УДК 681.3

Олецький О. В.

ПОБУДОВА ФОРМАЛІЗОВАНОГО ОПИСУ ГРАФА «ОНТОЛОГІЯ–ДОКУМЕНТ» ЯК МОДЕЛІ ІНФОРМАЦІЙНОГО НАПОВНЕННЯ ТЕМАТИЧНОГО ПОРТАЛУ

Запропоновано уточнення та формалізацію надалі онтологічно-орієнтованої моделі інформаційного наповнення тематичного порталу, розглянутої в раніших роботах. Модель подано як семантичну мережу, розділену на три компоненти, які описують онтологію предметної області, простір документів інформаційної системи, а також зв'язки між ними.

Ключові слова: онтологія, інформаційний пошук, тематичний портал, семантична мережа.

Вступ

Проблема підвищення якості інформаційного пошуку [4; 5 та ін.] за всієї актуальності далека від повного розв'язання. Навіть такі базові поняття, як релевантність та повнота пошукової видачі, формалізовані недостатньо.

У працях [6; 7] розвинуто підхід, спрямований на поліпшення якості інформаційного пошуку на тематичному порталі. Характерними рисами такого порталу є висока інформаційна зв'язність, тематична однорідність, достатньо висока структурованість та якість інформацій-

ного наповнення. Такі ресурси можуть набувати, зокрема, навчальних і науково-дослідницьких рис або віртуальних співтовариств, які можуть розвиватися на цій основі.

З огляду на це якість інформаційного пошуку на такому порталі може бути поліпшена завдяки максимальному врахуванню семантики, онтології предметної області. В основі підходу, який розглядається, лежить спроба побудови моделі інформаційного наповнення тематичного порталу на основі аналізу графа «онтологія–документ». Цей граф може бути занурений у контекст, пов'язаний із цілями відвідувачів порталу та з їхніми індивідуальними характеристиками.

Як базову модель розглянуто трійку $M = <W^*, D, L>$, де W – онтологія предметної області, W^* – розширена онтологія, наповнення онтології W конкретними екземплярами класів (фактично, база знань), D – множина документів; L – множина зв'язків між W^* та D. Власне онтологію описано як трійку $<Q, R, F>$, де Q – множина класів, які відповідають поняттям предметної області, R – множина зв'язків між ними, а F – множина функцій інтерпретації. Відповідно розширену онтологію описано як трійку $<Q^*, R^*, F^*>$, 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^* і $Q^* \times R^* \times F^*$. Тоді елементи D можуть бути значеннями функцій з F^*. По суті, така формалізація описує граф, вузли якого відповідають поняттям предметної області та інформаційним ресурсам, а дуги – зв'язкам між ними, причому ці зв'язки можуть бути різних типів.

Наведена модель має досить інтуїтивний характер, і цих інтуїтивних міркувань недостатньо для повноцінного аналізу та побудови на цій основі математично обґрунтованих мір подібності та релевантності документів. Ця робота спрямована на побудову більш формалізованих моделей.

Основний зміст дослідження

Будуватимемо модель інформаційного наповнення тематичного порталу на основі наступних міркувань. Нехай W – множина вузлів онтології предметної області. Будемо вважати, що $W=T\cup V$, де T – множина термінальних концептів (змістовно – множина термінів, які можуть траплятися в документах), V – множина метасимволів, тобто нетермінальних концептів, понять, що не можуть безпосередньо з'являтися в документах, але які впливають на їх формування та ранжування за релевантністю. Можна вважати, що $T\cap V=\varnothing$, оскільки у випадку, якщо слово з тексту водночас є важливим поняттям онтології, можемо просто формально перейменувати відповідний нетермінальний концепт.

Аналогічно множину документів D подамо як $D = DT\cup DV$, де DT – множина документів як таких, DV – множина категорій документів.

Введемо наступні множини можливих зв'язків (фактично, онтології зв'язків):

R – множина типів зв'язків між концептами предметної області (насамперед нетермінальними);

L – множина типів зв'язків між концептами предметної області та документами;

H – множина типів зв'язків між документами.

Тоді формалізовану модель інформаційного наповнення онтологічно-орієнтованого тематичного порталу можна записати як

$$M = W^*\cup Z^*\cup D^*, \qquad (1)$$

де W^* – множина кортежів (w_1, r, w_2); $w_1, w_2 \in W$, $r \in R$;

Z^* – множина кортежів (w, l, d); $w \in W$, $l \in L$; $d \in D$;

D^* – множина кортежів (d_1, h, w_2); $d_1, d_2 \in D$, $h \in H$.

На цій основі можна уточнювати поняття релевантності. Наприклад, у деякому наближенні можна прийняти, що документ $d\in DT$ є релевантним до ключового слова $t \in T$, якщо існують такі $r \in R$, $v \in V$, $l \in L$, $h \in H$, $d' \in DV$, що (t, r, v), (v, l, d'), $(d', h, d) \in M$.

Звичайно, важливішим є питання про механізми обчислень кількісних мір релевантності на базі моделі (1) або подібних моделей, і на цій основі – про ранжування документів або концептів предметної області за ступенем їхньої відповідності запитові користувача. Більш загально: задача отримання мір близькості між вузлами моделі (1) може бути розв'язана на основі знаходження ланцюжка переходів між ними та відповідної інтерпретації вагових коефіцієнтів переходів. У [6; 7] наведено певні міркування з приводу побудови подібних мір на основі зваженої векторно-просторової моделі й теоретико-множинних мір близькості, а також комбінованих мір релевантності на базі відомих підходів, описаних у [3; 4 та ін.]. У найзагальніших рисах, ці підходи можна описати так.

1. Застосування зваженої векторно-просторової моделі пошуку. Матрицю «документ–термін» природно розглядати як окремий випадок матриці даних у деякому просторі ознак, широко відомої в математичній статистиці та в розпізнаванні

образів. Дійсно, така матриця даних може мати вигляд $Q = \{q_{ij}\}$, q_{ij} – міра зв'язку між елементом $T_i \in T$; $W_j \in W$; T та W – деякі множини елементів. У «класичній» векторно-просторовій моделі використовують множини документів і термінів, але ніщо не заважає залучати до розгляду інші категорії елементів, а також різноманітні міри близькості між векторами матриці. Звичайно ключовим залишається питання: як саме слід формувати міри зв'язку q_{ij}. Зокрема, онтологічний аналіз може бути врахований у використанні та узагальненні традиційного матричного підходу до побудови мір релевантності між поняттями предметної області й документами, що їм відповідають. Узагальнену матрицю «термін–документ» можна подати як

$$U = KC, \qquad (2)$$

де K – матриця зв'язків між термінами (поняттями), C – матриця, елементи якої відповідають кількостям входжень терміна до документа. У частковому випадку, якщо $K = E$, то $U = C$.

Елементи матриць Q та H можна розглядати як незалежні параметри моделі. Проте методика може стати гнучкішою, якщо пов'язати ці параметри з різними типами зв'язків. Точніше, нехай $L = \{l_k\}$ – онтологія зв'язків між вузлами графа і $\lambda(l_k)$ – вага зв'язку l_k. Як зазначали, ці ваги можуть залежати від мети і характеристик відвідувача. Якщо вузли w_i та t_j пов'язані зв'язком l_k, то в найпростішому випадку можна покласти $q_{ij} = \lambda(l_k)$.

2. Теоретико-множинний аналіз споріднених елементів. Як базовий тут прийнято розглядати такий підхід: якщо R_a – множина елементів, пов'язаних з елементом a, а R_b – множина елементів, пов'язаних з елементом b, то мірою подібності між елементами a і b виступає співвідношення $\frac{|R_a \cap R_b|}{|R_a \cup R_b|}$. Очевидним розвитком цього підходу стає врахування вагових коефіцієнтів, пов'язаних із тим чи іншим типом зв'язків.

3. Комбіновані міри релевантності. Позначимо через $m_q(w, d)$, де $w \in W$, $d \in D$, міру релевантності документа d поняттю w за зв'язком q. Природно залучити до розгляду комбіновану міру релевантності документа d поняттю w, усереднену за всіма зв'язками з урахуванням їхніх вагових коефіцієнтів:

$$R(w,d) = \sum_{q \in Q} \alpha_q m_q(w,d) \qquad (3)$$

де α_q – вага (змістовно – міра важливості) q-го типу зв'язків. Для добору параметрів співвідношення (2) природним є застосування генетичних алгоритмів [8 та ін.], деякі підходи до розв'язання цієї проблеми в загальних рисах описано у праці [9]. Іншою можливістю для добору коефіцієнтів моделі мають стати методики Data Mining та Web Usage Mining [1; 2 та ін.]. Серед тих, які видаються найперспективнішими для оптимізації онтологічно-орієнтованого пошуку на тематичному порталі, слід звернути увагу на:

– побудова дерев рішень для формування наборів правил «якщо A, то B»;
– пошук асоціацій і алгоритм Apriori для виявлення закономірностей типу «A і B часто зустрічаються разом»;
– кластерний аналіз задля розбиття профілів відвідувачів та історій їхніх навігацій на кластери.

Висновки

Тут описано формалізовану модель інформаційного наповнення тематичного порталу, яка в найпростішому випадку складається з трьох окремих порівняно незалежних фрагментів. Модель, описана як відношення (1) – це, по суті, семантична мережа з явно виділеними компонентами, що відповідають предметній області, множині документів, а також зв'язкам між ними. На цій основі розглянуто можливість побудови мір релевантності й подібності між вузлами мережі та концептами предметної області – термінальними і нетермінальними. Але побудова більш формалізованих та обґрунтованих методик має бути предметом подальших досліджень.

Подальший розвиток моделі (1) може бути пов'язаний з такими напрямками:

– перехід до фреймово-семантичного подання завдяки тому, що вузли мережі можуть описуватися на основі відповідних фреймів та/або класів, а дуги – семантичні зв'язки між ними;
– чіткіше розділення понять фрейму як однієї з моделей знань у рамках теорії штучного інтелекту та класу як основного поняття об'єктно-орієнтованого аналізу, проектування і програмування. Деякі міркування з приводу різниці між фреймовими та об'єктними моделями можна знайти у [3]. Наприклад, для фреймів множинна класифікація, тобто ситуація, коли той самий об'єкт одночасно належить кільком різним фреймам, є цілком нормальним явищем. Так, поняття «*зелене яблуко*» водночас стосується і фрейма «*яблуко*», і фрейма «*зелений предмет*». Проте у рамках об'єктної моделі множинна класифікація вважається неприпустимою;
– чітке розділення категорій «елемент належить до множини» і «елемент є екземпляром деякого класу та/або деякого фрейму».

Література

1. Барсегян А. А. Технологии анализа данных : 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.
2. Гончаров М. Web Mining – добыча знаний из World Wide Web. – Режим доступу: http://www.spellabs.ru. – Назва з екрана.
3. Грэхем И. Объектно-ориентированные методы. Принципы и практика / И. Грэхем. – М. : Изд. дом «Вильямс», 2004. – 880 с.
4. Ландэ Д. В. Поиск знаний в Интернет / Д. В. Ландэ. – М. : Изд. дом «Вильямс», 2005. – 272 с.
5. Маннинг К. Д. Введение в информационный поиск / К. Д. Маннинг, П. Рагхаван, Х. Шютце. – М. : ООО «И. Д. Вильямс», 2011. – 528 с.
6. Олецький О. В. Організація онтологічно-орієнтованих засобів автоматизованого добору інформаційних ресурсів на тематичному порталі / О. В. Олецький // Наукові записки НаУКМА. Комп'ютерні науки. – 2009. – Т. 99. – С. 66–69.
7. 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.
8. Олецький О. В. Принципи застосування генетичних алгоритмів до задачі онтологічного інформаційного пошуку / О. В. Олецький // Наукові записки НаУКМА. Комп'ютерні науки. – 2010. – Т. 112. – С. 49–54.
9. Рутковская Д. Нейронные сети, генетические алгоритмы и нечеткая логика / Д. Рутковская, М. Пилиньский, Л. Рутковский – М. : Горячая линия. – Телеком, 2004. – 452 с.

O. Oletskyy

A FORMALIZED DESCRIPTION OF THE GRAPH «ONTOLOGY–DOCUMENT» AS A MODEL OF THE CONTENT OF THE WEB-PORTAL

Further development of the formalized model of an ontology-oriented model for the content of the web-portal is proposed. The model is described as a semantic network divided to three components: ontology of the subject domain, document space and relations between them.

Keywords: ontology, information search, web-portal, semantic network.

Матеріал надійшов: 6.04.2012