УДК 681.3:658.5

О.В. Олецький

ПРИНЦИПИ АНАЛІЗУ МАТЕМАТИЧНИХ МОДЕЛЕЙ ФОРМУВАННЯ МІР ПОДІБНОСТІ ТА РЕЛЕВАНТНОСТІ В РАМКАХ ТЕМАТИЧНОГО ВЕБ-ПОРТАЛУ

Пропонується комбінований підхід до формування онтологічно-орієнтованих мір подібності документів та їх відповідності запитам користувача в рамках тематичного веб-порталу. Підхід ґрунтується на організації хвильового процесу поширення активації на формальній моделі інформаційного наповнення порталу, яка описується як четвірка “онтологія-артефакт-користувач-проект”. Обговорюється зв’язок цього підходу з класичними мірами, зокрема з векторно-просторовою моделлю та з пошуковим процесом, який лежить в основі PageRank.

Вступ. Пошук інформації, необхідної для розв’язання тих чи інших задач користувача, займає центральне місце в інформаційних системах, зокрема веб-орієнтованих. Це стосується як спеціалізованих тематичних порталів, так і пошукових систем загального призначення. Загальновідомо, що сучасні методи інформаційного пошуку є недосконалими, і проблема підвищення якості та ефективності пошуку є дуже гострою та актуальною. Сьогодні інтенсивно розвиваються різні підходи до її вирішення [1-3 та ін.], але вони здебільшого носять евристичний характер. Навіть такі базові поняття, як ”релевантність документа запитові”, “схожість документів”, формалізовані недостатньо. Тому виникає потреба в розвитку теоретичних моделей, на основі яких можна було б більш формалізовано описувати пошук та пов’язані з ним процеси. Центральне місце тут мають займати підходи, які б дозволяли будувати достатньо обґрунтовані та інтегровані кількісні міри подібності та релевантності. При цьому такі міри мають максимально враховувати семантику та онтологію предметної області.

Серед найбільш відомих підходів до формування та аналізу мір подібності документів та їх релевантності запитові слід відмітити наступні:

1. Булева та векторно-просторова моделі пошуку [1, 4], які широко використовуються в сучасних пошукових системах. Але матрицю “документ-термін”, яка лежить в основі класичної векторно-просторової моделі, по своїй суті 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 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}.

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 $\frac{|R_a \cap R_b|}{|R_a \cup R_b|}$.

3. Очевидним розвитком цього підходу стає врахування вагових коефіцієнтів, пов'язаних з тим чи іншим типом зв'язків. Відповідна методика інтенсивно розвивається, зокрема в агентних системах; може знаходити ефективне застосування в системах електронної комерції [3, 5 та ін.].

4. Латентно-семантичний аналіз, спрямований на виявлення прихованих закономірностей в інформаційних масивах [1, 6, 7], та тісно пов'язаний з аналізом головних компонент та перетворенням Карунена-Лоева [8-11 та ін.].

5. Статистично-ймовірносні методики аналізу текстів та виявлення їх відповідності потребам користувачів [1, 12 та ін.].

6. Методики аналізу текстів на основі нейронних мереж [13, 14 та ін.].

7. Методики оцінки авторитетності джерел на основі аналізу графу гіпертекстових посилань; йдеться перш за все про алгоритми PageRank та HITS [1 та ін.].

Викладення основного матеріалу. Складність та недостатня формалізованість проблеми формування мір подібності та релевантності понять і документів та ранжування документів за цими мірами призводить до необхідності розвитку певних комбінованих підходів, які б дозволяли враховувати та інтегрувати різні методики. В роботі пропонується такий комбінований підхід, в основі якого лежить певний процес динамічного підрахунку мір релевантності на основі певної формалізованої моделі веб-порталу. Ця модель має максимально враховувати онтологію предметної області та залучати до розгляду такі основні компоненти:

- основні категорії: зокрема, онтологія предметної області, множина артефактів, множина користувачів, множина понять;
- зв'язки між цими категоріями;
- процедури та механізми, які забезпечують підрахунок мір близькості.

Модель інформаційного наповнення веб-порталу будується на основі формальної моделі онтології [2] з урахуванням принципів, сформульованих в [15, 16]. Основою служить формалізована модель "онтологія-артефакт", в яку "занурюються" інші категорії сутностей. Модель "онтологія-артефакт" описується 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** x *R** x *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 з функціями інтерпретації.

В 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. Оскільки мова йде про пошук на тематичному порталі, природно розглядати сутності, пов'язані з користувачами та задачами, які ними розв'язуються.

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 одного боку, з користувачами та проектами і роботами, в яких вони беруть участь – з іншого. Якщо повернутися до описаної вище узагальненої векторно-просторової моделі, то можна розглядати такий підхід: координатами простору ознак виступають множини понять предметної області та пов'язані з ними артефакти, а коефіцієнти q_{ij} визначаються не тільки зв'язками власне між цими компонентами, але залежать також від користувачів та задач.

Процес динамічного розрахунку мір релевантності може ґрунтуватися на ідеї хвильового процесу поширення активації між відповідними вузлами [17]. Така 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 [18].

Базовий підхід полягає в наступному. Початковий запит призводить до початкової активації певного набору вузлів моделі. Далі активуються пов'язані з ними вузли і т.д.; при цьому коефіцієнти відповідності окремих вузлів запиту мають динамічно перераховуватися. В моделі "онтологія-артефакт-користувач-проект" активація вузлів може здійснюватися на чотирьох множинах:

- на множині документів;
- на графі, який задає розширену онтологію;
- на множині користувачів;
- на множині окремих завдань і цілих проектів.

Типовим результатом пошуку, який здійснюється на основі вищенаведених принципів, має стати формування множини понять і документів, в тій чи іншій мірі релевантних запиту, з урахуванням вагових коефіцієнтів, пов'язаних з функціями інтерпретації. Кожний елемент цієї множини буде поданий у вигляді *(u, m_1(u), m_2(u), ...),* де *u* – знайдений вузол, $m_i(u)$ - міра релевантності цього вузла, обчислена за *i*-м критерієм, можливо, недостовірна або нечітка. Важливою є проблема комбінування критеріїв, яка полягає в переході від кількох мір релевантності документу за різними критеріями до однієї комбінованої міри релевантності.

Важливими є не тільки пошук на уже сформованій моделі, але й самоорганізація, формування такої моделі. 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 [19, 20].

В цьому контексті слід звернути увагу на те, що 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. Очевидно, що якщо певним чином змінити схему і параметри цього процесу, то в рамках загальної комбінованої методики поширення активації можна отримати ціле сімейство подібних мір. 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.

Основна проблема, яка виникає в процесі пошуку, пов'язана зі значними обсягами інформації, і відповідно – зі значною часовою складністю. Тому ключовим є наступне питання: як спрямувати процес поширення активації в потрібному напрямку та яким повинен бути критерій зупинки цього процесу?

Для розв'язання перелічених проблем видається доцільним застосовувати методики випадкового керування інформаційним пошуком [21]; зокрема генетичні алгоритми, які добре зарекомендували себе для розв'язання ряду перебірних задач [22, 23]. В 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 підбору параметрів хвильового процесу поширення активації.

Корисна інформація, яка накопичується в процесі пошуку, має максимально повно враховуватися, тобто стратегія пошуку повинна динамічно коригуватися. Одна з можливих стратегій такого пошуку полягає в наступному [21]: множина можливих рішень W розбивається на підмножини

$W_1, \ldots, W_n$. Задача полягає у виборі найбільш перспективної підмножини W^*, на якій має здійснюватися подальший пошук. До уваги повинні братися:

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

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

$$E(h(W`)p(W`)+(1-h(W`))(1-p(W`))),$$

де E – символ математичного очікування.

Висновки. У роботі наведені деякі підходи до побудови формалізованих моделей, на основі яких можна формувати більш обґрунтовані міри подібності понять та документів. Уточнення та розвиток цих підходів вимагає подальших досліджень, результати яких будуть опубліковані у подальших роботах автора.

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., Рогушина Ю.В.* Агентні технології. - К.: Київ. нац. торг.-екон. ун-т, 2005. – 338 с.
4. *Сэлтон Дж.* Автоматическая обработка, хранение и поиск информации. – М.:Сов.радио, 1973. – 560 с.
5. *Плескач В.Л., Рогушина Ю.В.* Перспективы онтологического анализа как составляющей интеллектуального поиска на рынке информационных услуг. //Теоретичні та прикладні аспекти побудови програмних систем. Матеріали міжнародної конференції TAAPSD'2006. Київ, грудень 2006 р. – С. 42–45.
6. *Berry M.W., Dumais S.T., O'Brein G.W.* Using linear algebra intelligent intelligent information retrieval. //SIAM Review. – 1995. – 37(4). – P. 573–595.
7. *Дерецький В.О.* Використання методу латентного семантичного індексування для автоматичної побудови тематичної онтології документа. //Теоретичні та прикладні аспекти побудови програмних систем. Матеріали міжнародної конференції TAAPSD'2005. Київ, грудень 2005 р. – С.73–77.
8. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.
9. *Фукунага К.* Введение в статистическую теорию распознавания образов. - М.: Наука, 1979. - 368 с.
10. *Олецкий А.В.* О применении интегрального разложения Карунена - Лоэва при моделировании динамических систем. // УСиМ, 1999, №2. - С.12-15.
11. *Олецкий А.В.* Основные свойства и практические применения базовой квадратурной схемы интегрального разложения Карунена-Лоэва. // Моделювання та інформаційні технології. Вип.27. – Київ, 2004. – С. 113–120.
12. *Люгер Дж.Ф.* Искусственный интеллект: стратегии и методы решения сложных проблем. - М.: Изд. дом "Вильямс", 2003. – 864 с.

13. *Каллан Р.* Основные концепции нейронных сетей. - М.: Изд. дом "Вильямс", 2001. – 288 с.
14. *Головко В.А.* Нейронные сети: обучение, организация и применение. – М.: ИПРЖР, 2001. – 256 с.
15. *Олецький 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.
16. *Діренко І.С., Олецький 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.
17. *Глибовець М.М.* Моделі та методи створення і супроводу високопродуктивного розподіленого навчального середовища. Автореферат дисертації на здобуття наукового ступеня доктора фізико-математичних наук. – Національний університет "Києво-Могилянська Академія", Київ, 2006.
18. *Олецький 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.
19. *Тарасов В.Б.* От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. – М.: Эдиториал УРСС, 2002. – 352 с.
20. *Джонс М.Т.* Программирование искусственного интеллекта в приложениях. – М.: ДМК Пресс, 2004. – 312 с.
21. *Глибовець М.М., Олецький 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.
22. *Рутковская Д., Пилиньский М., Рутковский Л.* Нейронные сети, генетические алгоритмы и нечеткая логика. – М.: Горячая линия - Телеком, 2004. – 452 с.
23. *Глибовец Н.Н., Медведь С.А.* Генетические алгоритмы и их использование для решения задач составления расписания//Кибернетика и системный анализ, 2003, №1. – С.95–108.