ОЛЕЦЬКИЙ О.В.

ШЛЯХИ ІНТЕЛЕКТУАЛІЗАЦІЇ ЛОКАЛЬНОГО ІНФОРМАЦІЙНОГО ПОШУКУ НА ОСНОВІ АНАЛІЗУ ГРАФУ "ОНТОЛОГІЯ-ДОКУМЕНТ"

Проблема покращення якості інформаційного пошуку, підвищення рівня його інтелектуальності [1] стає все більш актуальною. При цьому не викликає сумнівів необхідність орієнтації такого пошуку на онтологію, семантику предметної області.

Мова йде перш за все про проблему локального інформаційного пошуку на веб-ресурсах, для яких характерні висока інформаційна зв'язність, тематична однорідність, достатньо висока структурованість та якість інформаційного наповнення. До таких ресурсів можна віднести, зокрема, тематичні портали, які можуть набувати навчальних та науково-дослідницьких рис, або віртуальних співтовариств, які можуть розвиватися на їх основі.

В роботах [2, 3 та ін.] розвивається підхід на основі аналізу формальної моделі інформаційного наповнення тематичного веб-порталу у вигляді графа "онтологія-документ". Як базова модель розглядається трійка $M=<W^*,D,L>$, 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 онтологія описується як трійка $<Q, R, F>$, де Q – множина класів, які відповідають поняттям предметної області, 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^* x R^* x 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 різним 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 дозволяє будувати певні евристичні правила для підвищення цілеспрямованості пошуку. По суті, ідея цих правил має полягати в аналізі запиту, його віднесення до тієї чи іншої ситуації, і в прийнятті рішення в залежності від цієї ситуації.

Навігаційний граф веб-ресурсу, вузли якого відповідають окремим документам, а дути - 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 області. D – множина артефактів інформаційної системи. Q – задана множина можливих типів зв'язків, зокрема між поняті ями предметної області, а також між поняттями предметної області та артефактами інформаційної системи. Позначимо через $r_q(w,d)$, де $q \in Q$, $w \in W$, $d \in D$, міру релевантності документа d поняттю w за зв'язком q.

Природно залучити до розгляду деяку комбіновану міру релеван шості документа d поняттю w, усереднену за всіма зв'язками з урахуванням їх вагових коефіцієнтів:

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

де α_q – вага (змістовно - міра важливості) q-го типу зв'язків.

В цьому контексті виникає питання про вибір власне мір релевантності $r_q(w,d)$. Для побудови таких мір близькості між вузлами моделі можна застосовувати ряд відомих підходів [1, 3, 4 та ін.]:

1. Булєва та векторно-просторова моделі пошуку, які широко використовуються в сучасних пошукових системах. Але матрицю "документ-термін", яка лежить в основі класичної

векторно-просторової моделі, по своїй суп 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. В "класичній" векторно-просторовій моделі використовуються множини документів та 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 при використанні та узагальненні традиційного матричного підходу до побудови мір релевантності між поняттями предметної області та документами, що їм відповідають. Узагальнену матрицю "термін-документ" можна подати у вигляді

$$V=HC, \qquad (2)$$

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

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

Можна запропонувати подальші узагальнення векторно-матричної моделі. Так. якщо розглядати сімейство матриць зв'язків між вузлами онтології O_i, кожна з яких відповідає i-му зв'язку з вагою v_i та сімейство матриць зв'язків між документами D_j, кожна з яких відповідає j-му зв'язку з вагою

τ_j, то можна розглядати середньозважену матрицю вагових коефіцієнтів

$$W=(\Sigma_i \nu_i O_i) \vee (\Sigma_j \tau_j D_j) \qquad (3)$$

2. Теоретико-множинний аналіз споріднених елементів. Як базовий тут прийнято розглядати наступний 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 з тим чи іншим типом зв'язків.

В роботі [3] розвивається підхід на основі моделювання можливої поведінки відвідувачів. Для підбору параметрів співвідношення (1) природним є застосування генетичних алгоритмів [5 та ін.], деякі підходи до розв'язання цієї проблеми в загальних рисах описані в роботі [2].

Нарешті, звертає на себе увагу можливість застосування методик Data Mining [6 та ін.], в першу чергу – Web Usage Mining [7], пов'язаний з аналізом відвідуваності веб-ресурсів та виявлення закономірностей, що пояснюють поведінку відвідувачів. Одну з найбільш типових задач Web Usage Mining можна в найбільш загальних рисах охарактеризувати наступним чином: знаючи історію навігації даного відвідувача, тобто послідовність сторінок $P_1, \ldots P_n$, 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 міру його зацікавленості в цій сторінці. В рамках онтологічно-орієнтованого підходу на базі моделі "онтологія-документ", 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'ясовуються найбільш пріоритетні типи зв'язків між вузлами графа "онтологія-документ", і на цій основі розставляються

персоналізовані вагові коефіцієнти, то залежать від характеристик відвідувачів;

- на основі аналізу історії переходів між вузлами графа "онтологія-документ" оцінюється вірогідність того, що, перебуваючи у вузлі *q* з певним значенням характеристики *a,* відвідувач перейде за посиланням, яке відповідає типу зв'язків

- оптимізація структури навігаційного графа, з метою скорочення послідовності переходів, які відвідувач має зробити, щоб досягти мети;
- ефективний підбір контекстної реклами, яка була б пов'язана з ресурсами з найвищою оцінкою релевантности 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 С вважає корисним документ W, то користувачеві Х, характеристики якого схожі на характеристики користувача А, для розв'язання задачі К, схожої на С, можна порекомендувати список документів, схожих на W);

Подібні методики інтелектуалізованого пошуку можуть виявитися корисними для спеціалізованих тематичних порталів (підбір матеріалів, які можуть найбільшою мірою зацікавити потенційних відвідувачів), систем контекстної реклами, навчальних порталів (підбір найбільш ефективних та адекватних навчальних матеріалів), системах електронної комерції (встановлення контактів між постачальниками та споживачами певної продукції) тощо.

ЛІТЕРАТУРА

1. Маннинг К.Д., Рагхаван І І. Шютце Х Введение в информационный поиск. - М : О О О « И Д. Вильяме». 2011. - 528 с
2. Отецький О.В. Онтологічно-орієнтований інформаційний пошук на основі хвильового пронесу поширення активації. //1 Нау коні записки Н а У К М А. 1.86. Комп'ютерні пауки. - К.. 2008. (.'.50-52.
3. Олсцький О.В. До проблеми моделювання потоку відвідувань на оптологічно-оріентованому тематичном) порталі. //Моделювання та інформаційні технології. Збірник наукових прань Спеціальний випуск. Т.2.- К.201(1. -€.321-326.

4 Ланд > Д.В. І Поиск знаний в Интернет. - М.: Изл. лом "Вильяме", 2005. 272 с.

5. Рутковекая Д., Пилиньский М.. Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткая логика. - М.: Горячая линия - Телеком, 2004. -452 с.
6 Барссгян А.А., Куприянов М.С., Степаненко В.В.. Холол И.И. Технологии анализа данных: Data Mining. Visual Mining, Text Mining, OLAP. - СПб: БХВ-Петербург, 2007. - 384 с.
7. Гончаров М. Web Mining—добыча знаний из World Wide Web. //Електронний ресурс: http://www.spellabs.ru.