УДК 681.3:658.5

О.В. Олецький

ПРО ДЕЯКІ ПІДХОДИ ДО ПОБУДОВИ ОНТОЛОГІЧНО-ОРІЄНТОВАНИХ МОДЕЛЕЙ ПОШУКУ І САМООРГАНІЗАЦІЇ ДЛЯ ТЕМАТИЧНИХ ПОРТАЛІВ

Розглядаються семантично-орієнтовані моделі, спрямовані на формалізований опис інформаційного наповнення тематичного порталу на основі четвірки “онтологія-артефакт-користувач-проект”, а також на опис процесів пошуку на порталі та самоорганізації порталу на основі хвильового поширення активації. Пропонується використання моделей на основі поєднання комбінованих мір релевантності як зважених сум мір релевантності за окремими зв’язками між вузлами онтології та документами, а також математичних співвідношень між мірами важливості окремих вузлів та зв’язків, які породжуються різними процесами самоорганізації.

Вступ. Однією з найбільш актуальних проблем сучасних інформаційних систем є проблема пошуку інформації, максимально адекватної запитові користувача, та ранжування знайдених результатів відповідно до певних критеріїв релевантності; особливо це стосується веб-орієнтованих пошукових сервісів. Існує ряд підходів до її розв’язання [1], але ці підходи є здебільшого евристичними та враховують лише окремі аспекти проблеми. Навіть такі базові поняття, як ”релевантність документа запитові”, “схожість документів”, формалізовані недостатньо. Крім того, ці підходи недостатньо орієнтовані на семантику предметної області. Очевидною є необхідність переходу від евристичних підходів до більш-менш формалізованих моделей. Не викликає сумнівів і те, що ефективність пошуку, зокрема його точність і повнота, можуть бути суттєво підвищені, якщо ці моделі будуть істотно спиратися на семантику, онтологію предметної області. Особливого значення це набуває для тематичних та навчально-консультаційних порталів, цифрових бібліотек тощо – тобто для систем, для яких характерні тематична однорідність і достатньо висока зв’язність інформаційних ресурсів. Зокрема, онтологічно-орієнтовані методики пошуку дозволяють підвищити ефективність експертного контекстного підбору матеріалів так, щоб в першу чергу показувалися матеріали, які можуть зацікавити відвідувача з найбільшою мірою впевненості.

В роботі [2] розглядається підхід, спрямований на формалізацію моделі інформаційного пошуку як процесу хвильового поширення активації на деякому графі. Цей граф по суті є моделлю інформаційного наповнення системи; в [2] робиться спроба його формалізованого опису на основі формальних моделей онтології та четвірки “онтологія-артефакт-користувач-проект”. В загальних рисах, до розгляду залучаються різні типи зв’язків між поняттями предметної області, з одного боку, і документами, що

зберігаються в системі – з іншого. Міри важливості цих зв'язків залежать також від характеристик відвідувачів і від задач, які їм потрібно розв'язувати.

Ключовим елементом цього підходу є хвильовий процес поширення активації між вузлами базового графа. Розглядаються два основні аспекти цього процесу:

- власне формування мір релевантності документів запитові на основі співставлення запиту з вузлами онтології та подальшим поширенням активації до сусідніх вузлів (з урахуванням параметрів зв'язків між ними);
- самоорганізація, налаштування базової конфігурації у вигляді набору параметрів моделі, які не залежать від запиту.

Звичайно, ці загальні формулювання мають бути уточнені за такими основними напрямками:

- розвиток базових моделей інформаційного наповнення порталу з урахуванням семантики предметної області;
- побудова ефективних методик процесу поширення активації.

Дана робота спрямована на дослідження цих питань.

Викладення основного матеріалу. Як базова розглядається модель "онтологія-артефакт". Вона описана в [2] як 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 $Q^* \times R^* \times F^*$. Тоді елементи D можуть бути значеннями функцій з F^*.

Ідея "занурення" інших категорій сутностей в загальних рисах формулюється так: 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. Це підводить до природного уточнення моделі - "онтологія-артефакт-користувач-проект", що дозволяє встановлювати відповідність між вузлами онтології (поняттями предметної області) та пов'язаними з ними документами, з користувачами та проектами і роботами, в яких вони беруть участь. В контексті підвищення ефективності інформаційного пошуку це найтіснішим чином пов'язано з урахуванням мети пошуку.

На основі такого підходу можна здійснювати подальші формалізації. Нехай W – множина понять предметної області, D – множина артефактів інформаційної системи, Q – задана множина можливих типів зв'язків,

зокрема між поняттями предметної області, а також між поняттями предметної області та артефактами інформаційної системи.

Нехай, далі, $r_q(w,d)$, де $q \in Q$, $w \in W$, $d \in D$ – міра релевантності документа d поняттю w за зв'язком q. Можна навести методики для розрахунку таких мір як на основі певних узагальнень класичної векторно-матричної моделі, так і на основі теоретико-множинного підходу. Зокрема, 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.

Такі міри мають очевидну нечітку інтерпретацію. Дійсно, твердження "міра релевантності документа запитові q дорівнює x" – це по суті те саме, що твердження "документ належить до множини документів, релевантних запитові q, зі ступенем належності x". Такий погляд дозволяє застосовувати до побудови мір подібності та релевантності розвинену теорію нечітких множин.

У [2] зазначається, що типовим результатом пошуку на основі описаної моделі має стати формування множини понять і документів, в тій чи іншій мірі релевантних запиту, з урахуванням формування вагових коефіцієнтів, пов'язаних з функціями інтерпретації. Кожний елемент цієї множини буде поданий у вигляді $(u, m_1(u), m_2(u), \ldots)$, де u – знайдений вузол, $m_i(u)$ - міра релевантності цього вузла, обчислена за i-м критерієм, можливо, недостовірна або нечітка. З метою подальшого уточнення та формалізації цього положення природно залучити до розгляду деяку комбіновану міру релевантності 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.

Таким чином, одним з ключових завдань навчання і самоорганізації пошукової системи полягає в підборі коефіцієнтів α_q формули (1) з урахуванням того, що ці коефіцієнти залежать від інших категорій сутностей, "занурених" в модель "онтологія-артефакт" – зокрема, від користувача і від мети пошуку.

Сьогодні інтенсивно розвиваються нові інтелектуальні методики розв'язання перебірних задач оптимізації та самоорганізації, зокрема на основі інформаційного керування випадковим пошуком та механізмів, що імітують природні еволюційні процеси. В першу чергу це генетичні алгоритми [3] для знаходження оптимальних (або субоптимальних) наборів параметрів, а також методики, характерні для т.зв. ройового інтелекту (в першу чергу - алгоритм мурашки) [4, 5].

Слід відмітити, що використання алгоритму мурашки по суті тяжіє до імітаційного моделювання, оскільки в даному контексті він дозволяє проімітувати поведінку великої кількості відвідувачів. Крім того, подібні методики тісно пов'язані з механізмом перерозподілу певного ресурсу між вузлами графа, а це фактично дозволяє побудувати модель процесу у вигляді деякої системи

диференційних або алгебраїчних рівнянь (подібно до того, як у вигляді таких рівнянь прийнято описувати поширення певної речовини в просторі).

Цю ідеологію можна простежити на прикладі PageRank – відомої методики ранжування сторінок. В основі PageRank по суті також лежить деякий процес пошуку, який полягає в наступному [1]. Користувач відкриває випадкову сторінку, а потім переходить на іншу за випадково вибраним гіперпосиланням, і т.д. Інколи цей процес йому набридає, і він знову вибирає випадкову сторінку. Тоді PageRank сторінки – це ймовірність того, що такий користувач перейде саме на неї.

Математична формалізація цього процесу призводить до моделі у вигляді системи лінійних алгебраїчних рівнянь відносно рангів окремих сторінок PR_i [1]:

$$PR_a = (1-d) + d\sum_{i=1}^{n}\frac{PR_i}{C_i}, \qquad (2)$$

де PR_a – PageRank a-ї сторінки;

C_i – відношення загальної кількості посилань на i-й сторінці до кількості посилань з i-ї сторінки на a-ту;

d – коефіцієнт затухання; зазвичай 0.85.

У цьому контексті ідеологія PageRank спирається на перерозподіл "авторитетності" сторінок; при цьому "авторитетність" сторінки пов'язується з кількістю зовнішніх посилань на цю сторінку. Замість цього можна розглядати моделі, що спираються на перерозподіл інших мір важливості – мір, пов'язаних з важливістю вузлів онтології, якістю документів та наявними експертними оцінками тощо. Крім того, можна певним чином змінити схему і параметри самого процесу. Зокрема, в реальних умовах користувач практично ніколи не починає пошук з випадкової сторінки – пошук завжди починається з певних визначених вузлів, і саме формування набору цих вузлів може стати одним з основних результатів процесу самоорганізації порталу. Отже, можна говорити про ціле сімейство подібних моделей.

Висновки. Таким чином, математичні моделі самоорганізації тематичного порталу та пошуку і ранжування сторінок на такому порталі можуть спиратися на комбінування співвідношення (1) і сімейства співвідношень, подібних до (2). При цьому слід враховувати і те, що міри близькості $r_q(w,d)$, що фігурують у співвідношенні (1), теж можуть формуватися різним чином, а також те, що ці міри можуть носити нечіткий або недостовірний характер.

1. *Ландэ Д.В.* Поиск знаний в Интернет. – М.: Изд. дом "Вильямс", 2005. – 272 с.

2. *Олецький О.В.* Онтологічно-орієнтований інформаційний пошук на основі хвильового процесу поширення активації. //Наукові записки НаУКМА. Т. 86. Комп'ютерні науки. – К., 2008. – С.50-52.

3. *Рутковская Д., Пилиньский М., Рутковский Л.* Нейронные сети, генетические алгоритмы и нечеткая логика. – М.: Горячая линия - Телеком, 2004. – 452 с.

4. *Тарасов В.Б.* От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. – М.: Эдиториал УРСС, 2002. – 352 с.
5. *Джонс М.Т.* Программирование искусственного интеллекта в приложениях. – М.: ДМК Пресс, 2004. – 312 с.