УДК 681.3:658.5

Олецький О. В.

ПРИНЦИПИ ЗАСТОСУВАННЯ ГЕНЕТИЧНИХ АЛГОРИТМІВ ДО ЗАДАЧІ ОНТОЛОГІЧНО ОРІЄНТОВАНОГО ІНФОРМАЦІЙНОГО ПОШУКУ

У статті сформульовано рекомендації щодо застосування генетичних алгоритмів для проблеми інформаційного пошуку на базі онтологічного аналізу. Подано класифікацію параметрів онтологічної моделі інформаційного наповнення, які необхідно оптимізувати; запропоновані співвідношення, оптимізація параметрів яких має здійснюватися в першу чергу. Розглянуто основні підходи до вибору цільової функції.

Ключові слова: генетичий алгоритм, інформаційний пошук на базі онтологічного аналізу.

Вступ

У роботах [1-4] розвивається підхід, спрямований на розробку ефективних методик онтологічно орієнтованого локального пошуку інформаційних матеріалів, які б найкраще відповідали цілям відвідувачів з урахуванням їхніх індивідуальних характеристик. Мова йде насамперед про веб-ресурси, для яких характерні висока інформаційна зв'язність, тематична однорідність, достатньо висока структурованість та якість інформаційного наповнення. До таких ресурсів належать, зокрема, тематичні портали, які можуть набувати рис навчальних та науково-дослідницьких ресурсів, або віртуальних співтовариств, які можуть розвиватися на їхній основі.

В основі запропонованої методики лежить побудова формальної моделі інформаційного наповнення системи і розгляд процесу пошуку як певного хвильового процесу на цій моделі. Основними елементами моделі стають побудова певного графу, вузлами якого є поняття предметної області та артефакти інформаційної системи, а дуги відповідають зв'язкам між ними, і «занурення» до базової моделі «онтологія-артефакт» інших категорій сутностей, зокрема – інформації про відвідувачів та про їхні цілі. Ідея такого «занурення» полягає в тому, що вагові коефіцієнти зв'язків між вузлами онтології та артефактами інформаційної системи стають залежними від індивідуальних характеристик відвідувачів та від цілей, які перед ними стоять. Альтернативний погляд на проблему може полягати у побудові багатокомпонентної онтологічної системи з такими основними компонентами [5]: онтологія предметної області; онтологія артефактів; онтологія функцій інтерпретації, які пов'язують вузли онтології предметної області з вузлами онтології артефактів; онтологія цілей; онтологія ресурсів.

При цьому ключового значення набувають методи оцінювання важливості та релевантності вузлів модельного графа та зв'язків між ними, а також вибір оптимальних алгоритмів пошуку. Для класу ресурсів, про який ідеться, це можна здійснити шляхом «ручного» експертного оцінювання, але розвиток автоматизованих методик здійснення такого оцінювання теж має велике значення. Проблема ускладнюється ще й тим, що самі критерії якості інформаційного пошуку (наприклад, [6-8 та ін.]) сформульовано недостатньо формалізовано і вимагають суттєвих уточнень.

Для вирішення задач оптимізації та самонавчання, пошуку оптимальних параметрів моделі добре зарекомендували себе генетичні алгоритми [9-12 та ін.]. В основі генетичного алгоритму лежить імітація еволюційних процесів, зокрема, репродукція – створення нових поколінь та селекція – відбір найкраще пристосованих рішень.

Важливо дослідити можливість застосування генетичних алгоритмів до задачі онтологічно-орієнтованого інформаційного пошуку, саме це є основним завданням нашої роботи.

Основна частина роботи

Для застосування генетичних алгоритмів до задачі інформаційного пошуку необхідно визначити:

– множину параметрів задачі *P:* параметри інформаційної моделі P_M, які впливають на результат розв'язку цієї задачі; очевидно, що для задачі онтологічно орієнтованого пошуку найбільш суттєве значення можуть мати міри

© *Олецький О. В.*, 2010

близькості між основними компонентами моделі інформаційного наповнення (насамперед між вузлами онтології та документами); комбіновані критерії, які об'єднують різні міри, а також множина параметрів P_s власне процесу пошуку;
– множину параметрів P_{GA}, які впливають на результати роботи власне генетичного алгоритму;
– критерій оптимізації (цільову функцію).

Для підвищення гнучкості системи та для проведення необхідних експериментальних досліджень вважаємо доцільним провести ще одну класифікацію множини параметрів P:
– множина параметрів P_R, які задаються в ручному режимі;
– множина параметрів P_A, які визначаються генетичним алгоритмом.

Очевидно, що $P_R \cup P_A = P$; $P_R \cap P_A = \varnothing$.

Для побудови власне мір близькості між вузлами моделі можна застосовувати ряд відомих підходів [3, 4, 6, 7, 8 та ін.]:

1. Булева та векторно-просторова моделі пошуку, які широко використовуються в сучасних пошукових системах. Але матрицю «документ-термін», яка лежить в основі класичної векторно-просторової моделі, природно розглядати як окремий випадок матриці даних у певному просторі ознак, широко відомої в математичній статистиці та у розпізнаванні образів. Дійсно, така матриця даних може мати вигляд $Q = \{q_{ij}\}$, де 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}. Зокрема, онтологічний аналіз може бути врахований при використанні й 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

$$V = HC, \tag{1}$$

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 кожна 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

$$W = (\Sigma_i\, v_i O_i)\; V\; (\Sigma_j\, \tau_j D_j) \tag{2}$$

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.

Можна розглядати комбіновані критерії. Зокрема, в [3, 4] розглядається деяка комбінована міра релевантності документа d поняттю w, усереднену за всіма зв'язками з урахуванням їхніх вагових коефіцієнтів:

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

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

Звичайно, наведені міри не є єдино можливими, і описаний підхід вимагає подальшого розвитку. Зокрема, формули (2) і (3) можуть бути інтегровані в один комбінований критерій.

Ще один можливий тип співвідношень, який розглядався в [3, 4], пов'язаний із міркуванням, що міри важливості подібних вузлів теж можуть бути близькими. Якщо, як і раніше, розглядати суміжність за різними типами зв'язків, то можна записати наступну систему співвідношень відносно мір важливостей вузлів:

$$m(u) = \sum_{v \in G(u)} \beta_{uv}\, m(v) \tag{4}$$

Тут $m(u)$ – міра важливості вузла u; $G(u)$ – множина вузлів, суміжних з u, β_{uv} – коефіцієнти, які характеризують міру важливості відповідного зв'язку.

Вибір критерію оптимізації (цільової функції) є важливою проблемою, яка на сьогодні досліджена недостатньо. Мова може йде насамперед про оптимізацію пошукової видачі, модель якої ми розглядаємо як послідовність пар $((d_1, \rho(d_1)), (d_2, \rho(d_2)), \ldots)$, упорядковану за спаданням

$\rho(d_i)$. Тут $\rho_i = \rho(d_i)$ – міра відповідності документа запитові. Часто нас цікавлять не точні чисельні оцінки, а лише ранг документа в пошуковій видачі. Тоді природно вводити певне відношення переваги між документами та поставити вимогу: якщо користувач оцінює документ d_1 вище, ніж d_2, а ранг d_1 у пошуковій видачі, навпаки, є нижчим за ранг d_2, то така ситуація розглядається як помилка, і метою генетичного алгоритму має стати мінімізація кількості таких помилок. Критичного значення в цьому контексті набуває врахування думок користувача – якою мірою результати пошуку задовольняють відвідувача з урахуванням рівня його компетентності. Деякі інші міркування з приводу можливих підходів до оптимізації пошукової видачі можна знайти, зокрема, в [7, 8].

Зрештою, важливе значення має оптимізація елементів множини P_{GA} – параметрів власне генетичного алгоритму. До таких параметрів належать насамперед наступні:

– розмір популяції;
– кількість найкращих особин, які переходять до наступної популяції автоматично (якщо реалізується елітарна стратегія);
– ймовірність мутації на кожній ітерації;
– кількість ітерацій.

Таким чином, можна говорити про певний дворівневий процес оптимізації – на першому рівні налаштовуються параметри генетичного алгоритму, а на другому – параметри задачі.

У роботах [13, 2, 3] розглядається підхід, спрямований на підвищення цілеспрямованості пошуку на основі генетичних алгоритмів. 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 власне 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

$$E(h(W`)p(W`)+(1-h(W`))(1-p(W`))), \qquad (5)$$

де E – символ математичного очікування.

Ця методика в загалом пов'язана із значною часовою складністю. Але генетичні алгоритми по суті неявно реалізують саме цю методику. Дійсно, схеми можливих рішень фактично задають розбиття на підмножини, а накопичення інформації здійснюється на основі того, що згідно з теоремою про схеми [9], відсоток найбільш пристосованих схем (тобто критеріїв вибору найбільш перспективних підмножин) у наступних поколіннях збільшується.

Таким чином, можна стверджувати, що подальше дослідження генетичних алгоритмів на основі критерію (5) дозволить точніше і формалізованіше описати принципи їхньої роботи та зробити їх більш математично обґрунтованими, цілеспрямованими і керованими [13].

Висновки

У роботі досліджено можливість застосування генетичних алгоритмів для автоматизованого налаштування параметрів онтологічної моделі інформаційного наповнення тематичного порталу з метою підвищення релевантності пошуку. Крім тематичних порталів, подібна методика можна застосувати для ресурсів навчального та наукового спрямування та для віртуальних співтовариств, які можуть розвиватися на їхній основі. Насамперед запропоновано здійснювати оптимізацію параметрів співвідношень (1) – (4), наведених у роботі. Розглянуто можливість для організації дворівневої схеми оптимізації, на першому рівні якої налаштовуються параметри власне генетичного алгоритму, а на другому – власне параметри наведених співвідношень. Суттєвою проблемою при цьому залишається вибір критерію оптимізації пошукової видачі. В роботі наведено деякі базові рекомендації до вирішення цієї проблеми, але їхнє уточнення та розвиток є предметом подальших досліджень.

1. Олецький О. В. Онтологічно орієнтований інформаційний пошук на основі хвильового процесу поширення активації / Олецький О. В. // Наукові записки НаУКМА. Т. 86 : Комп'ютерні науки. – К., 2008. – С. 50–52.
2. Олецький О. В. Принципи аналізу математичних моделей формування мір подібності та релевантності в рамках тематичного веб-порталу // Збірник наукових праць Інституту проблем моделювання в енергетиці ім. Г. Є. Пухова. Вип. 52. – К., 2009. – С. 37–42.
3. Олецький О. В. Про деякі підходи до побудови онтологічно-орієнтованих моделей пошуку і самоорганізації для тематичних порталів // Збірник наукових праць Інституту проблем моделювання в енергетиці ім. Г. Є. Пухова. Вип. 53. – К., 2009. – С. 35–39.
4. Олецький О. В. Організація онтологічно-орієнтованих засобів автоматизованого експертного підбору інформаційних ресурсів на тематичному порталі // Наукові записки НаУКМА. Т.99. Комп'ютерні науки. – К., 2009. – С.66–69.

5. Олецький О. В. Побудова багатокомпонентної онтологічної системи для автоматизованого експертного добору навчальних матеріалів // Теоретичні та прикладні аспекти побудови програмних систем. Матеріали міжнародної конференції TAAPSD'2009, 8–10 грудня 2009 р. – С. 83–88.
6. Ландэ Д. В. Поиск знаний в Internet / Д. В. Ландэ. – М. : Вильямс, 2005. – 272 с. – ISBN 5-8459-0764-0.
7. Сэлтон Г. Автоматическая обработка, хранение и поиск информации / Г. Сэлтон. – М. : Сов. радио, 1973. – 560 с.
8. Плескач В. Л. Агентні технології / Плескач В. Л., Рогушина Ю. В. – К. : Київ. нац. торг.-екон. ун-т, 2005. – 338 с.
9. Рутковская Д. Нейронные сети, генетические алгоритмы и нечеткая логика / Д. Рутковская, М. Пилиньский, Л. Рутковский. – М. : Горячая линия – Телеком, 2004. – 452 с. – ISBN 5-93517-103-1.
10. Зайченко Ю. П. Основи проектування інтелектуальних систем / Ю. П. Зайченко. – К. : Слово, 2004. – 352 с.
11. Николенко С. И. Самообучающиеся системы / С. И. Николенко, А. Л. Тулупьев. – М. : МЦНМО, 2009. – 288 с. – ISBN 978-5-94057-506-1.
12. Глибовец Н. Н. Генетические алгоритмы и их использование для решения задач составления расписания / Глибовец Н. Н., Медведь С. А. // Кибернетика и системный анализ. – 2003. – №1. – С. 95–108.
13. Глибовець М. М. Про деякі підходи до проблеми інформаційного керування випадковим пошуком / Глибовець М. М., Олецький О. В. // Dynamical System Modelling and Stability Investigation. Thesis of Conference Reports, May 22–25, 2007. – С. 370.

O. Oletsky

PRINCIPLES OF USING GENETIC ALGORITHMS FOR ONTOLOGY ORIENTED INFORMATION SEARCH

Recommendations for using genetic algorithms for ontology-based information search are formulated. Parameters of the ontology-based model of content are classified; relations, parameters of which should be optimized, are proposed. Some approaches for choosing optimization criterion are regarded.

Keywords: Genetic algorithm, information search, ontology analysis.

УДК 004.415.2

Парфірова Т. С.

ПОНЯТТЯ ВЗАЄМОДІЇ З ТОЧКИ ЗОРУ СУТНІСНОЇ ПЛАТФОРМИ

Проведено розгортання поняття взаємодії із застосуванням методології сутнісної платформи. Отримано модель взаємодії, достатньо загальну для охоплення конкретних типів взаємодій. Загальність моделі підтримується тим, що вона містить механізми для саморозгортання і конкретизації.

Ключові слова: сутнісна платформа, взаємодія, активна сутність, пасивна сутність, результат взаємодії, прикладні засоби.

У роботах В. Н. Редько та його наукової школи [1–3] сформульовано основи сутнісної платформи (СП), які стали продовженням попередніх робіт, направлених на побудову концептуально єдиної теорії програмування.

Нагадаємо основні поняття СП. Сутність – все, що існує, в найзагальнішому сенсі цього слова. Деякі сутності виступають сутями інших сутностей, тобто є їхніми прагматико-обумовленими властивостями. Під властивістю розуміють сутність, яка характеризує сутність. Поняття сутесутнісного відношення відіграє виключно важливу роль у побудові сутнісної платформи.

СП носить характер загальної теорії, яка, з одного боку, описує все, а з другого, її повнота носить не актуальний, а потенційний характер. Тобто яку б область дійсності не взяти, хоча СП не дає її теоретичної моделі в готовому вигляді, проте можна провести розгортання СП до адекватної теорії, яка описує цю область. Будь-яке розгортання в СП повинно відповідати принципу достатніх основ, введення абстракції може

© *Парфірова Т. С.*, 2010