УДК 681.3

АГЕНТНИЙ ПІДХІД ДО ТЕМАТИЧНОГО ПОШУКУ ІНФОРМАЦІЇ З ВИКОРИСТАННЯМ ОНТОЛОГІЙ

С. Ремарович

Інститут програмних систем НАН України,
03187, Київ-187, проспект Академіка Глушкова, 40.
Тел.: +38 044 526 6249; rem@isofts.kiev.ua .

Розвиток глобальної мережі, ріст обсягу тематико-орієнтованої інформації в різноманітних предметних областях, збільшення числа користувачів Internet зумовили велику кількість підходів та засобів до пошуку інформації. Розглянуто аналіз централізованих та децентралізованих симстем пошуку. Запропоновано агентний підхід до побудови розподілених пошукових систем. Розглянута функціональність системи e-Content, що забезпечує тематичний пошук інформації у розподілених гетерогенних джерелах.

Development of a global network, growth of volume of the thematic information in various subject domains, increase in number of users Internet, stipulated a plenty of approaches and means to information search. In work the analysis centralized and decentralized search systems are considered. It is offered agent the approach to construction of the distributed search systems. Functionality of the e-Content system, which provides the thematic information retrieval in the distributed heterogeneous sources, is considered.

Вступ

Комп'ютерні інформаційно-пошукові системи займають все більш істотне місце в науці і освіті. Росте потреба в оперативному доступі до наукових матеріалів, використання яких сприяє поліпшенню дослідницького процесу. Все більше уваги приділяється Internet, як широкому довідковому інструменту, а в останні роки, і як середовищу зберігання та опрацювання бізнесової, наукової та інших типів інформації. Інтернет, як джерело інформації, характеризується такими основними особливостями як:

− величезний обсяг інформації;
− динамічність інформації;
− великий обсяг віртуальної інформації;
− велика кількість мов і форматів даних.

На даний момент можна виділити три підходи до організації пошукових машин в мережі: пошукові сервіси, каталоги і відносно новий підхід, заснований на метапошукових технологіях, в основі яких використовуються методи Семантичної мережі [1]. В рамках підходів Семантичної мережі принципи обробки і керування інформацією істотно відрізняються від традиційних. Суть його полягає у розмітці інформації семантичними тегами та використанні семантичної розмітки в пошукових системах. В сучасних підходах робляться спроби, що дозволяють комбінувати існуючі алгоритми і методи з підходами Семантичної мережі. Реалізація пошукових підходів на базі агентного принципу дозволить забезпечити виконання пошукових процедур на великому просторі розподілених джерел інформації.

Системи пошуку за своєю архітектурою можна поділити на централізовані і децентралізовані. На практиці більше використовуються перші з них. Але централізована архітектура має наступні фундаментальні недоліки, які перешкоджають виживанню таких систем в багатоджерельному гетерогенному оточенні, а саме:

− зниження частки проіндексованого Інтернет. Опублікована в різних виданнях статистика показує, що пошукові системи з централізованою архітектурою втрачають свої позиції. Наприклад, найбільша пошукова система Alta Vista індексує найбільшу частку Інтернет зі всіх пошукових систем. Проте частка проіндексованого Інтернет падає з року в рік. Це пов'язано з тим, що обсяг документів, опублікованих в Інтернет, росте значно швидше, ніж ростуть можливості централізованої системи індексування;

− низька якість пошуку. Централізована система прагне охопити інформаційні потреби всіх можливих користувачів, у зв'язку з чим в індекс включаються дані про всі доступні системі документи. В результаті тематика проіндексованих документів змінюється в дуже широких межах. З другого боку, запити користувача, як правило, містять не більше двох ключових слів. У відповідь користувач централізованої системи пошуку одержує величезне число документів, велика частина яких відноситься до категорії «сміття». Система не в змозі виконати якісне ранжирування результатів, і користувач тоне в потоці нерелевантних документів.

Системи з децентралізованою архітектурою [2] займають поки значно менший сектор ринку інформаційного пошуку, але майбутнє належить їм. Основні компоненти системи пошуку з розподіленою архітектурою:

− тематичні індекси. В системі з децентралізованою архітектурою число індексів не обмежене. Кожний індекс покриває певну тематичну область, забезпечуючи і зберігання інформації про документи, і пошук;

− тематичні мережні роботи, мережні агенти. На відміну від мережного робота централізованої системи, скануючого весь Інтернет, тематичний робот орієнтований на певну тематику. Це дозволяє використовувати більш інтелектуальні алгоритми сканування, підвищити повноту представленої в індексі інформації із заданої тематики та зменшити складність пошукового агента;

− брокери, агенти-посередники. В розподіленій системі запит користувача прямує до брокера, задачею якого є оцінка тематичної належності запиту і вибір індексів, в які слід переправити запит для пошуку (задача маршрутизації запиту користувача). Як і тематичні індекси, брокери також можуть належати різним власникам, що конкурують один з одним. Різні брокери можуть спеціалізуватися на пошуку в різних більш менш широких групах тем;

− репозитарій. У зв'язку з відкритістю системи з розподіленою архітектурою має бути можливість власникам нових тематичних індексів реєструвати свої індекси для доступу до них брокерів. У репозитарій заноситься інформація описового, адміністративного характеру про індекс і, найголовніше, опис вмісту індексу у формі, придатній для читання брокером. Саме на основі цієї інформації брокер ухвалює рішення про відповідність того або іншого індексу заданому запиту.

Наведений опис основних компонент системи пошуку з розподіленою архітектурою показує, що ці системи здатні подолати ті недоліки, які були визначені для систем з централізованою архітектурою.

1. Агентний підхід до тематичного пошуку інформації

У пошуковій системі з розподіленою архітектурою різні тематичні індекси можуть належати різним власникам. З одного боку, це може привести до конкуренції і перекриття тематик різних індексів, а з другого боку, незалежність індексів сприяє залученню нових ресурсів у процесі індексації Інтернет, що необхідне для достатньо повного його охоплення. Тематичний індекс формується автоматично за допомогою тематичного мережного робота або інформаційного агента [3, 4]. Від точності його роботи залежить якість самого індексу і вартість його побудови і супроводу.

Інформаційний агент − це програмна компонента, яка здатна самостійно ухвалювати рішення і проводити автономні дії, направлені на досягнення мети, відповідно інтересам користувача, визначати значення релевантності документа і теми.

Агент для автоматичного формування колекції документів – є важливою компонентою системи пошуку з розподіленою архітектурою. Задача агента полягає в поповненні індексу новими посиланнями на документи, які релевантні його тематиці. Відомі два різних підходи до побудови таких агентів.

Перший з них − використання індексів існуючих універсальних пошукових систем. Цей підхід достатньо широко застосовується на практиці і має позитивні і негативні сторони. Позитивні сторони:

− повторне використання раніше отриманих даних. Сканування Інтернет − це дорогий процес, який приводить не тільки до великих витрат, але і зачіпає інтереси багатьох інших сторін власників сайтів, що індексуються, користувачів. Вельми безрозсудно сканувати багато разів Інтернет ради надання доступу до однієї і тієї ж інформації з великого числа конкуруючих пошукових систем. Ідеальною була б ситуація, при якій деяка велика міжнародна організація виконувала регулярне сканування всього Інтернет і надавала безкоштовний доступ до отриманої необробленої інформації всім охочим для подальшої обробки і використання. Проте таке рішення не масштабується, оскільки ресурси цієї гіпотетичної організації, скільки б вони не були великі, завжди обмежені, і їх зростання не може встигати за зростанням обсягу інформації, опублікованої в Інтернет;

− новизна інформації, що використовується. Мережні агенти універсальних пошукових систем прагнуть індексувати всі нові документи, не обмежуючи себе фіксованою і, можливо, вже застарілою тематикою. Індекс таких систем надає представницьку вибірку відносно недавно опублікованих документів, що можна використовувати при аналізі тенденцій в тій або іншій області;

− низька вартість отримання інформації. Фактично на даний момент отримання інформації у вигляді відповідей на запити, що автоматично генеруються, від комерційних пошукових систем безкоштовно.

Негативні сторони:

− старіння індексу;

− закритість методики отримання інформації, що використовується. Алгоритми сканування Інтернет і пошуку в індексі є комерційною таємницею, і, отже, індекс комерційних універсальних систем представляється системам наступного рівня, які його використовують, у вигляді чорного ящика. Не можливо робити якісь об'єктивні висновки про характер розподілу інформації в Інтернет на основі непрямого (через систему пошуку) аналізу індексу комерційної системи;

− ненадійність доступу до інформації. Комерційні пошукові системи очевидно не зацікавлені в тому, щоб їх індекс аналізувався б автоматичними системами. У будь-який момент експерименти в цій області можуть бути заборонений у зв'язку з порушенням тих або інших прав комерційних пошукових систем.

Другий підхід до побудови агента пов'язаний з обходом Інтернет і заснований на використанні посилань на нові документи з раніше завантажених документів. Цей підхід також широко використовується, хоча і вимагає великих обчислювальних ресурсів.

Позитивні сторони:

− об'єктивність. У даному випадку інформація витягується безпосередньо з мережі, що забезпечує її об'єктивність;

− керованість процесу отримання інформації. На відміну від непрямого доступу до індексу комерційної системи через формування спеціальних запитів, доступ до інформації у даному підході більш прозорий. Є прямий зв'язок між алгоритмом сканування Інтернет і тематичною спрямованістю завантажуваних документів. Це дозволяє підбирати параметри алгоритму, що мінімізують середнє відхилення тематики завантажуваних документів від заданого тематичного напряму.

Негативні сторони:

− висока вартість. У даному випадку мережний робот реально сканує Інтернет, що приводить до великих витрат навіть при використанні обмеженого (заданою тематикою) пошуку.

З розширенням Інтернет мережі головна складність роботи пошукових систем полягає у забезпеченні релевантних відповідей на пошукові запити користувачів, тобто видати користувачам посилання на ті ресурси, які, на їх думку, відповідають тому, що вони шукали. Відповідно до своєї первинної концепції WWW була середовищем, у значній мірі орієнтованим на участь людей. Коли з часом виникла необхідність в обробці Web-контента різними роботами, агентами, то стало зрозуміло, що інтерпретувати інформацію так само якісно, як людина, вони не можуть. Виправити ситуацію має Семантична мережа [5]. Головна відмінність Семантичної мережі полягає в тому, що кожна сторінка Семантичної мережі містить інформацію на двох мовах: на звичайній, зрозумілій людині і що показується браузером, і на спеціальній, інформація на якій прихована від людських очей, але зрозуміла інтелектуальним програмам-агентам, роботам.

Фундамент Семантичної мережі складають онтології, які є описом на деякій формальній мові понять певної предметної області і відносин між ними. Онтології розробляються і можуть бути використані при вирішенні різних задач, у тому числі для сумісного застосування людьми або програмними агентами, для можливості накопичення і повторного використання знань в предметній області, для створення моделей і програм, оперуючих онтологіями, а не жорстко заданими структурами даних, для аналізу знань в предметній області [6].

Онтологія − це набір визначень (формальною мовою) фрагмента декларативних знань, орієнтований на спільне багаторазове використання різними користувачами у своїх застосуваннях. В онтології вводяться терміни, типи і співвідношення (аксіоми), що описують фрагмент знання [7].

Інше визначення онтології – комплекс понять від самих загальних до найбільш конкретних, що охоплює повний спектр об’єктів і відносин, включаючи події і процеси, а також значення (атрибутів і відносин), обумовлені, якщо необхідно, у часі і просторі. Ця система сутностей зв’язується як універсальними залежностями типу “загальне – частка”, “частина – ціле”, “причина – наслідок” і т.п., так і специфічними для відповідного домену. Визначаючи сутності у онтології, можна використовувати різні апарати представлення знань, − наприклад, фрейми, слоти, що обумовлюють припустимість отримання їхніх значень. Як обмеження можуть виступати продукції, логічні, алгебраїчні, табличні й інші залежності. Таким чином, онтологія – це модель предметної галузі, що використовує всі доступні засоби представлення знань, релевантні для даної галузі.

Розглянемо архітектуру мультиагентної системи, що є об’єктом нашого дослідження. В цій мультиагентной системі можна виділити такі типи агентів: агент користувача, пошуковий агент, агент формування запиту, агент онтології, агент інтеграції сторінок, агент переваг користувача, агент реєстру IP(інформаційних ресурсів). Структура системи, що пропонується, показана на рис. 1.

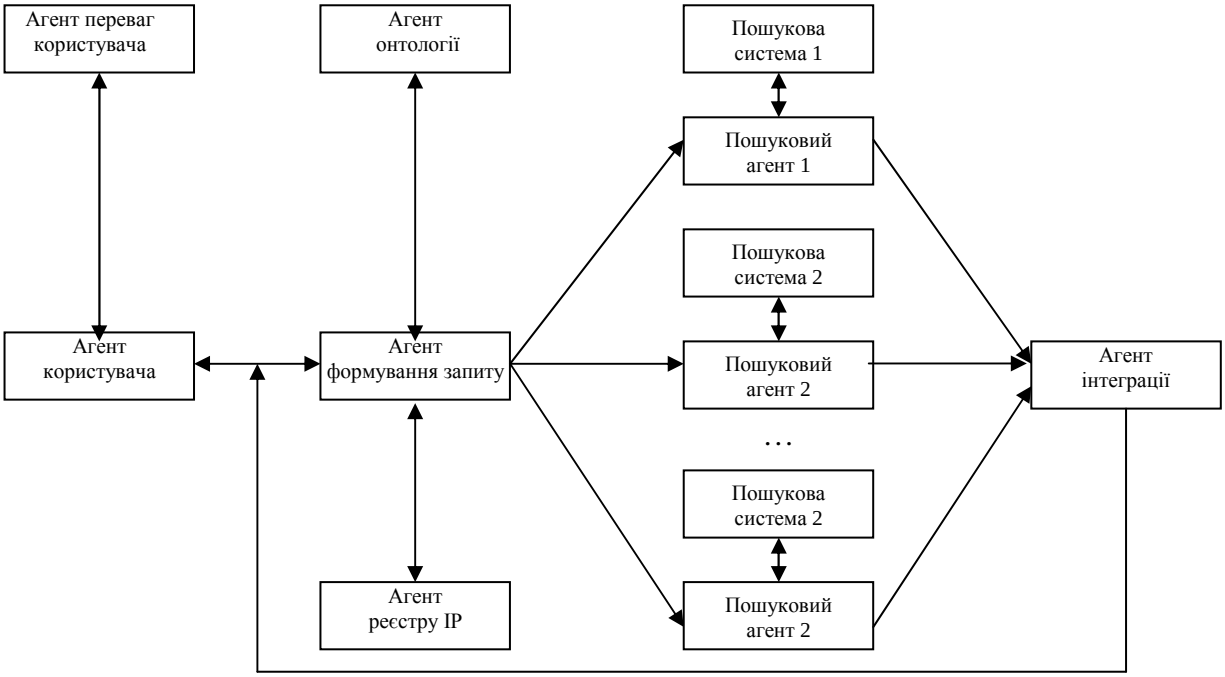

Рис. 1. Структурна схема мультиагентної пошукової системи

Архітектура є відкритою і модульною, що дозволяє легко включати нові онтології і нові інформаційні ресурси.

Агенти користувача і переваг. Призначений для користувача агент взаємодіє з користувачем, щоб виявити його переваги, які управляються агентом переваг. Ці переваги включають: відносні значення, які приписані термам, що використовуються для формування запитів, рейтинг Web-машин пошуку, правила ранжирування результатів обчислення і інші переваги. Агент переваг також вивчає переваги користувача, які засновані на досвіді і зворотному зв'язку з попередніми запитами.

Агент онтології. Агент онтології має доступ до онтологій, щоб визначити концепти семантичного пошуку. Онтології можуть бути визначені на мові OWL. Використання лексичної бази даних WordNet дозволяє удосконалити запит. Наприклад, WordNet може забезпечити колекцію синонімів для терміну.

Агент формування запиту. Користувач надає початковий запит агенту формування запиту. Цей агент, у свою чергу, консультується з агентом онтології, щоб удосконалити або узагальнити запит, грунтуючись на семантичному посередництві, забезпеченому доступними сервісами онтології. Як тільки запит був визначений за допомогою взаємодій між призначеним для користувача агентом і агентом онтології, агент формування запиту розділяє запит на підзапити, націлюючись на відповідні джерела даних. Необхідно забезпечити агента формування запиту правилами і політиками, щоб допомогти йому робити більш інтелектуальні рішення про оптимізацію запиту. Крім того, система може мати репозиторій оброблених запитів. Ця інформація використовуватиметься агентом формування запиту як база випадків і результати якої можуть використовуватися багато разів.

Агент інтеграції. Агент інтеграції відповідальний за компіляцію результатів підзапитів від різних джерел, ранжирування їх згідно переваги користувача, яка забезпечується агентом переваги.

Кожний пошуковий агент взаємодіє з визначеною пошуковою системою, передає їй запити і повертає результати її роботи агенту інтеграції.

Агент реєстру джерел. Основою для визначення та специфікації джерел електронних ресурсів є реєстр електронних інформаційних ресурсів. Агент реєстру допомагає агенту формування запиту оптимізувати запит, розділити його і передати відповідним агентам пошукових ситем.

Для кожного документа мережний агент обчислює оцінку близькості документа до тематики, заданої користувачем. Ця оцінка використовується для зменшення кількості нерелевантних документів. Метод обчислення оцінок грунтується на методі обчислення відстаней у рамках векторної моделі документів [8], що використовується в задачі інформаційного пошуку. Для обчислення оцінок агенту необхідно мати деякий опис тематики − тематичний фільтр, який має бути наданий користувачем. Замість оцінюваного документа при обчисленні оцінок використовується його профайл, який будується таким чином. З документа видаляються загальновживані слова, що не мають ніякої тематичної спрямованості (займенники, прийменники і т. п.). Для

всіх слів, що використовуються, виділяються їх основи (так звані терми) і обчислюються частоти їх використання в документі. Підсумковий профайл документа d є вектором пар: (t, $t\ f_{t,d}$) − терм t і частота його використання $t\ f_{t,d}$. Заданий користувачем тематичний фільтр теж є вектором пар: (t, w^t_{filter}) − терм t і його значущість w^t_{filter}. Користувач також зазначає поріг рекомендації T, який використовується агентом для ухвалення рішення про рекомендацію даного документа. Агент обчислює оцінку тематичної релевантності документа згідно наступній формулі:

$$r(d)=\Sigma t\ f_{t,d} \cdot w^t_{filter}\ . \qquad (1)$$

Класичними критеріями оцінки методів фільтрування є точність, тобто частка релевантних документів у загальному числі рекомендованих (що пройшли через фільтр), і повнота, тобто частка рекомендованих у загальній кількості релевантних документів. Проте, оскільки остаточне рішення про тематичну релевантність документа ухвалюється користувачем, то метою мережного агента є не самостійний відбір релевантних документів, а зменшення кількості документів за рахунок відсіву явно не відповідних документів. Тому основний критерій якості методу фільтрування − це не точність, а повнота і частка відфільтрованого шуму, тобто частка знайдених фільтром нерелевантних документів за відношенням до загального числа нерелевантних документів.

2. Реалізація агентного підходу до тематичного пошуку інформації у системі E-content

Пошукова система E-Content призначена для об’єднання наявних або створюваних цифрових ресурсів наукових установ і забезпечення інтегрованого представлення, пошуку та доступу до інформаційних ресурсів. Система забезпечує збір, нормалізацію, індексування і актуалізацію корпоративних цифрових ресурсів. Основою для визначення та специфікації джерел електронних ресурсів є реєстр електронних ресурсів.

Реалізована компонента агента онтології у системі E-Content дозволяє будувати тематичну онтологію, яка використовується для формування та виконання запиту пошуковими агентами системи (рис. 2).

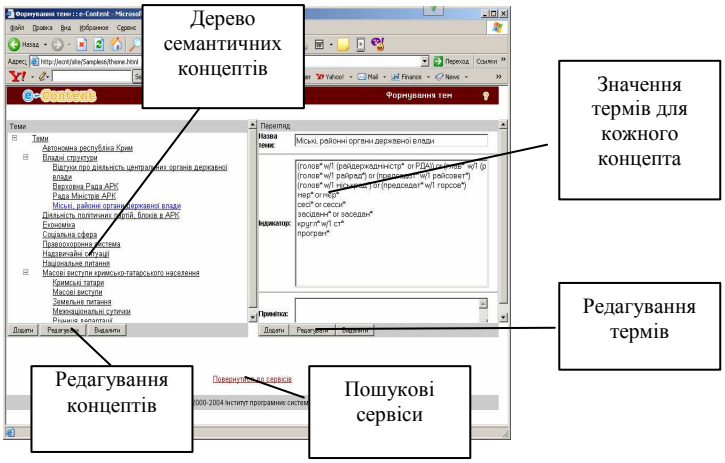

Рис. 2. Формування тематичної онтології

У системі підтримується можливість формувати тематичну онтологію двома шляхами:

− редактор тематичних онтологій, в якому користувач визначає основні об’єкти онтології (семантичні концепти, терми пошуку і т.д.);

− формування тематичної онтології за визначеним текстом або множиною текстів.

Користувач зазначає один або декілька тестових файлів, на основі яких автоматично будуть визначені теми та відповідні їм пошукові образи (рис. 3). Графічний інтерфейс прискорює та спрощує процес формування тематичної онтології у відповідності до потреб користувача.

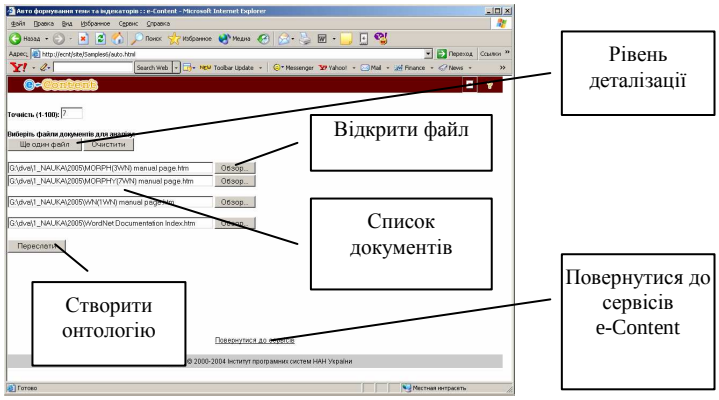

Рис. 3. Формування тематичної онтології за документом-зразком

Висновки

Інтернет, як джерело інформації, характеризується величезним обсягом інформації, динамічністю та дублюванням інформації. Розробка нових підходів та методів до побудови пошукових систем залишається актуальною на сучасному етапі. В роботі розглянуто архітектури пошукових систем, запропоновано використання методів Семантичної мережі при побудові – тематичні онтології. Розглянута архітектура мультиагентної пошукової системи. Спеціалізовані агенти системи дозволять підвищити якісний рівень пошукових систем.

Основні результати досліджень реалізовані в системі e-Content**,** яка використовує тематичні онтології для виконання пошукового сервісу. Система e-Content використовується як пошуковий сервіс у корпоративному порталі НАН України www.nas.gov.ua.

1. *Andon Ph., Deretsky V.* Control Oriented Ontology and Process Description for Cooperation Agents in Information Retrieval // Sixth International Scientific Conference „Electronic Computers and Informatics ECI'2004". − Kosice – Herlany, Slovakia; September 22-24, 2004. − P. 14 – 18.
2. *Patel, L. Petrosjan,* and *Rosenstiel W.*, editors. OASIS: Distributed Search System in the Internet. St. Petersburg State University Published Press, St. Petersburg, 1999.
3. *Lieberman Henry*. Letizia: An agent that assists web browsing. In Chris S. Mellish, editor, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence // Morgan Kaufmann publishers Inc. − San Mateo, CA, USA, 1995. − P. 924 − 929.
4. *Martin E. Meller*. Machine learning based user modeling for www search, http://citeseer.nj.nec.com.
5. *Tim Berners-Lee, James Hendler and Ora Lassila*. The Semantic Web http://www.sciam.com/article. cfm?articleid =000A0919
6. *Дерецький В.О.* Підхід до автоматичної побудови тематичної онтології документу для удосконалення інформаційного пошуку. − 2005. − № 3. − С. 76 − 82.
7. *OWL* Technical Committee. Web Ontology Language (OWL). http://www.w3.org/TR/2004/WD-owlref
8. *Salton G., Buckley C.* Term-Weighting Approaches in Automatic Text Retrieval // Information Processing and Management. − 1988. V. 24. − P. 513 − 523.