянутих сторінок, R_s – рейтинг переглянутих сторінок.

Рейтинг сторінки визначається рядом факторів:

1. Змістом сторінки, її інформативністю, наявністю навігаційних елементів;
2. Тривалістю перебування на ній інших відвідувачів;
3. Кількістю повторних заходів на сторінку одного і того же користувача протягом певних періодів часу;
4. Частота визначення користувачами сторінки як фаворита.

Елементи множини С можна поділити на дві підгрупи: відомі ($К_е$), невідомі (U_e)

У більшості випадків елементи множини U_e можна визначити на основі елементів множини $К_е$.

Відображення $К_е$ в U_e робиться на основі логів (log – англ., журнал реєстрації, протокол), які генеруються веб-серверами, анкет відвідувачів, спеціальних програмних агентів, ручного збору інформації.

З логів зокрема можна визначити такі характеристики: G_c, T_c, P_c. Анкети відвідувачів дають основну інформацію, з якої визначається індивідуальна характеристика (Id_c). Анкети формуються адміністраторами сайтів. За допомогою програмних агентів можна сформувати довідник структури сайту.

Базовою інформацією для аналізу поведінки користувача є журнал реєстрації відвідувачів (лог-файл), що формується Веб-сервером. Лог (Lg) являє собою множину записів, кожний з яких фіксує певну дію відвідувача на сайті: $Lg = \{Lg_i\}$.

$$Lg_i = \left(Id, Rf, Sq, Hs, Ip, Pip, Th, Tm, Ck, Mp\right),$$

де *Id* – унікальний код, *Rf* – адреса, звідки зайшов відвідувач, *Sq* – рядок запиту, *Hs* – сторінка, на яку зайшов відвідувач, *Ip* – IP-адреса відвідувача, *Pip* – IP проксі-проксі сервера відвідувача, *Th* – технічні характеристики комп'ютера відвідувача, *Tm* – час здійснення запиту, *Ck* – кукізи (cookie – англ., в Інтернет – невеликий фрагмент даних про історію звернення конкретного відвідувача до даного сайту, що автоматично створюється сервером на комп'ютері користувача) , *Mp* – метод здійснення запиту.

На базі цих даних можна визначити ряд додаткових характеристик. Наприклад, на основі IP-адреси можна визначити географічне розташування відвідувача; знаючи рядок запиту, можна визначити, чи заносив відвідувач сторінку до фаворитів та ін.

Для кращої формалізації поведінки відвідувача сайту введемо множину кліків (clicks) $Cl = \{Cl_i\}$, яка в більшості випадків може бути тотожна множині *Lg*. Проте в ряді випадків елементи множини *Cl* містять деяку додаткову інформацію порівняно з елементами множини *Lg*.

Переважно аналіз функціонування сайту та поведінки відвідувача сайту зводиться до визначення різноманітних показників активної компоненти та її взаємодії з пасивною. Проте, безсумнівно, є важливим і дослідження самого Веб-серверу на основі того ж масиву даних, що і в попередньому випадку. Зокрема, такі дослідження дозволяють визначити правильну структуру сайту, уточнити розділи. Крім того, визначається доцільність існування елементів сторінки (зокрема, навігаційних та інтерактивних).

Загалом, аналізуючи статистичні дані, можна виділити такі класи досліджень:

1. Технічні характеристики клієнта.
   a) операційна система;
   b) Інтернет – броузер та його характеристики – мова інтерфейсу, формати файлів, з якими він може працювати;
   c) IP адреса – дозволяє визначити провайдера і орієнтовне географічне місцеположення (див. далі);
2. Навігація.
   a) визначення ваги внутрішніх посилань, рейтинги внутрішніх посилань – такі рейтинги корисні для розробників сайтів: знаючи, як саме рухаються відвідувачі сайту, можна відповідно розміщувати в потрібних місцях важливу інформацію.
   b) ресурс, з якого відвідувач полишив сайт, – зростання кількості відвідувачів, що полишив сайт з якогось певного ресурсу (групи ресурсів), свідчить про необхідність переглянути його зміст.
3. Географічне розташування.
   a) провайдер відвідувача – визначається через маску мережі провайдера; для цього необхідно сформувати довідник провайдерів та масок їх мереж;
   b) регіон розташування – визначається через провайдера відвідувача.
4. Джерело входу на сайт.
   a) пошукові машини, їх рейтинг – дозволяє вести цілеспрямовану рекламну кампанію, передбачає проведення декодування та аналіз ключових слів, за якими було знайдено сайт.
   b) вхід із сайтів партнерів (банерів) – дозволяє визначити пріоритети розвитку співпраці із партнерами.

5. Web-сайт.
   a) помилка при вході – визначається через статистичну інформацію сервера, на якому розміщено сайт, дозволяє проаналізувати причину появи та передбачити шляхи її вирішення в майбутньому;
   b) рейтинги ресурсів;
   c) рейтинги категорій. Ресурси сайту належать до визначених категорій. аналіз по категоріях є аналогічним до аналізу по ресурсах, проте забезпечує вищий рівень абстрагування. Це дозволяє виявити причини популярності одних та непопулярності інших ресурсів, врахувати їх у подальшій роботі;
6. Сесії.
   a) тривалість сесії в часі;
   b) кількість запитів у сесії;
   c) "ціна" сесії – наскільки "корисним" був відвідувач: тільки ознайомився із сайтом, залишив відгук, зробив замовлення тощо.
7. Обсяги переданої інформації.

Технічно аналіз за вищепереліченими напрямами, якщо можливо здійснюється в часовому розрізі (години, доби, тижні тощо), формуються рейтинги.

Далі наведемо ряд прикладів формалізації задач аналізу поведінки відвідувача сайту.

Приклад 1

Регіональні уподобання аудиторії (за розділами) визначаються як така функція:

$$F_{geo}(R_{geo}): Lg \to P(A(R_{geo}), Pt),$$

де R_{GEO} – правило поділу аудиторії за регіонами (задається наперед); $P(A(R_{GEO}), Pt)$ – відношення, що характеризує зацікавленість частини аудиторії $A(R_{GEO})$ у розділі Pt. Очевидним методом задання цього відношення є розподіл імовірності звернення відвідувача з відповідної частини аудиторії до відповідної частини сайту –

$$P(A(R_{GEO}), Pt_j) = \left\{\Pr\left(G_i, Pt_j\right)\right\}_{Pt_j \in Pt}^{G_i \in G_c},$$

де $\Pr\left(G_i, Pt_j\right)$ – імовірності звернення відвідувача C_i до частини сайту Pt_j.

Приклад 2

Активність відвідувачів на сайті впродовж певного періоду (день, тиждень, місяць, рік) можна відобразити функцією

$$F_{Time}: Lg \to P(A, Tm).$$

$P(A, Tm)$ – відношення, що характеризує зацікавленість аудиторії A сайтом в періоди Tm. Можливим методом задання цього відношення може бути розподіл імовірності звернення відвідувача аудиторії до сайту в певний час –

$$P(A, Tm) = \left\{\Pr\left(A, Tm_i\right)\right\}_{Tm_i \in Tm},$$

де $\Pr\left(A, Tm_i\right)$ – імовірність звернення аудиторії A до сайту в час Tm_i.

Приклад 3

Визначимо функцію, яка відображає популярність засобів навігації серед відвідувачів:

$$F_{Ng}: Lg \to P(Cl, Pg(R_{nav})),$$

де R_{nav} – правило поділу сторінки на навігаційні елементи; $P(Cl, Pg(R_{nav}))$ – відношення, що характеризує використання навігаційного елементу $Pg(R_{nav})$ при кліканні Cl. Очевидним

методом задання цього відношення є імовірність клікання на певному навігаційному елементі –

$$P(Cl, Pg(R_{nav})) = \left\{\Pr\left(Cl_j, Ln_i\right)\right\}_{Ln_j \in Ln}^{Cl_j \in Cl},$$

де $\Pr\left(Cl_j, Ln_i\right)$ – імовірність клікання Cl_j на навігаційному елементі Ln_i.

Приклад 4.

Побудуємо функцію, яка визначає типові шляхи навігації по сторінках:

$$F_w : Lg \to \left\{(W_i, \Pr(W_i))\right\}_{\Pr(W_i) \geq const},$$

де $\left\{(W_i, \Pr(W_i))\right\}_{\Pr(W_i) \geq const}$ функція визначає набір навігаційних шляхів *W*, ймовірність виникнення яких є вища за *const* – константа, що задає нижню межу імовірності виникнення даного шляху.

Приклад 5

Побудуємо функцію, яка визначає, за якими ключовими словами заходять на певні розділи сайту.

$$F_{Kw} : Lg \to \left\{(Pt_i, \Pr(Kw_i))\right\}_{\Pr(Kw_i) \geq const},$$

де $\left\{(Pt_i, \Pr(Kw_i))\right\}_{\Pr(Kw_i) \geq const}$ функція визначає розділи *Pt* сайту, імовірність потрапляння на які за ключовими словами *Kw*, вища за *const* – константа, що задає нижню межу імовірності потрапляння.

ВИСНОВКИ

У результаті проведеного дослідження було дано формальне визначення основних понять задачі інтелектуального аналізу поведінки відвідувачів Веб-сайту. Було визначено такі поняття, як веб-сайт, веб-сторінка, відвідувач, лог-файл та ін. Наведено приклади формалізації ряду задач на основі поданих визначень.

Подане формальне визначення дозволяє надалі формалізувати задачі, пов'язані з аналізом поведінки відвідувачів, побудувати алгоритми аналізу лог-файлів.

1. Буров Є.В., Пелещишин А.М. Оптимізація розміщення даних у Web-системах. // Вісн. Держ. ун-ту "Львівська Політехніка". – 1998. – № 330. – С. 17–27. 2. Васкевич Д. Стратегии Клиент/Сервер. – 2-е изд. – К.: Диалектика, 1996. – 384 с. 3. Галайко В.М. Розроблення інтелектуальних Web-систем. // Вісн. Держ. ун-ту "Львівська Політехніка". – 1998. – № 330. – С. 53–62. 4. Зайцев С.С. Описание и реализация протоколов сетей ЭВМ. – М.: Энергия, 1980. – 155 с. 5. Getting Back to Back: Alternate Behaviors for a Web Browser's Back Button, Saul Greenberg, Andy Cockburn, // http://www.cpsc.ucalgary.ca/grouplab/ papers/ greenberg99getting.pdf. 6. Evolving Visit Behavior in Clickstream Data, Wendy W. Moe and Peter S. Fader, 2001.