УДК 681.3

Олецький О. В.

ЗАСТОСУВАННЯ ГРАФУ «ОНТОЛОГІЯ–ДОКУМЕНТ» ДО ЗАДАЧІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ПОВЕДІНКИ ВІДВІДУВАЧІВ ВЕБ-РЕСУРСІВ

Розглянуто можливості, пов'язані з інтелектуальним аналізом поведінки відвідувачів тематичного веб-порталу та використанням результатів такого аналізу. Запропоновано підхід у використанні формалізації інформаційного наповнення порталу у вигляді графа «онтологія–документ».

***Ключові слова**:* інтелектуальний аналіз даних, Web Usage Mining, онтологія, веб-портал.

Вступ

Сьогодні інтенсивно розвиваються методики Data Mining, тобто інтелектуального аналізу даних, пошуку закономірностей, які пояснюють наявні дані, видобутку знань з сирої інформації [4]. Як самостійний напрямок виокремлюється Web Usage Mining [1], пов'язаний з аналізом відвідуваності веб-ресурсів та виявлення закономірностей, що висвітлюють поведінку відвідувачів.

Одну з найтиповіших задач *Web Usage Mining* можна охарактеризувати так у загальних рисах: знаючи історію навігації даного відвідувача, тобто послідовність сторінок $P_1,... P_n$, що були переглянуті цим відвідувачем, виявити закономірності здійснених переходів, і на основі цього – вірогідність того, що він перейде за деяким посиланням на сторінку q, а також оцінити ступінь його зацікавленості в цій сторінці.

На інформативніший аналіз можна сподіватися, якщо враховувати зв'язки між профілями відвідувачів і фактичними переходами, а також семантичний контекст самих сторінок, тобто їх тематику та семантичні околи, множини ресурсів, споріднених з ними за тим чи іншим критерієм.

Таким чином, на досить загальному рівні можна розглядати:

– множину можливих моделей, які дають змогу описувати характеристики відвідувачів;

– описи відвідувачів, отримані на основі тих чи інших моделей;

– множину можливих моделей, які описують інформаційні ресурси;

– описи ресурсів, отримані на основі відповідних моделей;

– послідовності переходів, здійснених різними відвідувачами;

закономірності, які описують зв'язки між описаними компонентами; саме виявлення і аналіз цих зв'язків і є основною задачею інтелектуального аналізу даних в рамках *Web Usage Mining*.

Основний вміст роботи

Перейдемо до конкретніших формулювань. У дослідженнях [3; 2 та ін.] розвивається підхід, спрямований на формалізацію зв'язків між поняттями предметної галузі, з одного боку, та окремими інформаційними ресурсами, з іншого боку, на основі формальних моделей онтологій. У рамках цього підходу виокремлено різні типи зв'язків між окремими концептами предметної галузі та між концептами і документами; кожному типу зв'язків надано свій ваговий коефіцієнт. Такий підхід може виявитися особливо корисним для реалізації алгоритмів локального пошуку на тематичних порталах, для яких характерна прив'язка до певної предметної галузі, висока інформаційна зв'язність, тематична однорідність, достатньо висока структурованість та якість інформаційного наповнення. Очевидно, що інформаційний пошук на таких ресурсах має суттєво спиратися на семантику, онтологію предметної галузі.

Модель інформаційного наповнення вебресурсу розглядається як граф «онтологія–документ», що будується на основі формальних моделей онтології. Як базова модель розглядається трійка $M=<W^*,D,L>$, де W – онтологія предметної галузі, W^* – розширена онтологія, наповнення онтології W конкретними екземплярами класів (фактично – база знань), D – множина документів; L – множина зв'язків між W^* та D. Власне, онтологія описана як трійка $<Q, R, F>$, де Q – множина класів, які відповідають поняттям предметної галузі, R – множина зв'язків між ними, а F – множина функцій інтерпретації. Відповідно, розширена онтологія описується як трійка $<Q^*, R^*, F^*>$, де Q^* – множина класів разом з їх екземплярами, R^* – множина зв'язків між цими елементами, а F^* – множина функцій інтерпретації, визначених у найпростішому випадку на елементах з Q^*, R^* та $Q^* \times R^* \times F^*$. Тоді елементи D можуть бути значеннями функцій з F^*. По суті, така формалізація описує граф, вузли якого відповідають поняттям предметної галузі та інформаційним ресурсам, а дуги – зв'язкам між ними, причому ці зв'язки можуть бути різних типів.

Далі, якщо w є елементом розширеної онтології, а d – артефактом інформаційної системи, то функції інтерпретації f та відповідні вагові коефіцієнти можуть формуватися на основі цих категорій сутностей. Таким чином, здійснюється перехід до моделі «онтологія – артефакт – користувач –проект», в якій міри важливості зв'язків залежать від характеристик та цілей відвідувачів. Альтернативний погляд на проблему може полягати в побудові багатокомпонентної онтологічної системи, окремі компоненти якої відповідають окремим категоріям сутностей. Така класифікація дає змогу будувати певні евристичні правила для підвищення цілеспрямованості пошуку. По суті, ідея цих правил має полягати в аналізі запиту, його зарахування до тієї чи іншої ситуації і в прийнятті рішення залежно від цієї ситуації.

Навігаційний граф веб-ресурсу, вузли якого відповідають окремим документам, а дуги – гіпертекстовим посиланням, може формуватися динамічно на основі аналізу графа «онтологія–документ». Мова може йти про оптимізацію структури веб-ресурсу, динамічне формування структури гіпертекстових посилань та ін.

Нехай W – множина понять предметної галузі, D – множина артефактів інформаційної системи, Q – задана множина можливих типів зв'язків, зокрема між поняттями предметної галузі, а також між поняттями предметної галузі та артефактами інформаційної системи. Позначимо через $r_q(w, d)$, де $q \in Q$, $w \in W$, $d \in D$, ступінь релевантності документа d поняттю w за зв'язком q.

Природно залучити до розгляду деяку комбіновану міру релевантності документа d поняттю w, усереднену за всіма зв'язками, враховуючи їхні вагові коефіцієнти:

$$R(w,d) = \sum_{q \in Q} \alpha_q r_q(w,d), \qquad (1)$$

де α_q – вага (змістовно – міра важливості) q-го типу зв'язків. Природно ставити питання про автоматизований підбір параметрів співвідношення (1), зокрема на основі методик *Data Mining*.

Таким чином, в рамках онтологічно-орієнтованого підходу, що описується, можна розглядати такі постановки задач *Web Usage Mining*:

– множина відвідувачів розбивається на кластери або за власними профілями, або за історією навігації; для кожної групи з'ясовано найпріоритетніші типи зв'язків між вузлами графа «онтологія–документ», і на цій основі розставлено персоналізовані вагові коефіцієнти, що залежать від характеристик відвідувачів;

– на основі аналізу історії переходів між вузлами графа «онтологія–документ» оцінена вірогідність того, що, перебуваючи у вузлі q з певним значенням характеристики a, відвідувач перейде за посиланням, яке відповідає типу зв'язків r;

– оптимізація структури навігаційного графа задля скорочення послідовності переходів, які відвідувач має зробити, щоб досягти мети;

– ефективний добір контекстної реклами, яка була б пов'язана з ресурсами з найвищою оцінкою релевантності, тобто з тими, які могли б з найбільшою вірогідністю зацікавити відвідувача, що в даний момент перебуває в деякому вузлі графа «онтологія – документ»;

– прийняття рішень за аналогією (наприклад, якщо користувач А для розв'язання задачі *C* вважає корисним документ *W*, то користувачеві *X*, характеристики якого схожі на характеристики користувача А, для розв'язання задачі *K*, схожої на *C*, можна порекомендувати список документів, схожих на *W*);

– налаштування параметрів процесу навчання та самоорганізації порталу на основі поширення активації, який у загальних рисах описаний в [3];

– аналіз породжуючих моделей, які зумовлюють власне навігаційний процес, та аналіз можливого впливу параметрів цих моделей на оцінки мір релевантності інформаційних ресурсів; цей підхід у загальних рисах описано в [2].

Серед методик Data Mining, які видаються найперспективнішими для розв'язання перелічених завдань, слід звернути увагу на такі [4]:

– побудова дерев рішень з метою формування наборів правил «якщо *A*, то *B*»;

– пошук асоціацій і алгоритм Аргіогі з метою виявлення закономірностей типу «*A* і *B* часто зустрічаються разом»;

– кластерний аналіз, щоб розбити на кластери профілі відвідувачів та історії їх навігацій;

– генетичні алгоритми задля оптимізації параметрів співвідношення (1), а також оптимізації структури навігаційних графів.

Висновки

У дослідженні в загальних рисах описано, яким чином модель інформаційного наповнення веб-ресурсу, зокрема тематичного порталу, що будується на основі формальної моделі онтології у вигляді графа «онтологія–документ», може буде застосована до інтелектуального аналізу поведінки відвідувачів ресурсу. Наведена формалізація дає можливість виокремити параметри моделі, на які слід звернути увагу насамперед. Уточнення та подальші формалізації підходу стануть предметом наступних досліджень.

Література

1. Гончаров М. Web Mining – добыча знаний из World Wide Web [Електронний ресурс]. – Режим доступу: http://www.spellabs.ru. – Назва з екрана.
2. Олецький О. В. До проблеми моделювання потоку відвідувань на онтологічно-орієнтованому тематичному порталі. // О. В. Олецький / Моделювання та інформаційні технології. Збірник наукових праць. Спеціальний випуск. – 2010. – Т. 2. – С. 321–326.
3. Олецький О. В. Онтологічно-орієнтований інформаційний пошук на основі хвильового процесу поширення активації //О. В. Олецький / Наукові записки НаУКМА. Комп'ютерні науки. – 2008. – Т. 86. – С.50–52.
4. Технологии анализа даннях : Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – СПб. : БХВ-Петербург, 2007. – 384 с.

O. Oletsky

USE OF "ONTOLOGY–DOCUMENT" GRAPH FOR ANALYZING USERS BEHAVIOUR AT THE WEB-PORTAL

Facilities connected to intelligent analyzing behaviour of web portal users and using results of such analysis are regarded. An approach connected to formalizing information model of the portal as a graph «ontology – document» is proposed.

Keywords: Data Mining, Web Usage Mining, ontology, web-portal.

Матеріал надійшов 11 травня 2011 р.