9. Богатырёва Ю. А. Мультимножества: библиография, решетка мультимножеств / Ю. А. Богатырёва // Theoretical and Applied Aspects of Program Systems Development: international conference, December 8–10, 2009. – Kyiv. – 2009. – С. 13–20.

10. Буй Д. Б., Богатырева Ю. А. К вопросу о решетке мультимножеств / Д. Б. Буй, Ю. А. Богатырева // Десятый международный семинар «Дискретная математика и ее приложения», Москва, МГУ, 1–6 февраля 2010 г.

D. Buy, S. Polyakov

COMPOSITIONAL SEMANTICS OF THE RECURSIVE EXPRESSIONS AND THEIR GENERALIZATIONS IN LANGUAGES LIKE TO SQL

The paper describes a method for showing hierarchies in relation databases uses an adjacency list model. The paper introduces the adjacency lists sorts and their samples. The navigations queries are described as well as common table expression in their recursive format. Samples of such queries are shown. The paper defines formal semantic of the recursive common table expression.

Keywords: compositional semantics, recursive queries, SQL, common table expression, CTE.

УДК 004.8

Глибовець А. М., Шабінський А. С.

ОДИН ПІДХІД ДО ПОБУДОВИ ІНТЕЛЕКТУАЛЬНОЇ ПОШУКОВОЇ СИСТЕМИ

У роботі подано загальне бачення архітектури інтелектуальної пошукової системи наукових матеріалів. Звуження предметної області проведено для покращення відсіву пошукового «сміття» і структурованості бази знань.

Ключові слова: інтелектуальна пошукова система (ІПС), інформаційний пошук (ІП), семантичний пошук, пошук за індексом,онтологія, пошуковий агент,

Розвиток World Wide Web (WWW) спричинив суттєве збільшення об'єму інформації в Інтернеті. Постає питання: як розрізнити інформацію від знання та прискорити час її обробки? Зрозуміло, що тут може нам допомогти контекст знань.

Метою цієї статті є розробка архітектурних принципів функціонування та впровадження інтелектуальної пошукової системи, яка б дозволяла здійснювати оптимальний пошук документів наукового та науково-публіцистичного типу з використанням семантичної мережі.

Інтелектуальна пошукова система (ІПС) має ряд суттєвих переваг порівняно із традиційними пошуковими системами, зокрема базованими на пошуку за ключовими словами. Більшість переваг полягає у використанні покращеної моделі інформаційного пошуку (ІП), інтерактивної взаємодії з користувачем, участі експертів, автоматизованих засобів формування та підтримки бази знань, спробі «зрозуміти» інформаційну потребу користувача. Важливою функцією ІПС стає автоматичне динамічне накопичення знань та мета знань у процесі роботи системи. В цьому контексті досвід – це сукупність опрацьованих і впорядкованих знань, результат роботи експертів та опрацювання відгуків користувачів. Зрозуміло, що в цьому випадку здійснюваний ІПС-пошук має бути семантичним. Тобто система має оперувати не тільки вербальним описом сутностей предметних областей, а й семантичними поняттями, які вкладені в інформаційні одиниці та зв'язки між ними.

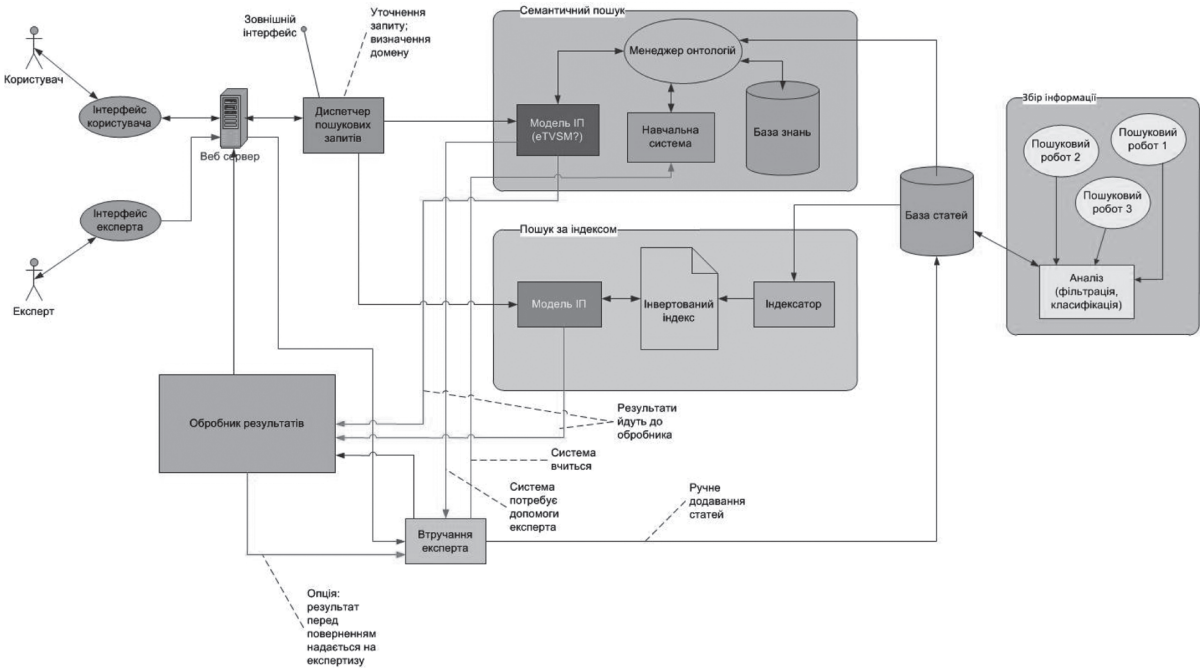

Рис. 1. Концептуальна архітектура ІПС

У нашому підході ІПС має відповідати таким вимогам:

- середовище пошуку подається онтологіями та механізмами інтерпретації;
- аналізувати інтерактивно інформаційні потреби користувача та проводити семантичний пошук;
- підтримувати самонавчання шляхом врахування динамічних змін середовища пошуку, автоматичного розширення онтологій предметних областей та за допомогою експертів і користувачів.

Особливістю нашої ІПС є робота з текстовою інформацією, переважно наукового та науково-публіцистичного типу. Джерелами інформації системи є WWW, тексти, додані авторами до бази даних статей системи, експертні знання та співпраця зі спеціалізованими електронними бібліотеками і репозитаріями. Основна роль експертів буде полягати в аналізі та класифікації специфічної інформації, уточненні та покращенні результатів пошуку, додаванні нової інформації, ручному редагуванні бази знань тощо.

У системі реалізуються паралельно два типи пошуку: семантичний та класичний за індексом, що надає більшу повноту та релевантність результатів пошуку.

Архітектура ІПС повинна забезпечувати можливість автоматизованої побудови онтологій певних предметних областей за участю експерта та накопиченого досвіду.

Концептуальна архітектура нашої ІПС представлено на рисунку 1.

Як видно з рисунку, для побудови ІПС використовується модульний підхід. Розглянемо основні функціональні частини системи: збір інформації, пошук за індексом, семантичний пошук, диспетчер пошукових запитів, оброблення результатів, експертне втручання.

Підсистема збору інформації

Підсистема збору інформації (ПЗІ) здійснює сканування ресурсів-джерел на наявність потрібної інформації, фільтрує, класифікує за тематикою та зберігає її.

Для реалізації цих функцій застосовуватимуться методи машинного навчання, алгоритми кластеризації та класифікації. Збір та аналіз інформації проводитиметься на базі агентно-орієнтованої архітектури, подібно до роботи [1]. За виділеними предметними областями закріплюються окремі пошукові агенти.

Пошуковий агент має виконувати такі функції: пошук посилань на нові джерела інформації, виокремлення й очистка тексту, визначення рівня науковості тексту, визначення належності тексту предметній області з подальшою передачею на аналіз.

Складовими частинами пошукового агента є модулі визначення рівня науковості тексту і відповідності предметній області. Зупинимося докладніше на цих модулях.

Головними ознаками наукового стилю є інформативність, понятійність і предметність, об’єктивність, логічна послідовність, узагальне-

ність, однозначність, точність, лаконічність, доказовість, переконливість, аналіз, синтез, аргументація, пояснення причинно-наслідкових відношень, висновки. У реалізованій нами підсистемі пошуковий агент працює з XML-файлом, структура якого разом із визначеним набором слів і словосполучень відображає критерії визначення текстів наукового типу (стилю). Рівень науковості тексту визначається частотою появи цих слів та їхніх словоформ (відмінків) у тексті аналізу. Проведений нами аналіз засвідчує те, що текст можна віднести до наукового, якщо частота появи в ньому виділених слів конфігураційного файлу більша за 1,5 %. Для покращення алгоритму в описаний аналіз можна внести додаткові параметри (важливість слова, спеціалізовані шаблони структурування).

Для визначення належності тексту предметній області пошуковий агент обробляє інший XML-файл, що характеризує специфіку предметної області. За проведеними підрахунками адекватне віднесення тексту до предметної області можливе за наявності в ньому більше 10 % термінів із конфігураційного XML-файлу. Планується заміна XML-файлу на онтологію предметної області, що значно покращить результати аналізу.

Підсистема пошуку за індексом

Ця підсистема реалізує традиційний підхід до пошуку інформації із застосуванням добре досліджених методів ІП на базі моделі прихованого семантичного аналізу (Latent Semantic Analysis, LSA) і т. п. Зокрема, кластеризація як метод машинного навчання без експерта може бути застосована для групування документів відповідно до інформаційної потреби [2]. Прихований семантичний аналіз дозволяє враховувати синонімічні зв'язки у текстах, імітуючи семантичний пошук без використання тезаурусів, онтологій тощо. Згідно з роботою [4], в якій проведено порівняльний аналіз різноманітних моделей, ІП LSA виявляється другою за кривою точності-повноти після моделі Enhanced Topic-based Vector Space Model (eTVSM), яка, проте, є онтологічно орієнтованою і буде використовуватися в нашій системі в модулі семантичного пошуку.

Зауважимо, що хоча традиційний пошук у нашій системі постулюється як менш ефективний у задоволенні інформаційної потреби, менш виразний, проте він володіє кращою обчислювальною ефективністю і додає повноту семантичному модулю пошуку інформації.

Підсистема семантичного пошуку

У семантичному пошуку використовуватиметься модель ІП Enhanced Topic-based Vector Space Model (eTVSM) [3, 4]. У моделі розроблено концепцію визначення відношень між поняттями шляхом усунення незалежності між категоріями і використання онтології як джерела знань про семантичні зв'язки між поняттями предметних областей. Спосіб визначення схожості документів побудовано на основі концепції семантичного змісту інтерпретації термінів. Найменшою одиницею інформації у eTVSM є термін. Термін може складатися із декількох слів і позначати будь-яке цілісне поняття. З кожним терміном пов'язана довільна кількість інтерпретацій. Незалежність термінів усуває цикли в структурі онтології [4].

Для покращення пошуку окремо будуватиметься база знань з кожної предметної області, що буде включати опис основних визначень, а також їхні взаємозв'язки. В цю базу знань будуть додані правила логічного виведення, які зможуть надавати розширену відповідь.

Онтологія предметної області будуватиметься автоматично. На початку з кожної предметної області вводиться каркас, що визначає загальну структуру. В процесі роботи цей каркас поповнюється автоматично знаннями, які формуються на знайдених текстах обраної предметної області.

Підсистема-диспетчер пошукових запитів

Ця підсистема проводить аналіз запиту користувача природною мовою з подальшою інтерпретацією для реалізації пошуку. Уточнення пошукової потреби для надання більш релевантних результатів реалізується за рахунок інтерактивного спілкування з користувачем.

Підсистема оброблення результатів

Ця підсистема отримує впорядковані результати роботи пошукових модулів. Результати традиційного пошуку покращуються за рахунок результатів семантичного пошуку.

Експертні втручання

Для кожної предметної області формується команда експертів. Експертна команда розробляє початкові словники та правила предметної області для пошукового робота, створює каркас онтології предметної області. Онтологія термінів будується на основі початкового словника. Застосовуючи методи машинного навчання, інформацію для наповнення онтологій можна отримувати із еталонного набору документів, причому тематична класифікація теж може бути автоматизована. У такому разі експертне втручання значно покращить результати роботи автоматизованих засобів, що, зрештою, позитивно вплине на якість результату в цілому.

На етапі ініціалізації експерти додають посилання на сховища ресурсів, з яких система має почати аналіз. Вони реагують також на запитання пошукового агента у випадках, коли ймовірність віднесення матеріалів до предметної області недостатня для автоматичного внесення, але значна, щоб їх просто відкинути. На цій базі розширюється словник предметної області й онтологія.

Аналогічно відбувається експертне втручання в процес побудови бази знань предметної області.

У випадку, коли на запит користувача система не може знайти релевантної відповіді, запит користувача передається експертам для опрацювання, і за необхідності – для ручного розширення бази знань.

Висновок

У роботі подано загальне бачення архітектури інтелектуальної пошукової системи наукових матеріалів. Звуження предметної області проведено свідомо для покращення відсіву пошукового «сміття» і структурованості бази знань. Запропоновану архітектуру можна використати при побудові депозитаріїв навчальних матеріалів НаУКМА.

1. Espinasse B. Agent and Ontology Based Information Gathering on Restricted Web Domains with AGATHE / Bernard Espinasse, Sébastien Fournier, Fred Freitas // Proceedings of the 2008 ACM symposium on Applied computing. – New York : ACM, 2008. – P. 2381–2386. – ISBN 978-1-59593-753-7.
2. Manning Ch. Introduction to Information Retrieval / Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. – New York : Cambridge University Press, 2008. – 496 p. ISBN-10: 0521865719.
3. Kuropka D. Modelle zur Repräsentation natürlichsprachlicher Dokumente / Kuropka, D. –Berlin : Logos Verlag, 2003.
4. Polyvyanyy Artem. Evaluation of a Novel Information Retrieval Model: eTVSM. – Potsdam : HPI, 2007.

A. Glybovets, A. Shabinskiy

ONE APPROACH TO THE CONSTRUCTION OF INTELLECTUAL SEARCH SYSTEM

The paper presents a general vision of the intellectual architecture of research materials search engine. Narrowing of subject area held to improve screening search «garbage» and structured knowledge base.

Keywords: XML, World Wide Web (WWW), Latent Semantic Analysis (LSA), Enhanced Topic-based Vector Space Model (eTVSM)