

ОНТОЛОГІЧНО ОРІЄНТОВАНИЙ ІНФОРМАЦІЙНИЙ ПОШУК НА ОСНОВІ ХВИЛЬОВОГО ПРОЦЕСУ ПОШИРЕННЯ АКТИВАЦІЇ

Розглядається підхід, спрямований на формалізований опис інформаційного пошуку як хвильового процесу поширення активації на основі формальної моделі інформаційної системи. Для інформаційного наповнення тематичних порталів, інтегрованих з віртуальними спільнотами, пропонується формальна модель на базі четвірки «онтологія - артефакт - користувач - проект». Обговорюються можливості, пов'язані з застосуванням алгоритмів ройового інтелекту та генетичних алгоритмів.

Підвищення якості інформаційного пошуку, зокрема його повноти і релевантності, сьогодні стає дедалі актуальнішим. Існує значна кількість підходів до вирішення цієї проблеми [1-3 та ін.], але вони здебільшого мають евристичний характер і залучають до розгляду лише окремі аспекти пошуку. Кількісні міри, які дозволили б охарактеризувати відповідність певного документа інформаційним потребам користувача, тобто власне релевантність окремого документа, на сьогодні розвинені недостатньо. Тому виникає потреба в розвитку теоретичних моделей, на основі яких можна було б більш формалізовано описувати пошук та пов'язані з ним процеси.

Очевидною є необхідність максимальної орієнтації пошуку на використання семантики та онтології предметної області [2]. Особливо природною є така зорієнтованість для тематичних порталів, на базі яких розвиваються віртуальні співтовариства користувачів. Серед магістральних напрямів розвитку таких онтологічно орієнтованих порталів, інтегрованих із віртуальними спільнотами, слід відзначити:

1. Підвищення інформаційної зв'язності порталу для швидкого переходу від поточного ресурсу до матеріалів, пов'язаних із ним за певним критерієм. Критично важливим тут є те, що перелік таких зв'язків повинен не бути статичним, а формуватися динамічно - залежно від поточної ситуації і на основі семантики предметної області та відповідних онтологій; деякі принципи побудови таких порталів на основі онтологій предметної області обговорювалися в [4; 5].

2. Формування віртуальної спільноти на основі персональних робочих середовищ окремих користувачів (які можна охарактеризувати як персональні робочі кабінети). Такі середовища мають бути максимально адаптованими до індивідуальних потреб; поєднувати в собі засо-

би керування особистими документами, планування завдань тощо.

3. Відповідно до цього формалізована модель віртуальної спільноти має описувати, як мінімум, такі компоненти [6; 7]: організацію окремого робочого кабінету; зв'язки між окремими персональними середовищами та можливу взаємодію між ними; вплив загальної інфраструктури порталу на створення, наповнення, функціонування та підтримку персонального робочого середовища.

4. Ефективний інформаційний пошук має залучати до розгляду не просто міри релевантності документа запиту, а індивідуальні міри релевантності для кожного користувача з урахуванням завдань, які перед ним стоять, та проблем, які його дійсно цікавлять. Деякі підходи до організації такого пошуку обговорювалися зокрема в [6; 7].

5. У цьому контексті видається доцільним зближення підходів, у рамках яких можна описати власне процес пошуку, з одного боку, та формування віртуальної спільноти, з іншого.

Тому в роботі розглядається підхід, у рамках якого будується деяка формалізована модель інформаційного наповнення порталу. На базі цієї моделі пропонується розглядати процеси самоорганізації порталу і пов'язану з ним динамічну побудову мір релевантності. Ефективні методи пошуку можуть бути запропоновані саме на основі цього процесу.

Модель інформаційного наповнення порталу є багатокомпонентною. Основою слугує формалізована модель «онтологія - артефакт», у яку «занурюються» інші категорії сутностей. Модель «онтологія - артефакт» описана в [5; 7] як трійка $M = \langle W^*, D, L \rangle$, де W - онтологія предметної області, W^* - розширена онтологія, наповнення онтології W конкретними екземпляра-

ми класів (фактично - база знань), D - множина документів; L - множина зв'язків між W^* та D . Власне онтологія описується як трійка $\langle Q, R, F \rangle$, 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 $\langle Q^*, R^*, F^* \rangle$, 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^* . Тобто документ d вважається релевантним відносно W^* , якщо існують хоча б один вузол w та функція інтерпретації f , такі що $d = f(w)$. Тоді міри релевантності будуть безпосередньо пов'язані з ваговими коефіцієнтами зв'язків, пов'язаних із функціями інтерпретації.

У цьому випадку ідею «занурення» інших категорій сутностей можна в загальних рисах сформулювати 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. Формально сутності, які занурюються в цю модель, можуть бути довільними. Оскільки мова йде про тематичний портал, інтегрований з віртуальною спільнотою, природно розглядати сутності, пов'язані з користувачами та задачами, які ними розв'язуються; окремі задачі об'єднуються в цілі проекти.

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.

Ця модель може бути застосована до розрахунку міри релевантності на основі хвильового процесу поширення активації між відповідними вузлами [8]. У [9] відзначається, що методика хвильового пошуку може виявитися досить перспективною саме для онтологічно орієнтованих тематичних порталів - порталів знань, для яких характерною є тематична однорідність і достатньо висока зв'язність інформаційних ресурсів.

У нашій моделі активація вузлів може здійснюватися на чотирьох множинах:

- на множині документів, при цьому міри спорідненості можуть задаватися явним чином або визначатися динамічно (наприклад, на основі просторово-векторної моделі [1; 10]);
- на графі, який задає розширену онтологію;
- на множині користувачів;
- на множині окремих завдань і цілих проектів.

Важливими є не тільки пошук на уже сформованій моделі, але й самоорганізація, формування окремих персональних середовищ та віртуальних спільнот. Принципи, які можуть лягти в основу формування коефіцієнтів зв'язків і мір активації в ході такої самоорганізації, багато в чому аналогічні загальним принципам ройового інтелекту, зокрема - алгоритму мурашки [11; 12].

Типовим результатом пошуку, який здійснюється на основі вищенаведених принципів, має стати формування множини понять і документів, тією чи іншою мірою релевантних запиту, з урахуванням формування вагових коефіцієнтів, пов'язаних із функціями інтерпретації. Кожний елемент цієї множини буде поданий у вигляді $(u, m_1(u), m_2(u), \dots)$, де u - знайдений вузол, $m_i(u)$ - міра релевантності цього вузла, обчислена за i -м критерієм, можливо, недостовірна або нечітка [7]. Припускається навіть динамічне породження нових критеріїв, якщо задати процедуру такого породження.

Основна проблема, яка виникає у процесі такого пошуку, пов'язана зі значними обсягами інформації і, відповідно, - зі значною часовою складністю. Тому ключовим є таке питання: як спрямувати процес поширення активації в потрібному напрямку та яким має бути критерій зупинки цього процесу? Важливою є також проблема комбінування критеріїв, яка полягає в переході від кількох мір релевантності документа за різними критеріями до однієї комбінованої міри релевантності.

Для розв'язання перелічених проблем видається доцільним застосовувати методики випадкового керування інформаційним пошуком [13]; зокрема генетичні алгоритми, які добре зарекомендували себе для розв'язання ряду перебірних задач [14; 15]. У контексті, який розглядається, можна виділити як мінімум два аспекти застосування цих алгоритмів:

- власне для вибору найбільш перспективної підмножини документів, серед яких документи, потрібні користувачеві для вирішення його конкретної задачі, будуть міститися з максимальною ймовірністю;
- для експериментального добору параметрів хвильового процесу поширення активації.

Корисна інформація, яка накопичується у процесі пошуку, має максимально повно враховуватися, тобто стратегія пошуку повинна динамічно коригуватися. Одна з можливих стратегій такого пошуку полягає в наступному [13]: множина можливих рішень Ω розбивається на підмножини $\Omega_1, \dots, \Omega_n$. Задача полягає у виборі найбільш перспективної підмножини Ω^* , на якій має здійснюватися подальший пошук. До уваги повинні братися:

- коефіцієнт звуження області пошуку $h(\Omega^*) = |\Omega^*|/|\Omega|$;
- оцінка $p(\Omega^*, I)$ перспективності підмножини Ω^* , ця оцінка залежить, крім самої підмножини Ω^* , від наявної інформації I ; остання, своєю чергою, складається з апіорної оцінки перед розбиттям та результатів апостеріорної перевірки, в якій з підмножин знаходиться потрібне рішення.

Тоді проблема вибору найбільш перспективної підмножини Ω^* може бути математично сформульована як задача мінімізації критерію

$$E(h(\Omega)p(\Omega) + (1 - h(\Omega))(1 - p(\Omega))), \quad (1)$$

де E - символ математичного очікування.

Слід зазначити, що ідеологія роботи генетичних алгоритмів тісно пов'язана з формулою (1). Дійсно, схеми можливих рішень генетичного алгоритму фактично задають розбиття на підмножини, а накопичення інформації здійснюється на основі того, що згідно з теоремою про схеми [14] відсоток найбільш пристосованих схем

(тобто критеріїв вибору найбільш перспективних підмножин) у наступних поколіннях зростає.

У рамках цього підходу можна розв'язувати також задачі, характерні для інтелектуального аналізу даних (Data Mining), наприклад [6; 7]:

- пошук потенційних партнерів на основі аналізу мір схожості між персональними роботами середовищами;
- динамічне формування груп користувачів, які мають спільні або схожі інтереси;
- прийняття рішень за аналогією (наприклад, якщо користувач A для розв'язання задачі C вважає корисним документ W , то користувачеві X , характеристики якого схожі на характеристики користувача A , для розв'язання задачі K , схожої на C , можна порекомендувати список документів, схожих на W);
- автоматичне визначення відповідних функцій інтерпретації формальної моделі та пов'язаних із ними вагових коефіцієнтів на основі документів, відібраних користувачем.

1. Ландэ Д. В. Поиск знаний в Интернет. - М.: Изд. дом «Вильямс», 2005. - 272 с.
2. Гаврилова Т. А., Хорошевский В. Ф. Базы знаний интеллектуальных систем. - СПб.: Питер, 2000. - 384 с.
3. Плескач В. Л., Рогущина Ю. В. Агентні технології. - К.: Київ. нац. торг.-екон. ун-т, 2005. - 338 с.
4. Олецкий О. В. Застосування формальних моделей онтологій для формалізації інформаційних потоків у системах управління контентом // Теоретичні та прикладні аспекти побудови програмних систем. Матеріали міжнародної конференції TAAPSD'2005. Київ, 7-9 грудня 2005 р.
5. Діренко І. С., Олецкий О. В. Система управління вмістом веб-ресурсів на основі онтологічно-документного моделювання // Теоретичні та прикладні аспекти побудови програмних систем. Матеріали міжнародної конференції TAAPSD'2006. Київ, грудень 2006 р. - С. 171-176.
6. Олецкий О. В. До проблеми організації онтологічно-орієнтованих персональних робочих середовищ в рамках віртуального співтовариства // Системи підтримки прийняття рішень. Теорія і практика. Матеріали 4-ї дистанційної науково-практичної конференції з міжнародною участю СППР'2008. Київ, червень 2008. - С. 81-83. - [www: http://conf.atsukr.org.ua/files/conf_dir_8/oletsky_sppr2008.pdf](http://conf.atsukr.org.ua/files/conf_dir_8/oletsky_sppr2008.pdf).
7. Олецкий О. В. Принципи побудови формалізованих онтологічно-орієнтованих моделей організації віртуальних спільнот у рамках веб-порталу // Теоретичні та прикладні аспекти побудови програмних систем. Матеріали міжнародної конференції TAAPSD'2008. Київ - Чернігів, 22-26 вересня 2008 р. - С. 209-214.
8. Глибовець М. М. Моделі та методи створення і супроводу високопродуктивного розподіленого навчального середовища: Автореф. дис. ... д-ра фіз.-мат. наук / Нац. ун-т «Києво-Могилянська академія». - К., 2006.
9. Олецкий О. В. До проблеми онтологічно-орієнтованого пошуку в інформаційних системах // Теоретичні та прикладні аспекти побудови програмних систем. Матеріали міжнародної конференції TAAPSD'2007. Бердянськ, 4-9 вересня 2007 р. - С. 73-77.
10. Сэлтон Дж. Автоматическая обработка, хранение и поиск информации. - М.: Сов. радио, 1973. - 560 с.
11. Тарасов В. Б. От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. - М.: Эдиториал УРСС, 2002. - 352 с.
12. Джонс М. Т. Программирование искусственного интеллекта в приложениях. - М.: ДМК Пресс, 2004. - 312 с.
13. Глибовець М. М., Олецкий О. В. Про деякі підходи до проблеми інформаційного керування випадковим пошуком // Dynamical System Modelling and Stability Investigation. Thesis of Conference Reports, May 22-25, 2007. - С. 370.
14. Рутковская Д., Пилинский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткая логика. - М.: Горячая линия - Телеком, 2004. - 452 с.
15. Глибовець Н. Н., Медведь С. А. Генетические алгоритмы и их использование для решения задач составления расписания // Кибернетика и системный анализ. - 2003. - № 1. - С. 95-108.

O. Oletsky

ONTOLOGIC INFORMATION SEARCH ON THE BASE OF WAVE PROCESS OF ACTIVATION PROPAGATION

An approach to treating the problem of information search as a wave process of activation propagation based on a certain model of information system is regarded. For information model of a subject portal integrated with virtual communities, the model «ontology - artifact - user - project» is proposed. Applying of genetic algorithms and swarm intelligence is discussed.