УДК 681.3

В. О. Дерецький

ПІДХІД ДО АВТОМАТИЧНОЇ ПОБУДОВИ ТЕМАТИЧНОЇ ОНТОЛОГІЇ ДОКУМЕНТА ДЛЯ УДОСКОНАЛЕННЯ ІНФОРМАЦІЙНОГО ПОШУКУ

Пропонується новий підхід для покращення існуючих інформаційних пошукових систем шляхом додання семантичних розширень, які посилюють якість послуг інформаційного пошуку.
Основою підходу є використання методу LSI (Latent Semantic Indexing), де з текстових докумен тів будується семантичні концепти конкретної тематичної онтології і яка представлена у якості пошукового контексту користувача. Підхід націлений на автоматичне створення тематичної онтології на основі текстових документів користувача. Створена онтологія використовуються засобами асистенту семантичного пошуку, які є проміжними між запитами користувача та пошуковими машинами.

Вступ

Для знаходження точної інформації в Інтернет-середовищі необхідно витратити надто багато часу і передивитися велику кількість Web-сайтів та Web-сторінок. Звичайно, пошукові машини Інтернет допомагають користувачеві прискорити процес інформаційного пошуку, але часто відсутність контексту пошукових слів перешкоджає ефективності послуг інформаційного пошуку. Ці проблеми вирішуються шляхом впровадження технології XML, яка була розроблена з метою формування опису структури та семантики даних, але більшість Web-сайтів все ще створюються з використанням технології HTML.

Іншою проблемою є знання про інтереси користувача або контекст в процесі інформаційного пошуку. Пошукові машини віднаходять та класифікують знайдену інформацію на основі ключових слів та їх характеристик. Отримується велика кількість сторінок, що містять ключові слова, проте сторінки можуть не мати необхідної інформації для користувача. Пошукові слова характеризуються множинами значень (проблема полісемії) [1]. Контекст, в якому ці слова з'являються, допоможе відрізнити найбільш відповідне значення запиту.

У підході пропонується використати відомості про потреби користувача шляхом підтримки моделі його інтересів на стороні клієнта та використання їх для фільтрації найбільш відповідних запиту результатів. Таку модель користувач може створити, застосовуючи текстові документи, контекст яких його цікавить.

Інший аспект запропонованої методики полягає в тому, що користувач може не мати достатньо знань про терміни певної предметної області, за якими здійснюється пошук інформації в Web-середовищі. Наприклад, про деталі деяких виробів чи матеріалів користувач може не мати уявлення, але ця інформація може міститись у відповідних документах, які можуть використовуватися у якості профілів інтересу користувача.

Якість результатів пошуку значно різниться залежно від якості пошукового запиту: може отримуватись як надто обмежений список посилань, так і надмірно велика кількість невідповідних посилань. В окремих випадках достатньо конкретизувати запит парою ключових слів.

Багато користувачів Інтернету шукають інформацію, яку не можна легко описати за допомогою кількох ключових слів. Найчастіше очікувані результати отримують за допомогою кількох пошукових запитів. Тексти, одержані внаслідок одного запиту, є контекстом для формування більш точного наступного запиту.

Запропонований підхід спрямований на автоматичне створення тематичної онтології з текстових документів користувача. Його метою є створення узгоджених із заданим документом чи мно-

жиною документів семантичних категорій та відповідних ключових слів.

Підхід базується на використанні методу LSI (Latent Semantic Indexing) [2], за яким з текстових документів будується предметно-орієнтована онтологія, яка використовується для пошукового контексту користувача. Документи читаються, оброблюються і створюється онтологія ієрархічної структури. Цю онтологічну структуру можна розглядати та модифікувати за допомогою графічного інтерфейсу користувача з метою побудови найбільш ефективних запитів. Побудовані семантичні інструменти приховують від користувачів складність створення семантичної структури та використання мови запиту конкретної пошукової машини. Засоби виконують стандартні дії, які робить більшість користувачів, щоб досягти кращих показників та отримати більш відповідні результати.

Засоби семантичного пошуку

У процесі досягнення цієї цілі використовуються сучасні статистичні методи інформаційного пошуку, зокрема метод латентного семантичного індексування (Latent Semantic Indexing – LSI), за якими робиться спроба зрозуміти статистичні посилання термінів шляхом заміни простору термів документа на значно менших розмірів простір концептів. В LSI це виконується використанням методу матричної декомпозиції – Singular Value Decomposition (SVD). Ефективність SVD у порівнянні з іншими методами описана в [2].

Для побудови конкретної онтології використовується послідовність наступних процедур: читання збірки відповідних текстових документів та вилучення онтологічної інформації статистичними методами шляхом попередньої обробки текстів документів (Pre-processing), нормалізації текстів (Normalization), формування семантичних концептів, що відносяться до значимих термінів, з використанням LSI та SVD.

Попередня обробка є процесом, в якому здійснюється здобуття значимих термінів та підраховується їх частоти під час читання чи завантаження текстового файлу. Над текстовим документом виконуються кілька процедур представлення тексту в необхідному форматі для впевненості у тому, що підрахована статистика є значимою. Ці процедури стосуються визначення загальних основ слів та відсічення відмічених, що семантично малозначні.

Нормалізація – це процес, за якого рахується нормалізована вага кожного слова, яке було отримано в результаті попередньої обробки. В попередній обробці документ аналізується в аспекті корелятивних слів та матриці частотності, відомої як матриця „терм-документ”, що створюється в результаті проведення аналізу.

З використанням методу латентного семантичного індексування (Latent Semantic Indexing – LSI) створена матриця розкладається на три матриці: термів *(U)*, одинична діагональна *(S)* та документів *(V)*. Після мінімізації об’єму матриць матриця термів розкладається на вектори термінів, визначених як концепти, що формуються як група відповідних термінів.

Побудова онтології документа є, по суті, побудовою концептуальних та термінологічних вузлів з матриці термів *(U)* і документів *(V)*. Концептуальний вузол представляє концепт і містить інформацію про його назву, терми, що відносяться до нього, та їх ваги в концепті.

Графічний інтерфейс дозволяє користувачеві легко переглядати та редагувати онтологію. Засоби створення онтології документа базуються на наступних етапах: введення (текстових документів); попередня обробка, нормалізація; індексування за методом LSI, що базується на SVD, побудова онтології документа (рис. 1).

Нижче розглянемо основні деталі визначених процедур.

Попередня обробка документів

Попередня обробка документів – це процес, за якого визначаються значимі терміни і точно рахується їх часто-

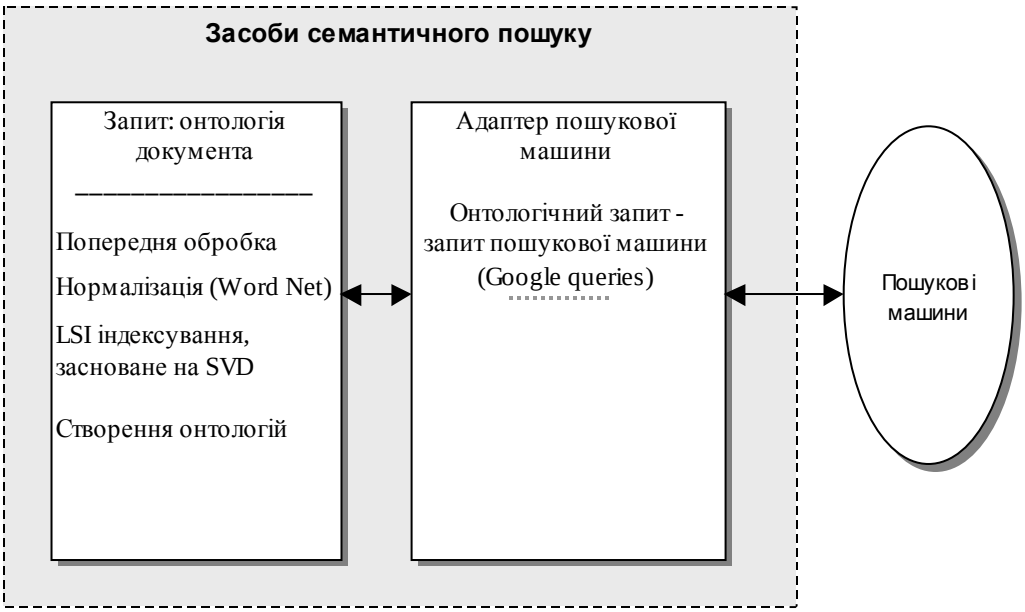

Рис. 1. Схема засобів семантичного пошуку

тність під час читання текстового файла. Попередня обробка включає фільтрування текстового документа, щоб полегшити статистичний аналіз. Під час попередньої обробки текстового документа система виконує наступні кроки:

числа і розділові знаки видаляються;

слова перевіряються та приводяться до загального формату символів (low-case words); наприклад, якщо текст містить слова "книга" і "Книга", то на виході програми частотність слова “книга” повинна дорівнювати двом. Виключаються артиклі, прийменники, сполучники та ін.

Ще один аспект попередньої обробки – це рахунок слів у конкретних граматичних формах (неправильні дієслова, іменники латинського походження і т.д.).

Остання процедура етапу попередньої обробки: "виділення основи” (stemming). Метою є обмежити зміни, що виникають, коли зустрічаються різні граматичні форми того ж самого слова (наприклад, "президент" – "президентський", "працювали" – "працівники " тощо).

Для попередньої обробки документів використовуються засоби лексичної бази даних WordNet (lexical database tools) [3, 4]. Функції морфологічної обробки WordNet обробляють широкий спектр морфологічних характеристик. Засоби Morphy використовують два типи обробки, щоб спробувати конвертувати форму слова в форму, яку можна знайти в базі даних WordNet. Доступ до морфологічного процесора WordNet Morphy можна отримати шляхом наступних функцій:

morphstr() є основним інтерфейсом користувача Morphy. Функція полягає в намаганні знайти основну форму слова (лему) або сполучення. В результаті виконання функція повертає покажчик до знайденої основної форми. Подальші виклики функції повертають основні форми слова тієї ж групи**.** Коли більше не знайдено основних форм, то повертається NULL;

morphword() намагається знайти основну форму слова в точно зазначеній позиції. Ця функція викликається *morphstr()* для кожного окремого слова.

Для попередньої обробки багатомовних документів планується скористатися підходом проектів WordNet Europe та WordNet Rus в яких зберігається основна концепція WordNet [1].

Нормалізація

Нормалізація – це процес, в якому рахується нормалізована вага кожного слова, отриманого після попередньої обробки [5]. При нормалізації в першу чергу здійснюється визначення кореляції слова. Першим кроком є визначення числа частотності кожного терму в документі. Потім рахується вага кожного терма за наступною формулою:

$$W(i,k) = \frac{fr(i,k)}{\sum_{j=1}^{nk} fr(j,k)},$$

де *W(i, k)* – вага *i*-го терма в *k-му* документі; *nk* – загальна кількість термів в документі; *fr (i, k)* – частота терма *i* в документі *k.* Ця вага відповідає терму в документі. Але вага терму повинна нормалізуватися відповідно до множини документів.

У наступному кроці нормалізація кожного окремого документа об'єднується з нормалізацією збірки документів. Треба зауважити, що термін може мати велику вагу просто тому, що документ, в якому він зустрічається, є замалий і тому розраховується частота, з якою він зустрічається по всій збірці документів. Нормалізована вага терму розраховується за формулою

$$NW(i,k) = \frac{W(i,k)}{\sqrt{\frac{nk}{j=1} W \frac{2}{(ik)}}}.$$

Цей процес є стандартною нормалізацією термів для документа [6].

Створення матриці „терм-документ"

На цьому етапі текстовий документ представляється як група значимих нормалізованих термінів. Ваги нормалізованих термінів разом формують матрицю *W*, де *W(i,k) = NW (i, k).* На цю матрицю посилаються як на термінологічну матрицю документа. Рядками термінологічної матриці є терміни, а документи представляються її стовпцями.

Таким чином створюється термінологічна матриця документа, яка описує частотність значимих термінів в кожному документі збірки. Для цього визначаються значимість термінів, що зустрічаються в різних документах. При обробці нового документа система повинна перевірити, чи ці нові терміни вже є граматичними формами деяких попередніх термінів.

Результатом цього процесу є термінологічна матриця документа, що містить значимі терміни всієї збірки документів у якості рядів, і документів збірки у якості стовпців.

Метод семантичного індексування

На цьому етапі визначається група концептів з термінологічної матриці документа, де концепт визначається як група відповідних термів. Це здійснюємо шляхом використання методу, що має назву методу латентного семантичного індексування (Latent Semantic Indexing – LSI), який включає процедуру декомпозиції матриці *W* з використанням методу Singular Value Decomposition (SVD) [6].

Метод LSI – це статистичний метод, який пов'язує терміни тексту в семантичну структуру без синтаксичного чи семантичного аналізу природномовних текстів та без ручного втручання людини. Використовуючи цей метод, кожний документ представляється не за термінами, а за концептами, які деякою мірою дійсно статистично незалежні, а терміни – не є такими. Терміни не можуть використовуватися у якості дескриптора документа, оскільки припускається, що вони незалежні. Але деякі терміни повторно зустрічаються в різних документах, і не повинні розглядатися як незалежні.

Концепція LSI досліджена та описана в [2].

Метод LSI використовує в свою чергу метод SVD – Singular Value Decomposition, ретельно описаний в [2, 6].

Метод матричної декомпозиції

Метод SDV – Singular Value Decomposition – відомий метод матричної декомпозиції [3]. Він розкладає матрицю *W*, наприклад: термінологічна матриця документів $m \times n$ з термінами *m* та документами *n*:

$W = U * S * V,$

де U – $m \times r$ матриця, що називається термінологічною; V – $r \times n$ матриця, яка називається матрицею документів і S – $r \times r$ діагональна матриця, що містить одиниці по діагоналі в порядку зменшення. У цій декомпозиції величини *i* відповідають вектору *ui* згідно стовпчика *i* у матриці *U* та *vi* рядку *i* у матриці *V*. Без втрат узагальнення для будь-якої частини документа можна допустити, що стовпчики матриці *U*, рядки матриці *V* та діагональні величини матриці *S* впорядковані таким чином, що величини упорядковані вниз по діагоналі в порядку зменшення. За допомогою методу LSI формується нова матриця.

$$W^s = U^s * S^s * V^s,$$

де W^s створено з W шляхом видалення всіх найбільших значень *s*; U^s – з U шляхом видалення всіх стовпчиків, які відповідають значенням *s*, що залишилися. V^s сформовано з V шляхом видалення *s* відповідно до рядків, де $s \le r$. Згідно [3], матриця W^s – це приблизна відповідність матриці W із зростаючою точністю, оскільки *s* наближається до *r*. У запропонованому підході видаляються всі *s*, значення яких нижче порогу, що визначається як процент найбільшого значення.

Матриця U^s – це матриця $m \times s$, що представляє кореляції між термінами у збірці документів. Кожний стовпчик цієї матриці *ui* є вектором, який розглядається як такий, що представляє концепт. Елементи *ui* дають кореляцію термінів до концепту. Частота концепту в документах представлена величиною *i*. Детальну інформацію про метод SVD можна отримати з [2].

Формування онтології документа

Побудова онтології документа є, по суті, побудовою концептуальних вузлів та термінологічних вузлів матриці термінів *U* та матриці документу *V*, отриманих з SVD.

Концептуальний вузол представляє концепт, що містить інформацію про свою назву та терміни, які належать до нього, і їх вагу в концепті. Назва концепту породжується автоматично з найбільш частотних його термінів, які пишуться через дефіс. Кожний стовпчик у матриці документа *U* відповідає концептуальному вузлу.

Термінологічний вузол представляє термін і містить інформацію про її назву, концепт, до якого він належить, та його вагу в різних концептах. Назва терміну генерується автоматично і вона сама є просто терміном. Кожний рядок в матриці документа *U* відповідає вузлу терміну.

Формування графа онтології

Онтологія представляється у вигляді графа. Маються два типи вузлів: концептуальні і термінологічні. Сформований граф використовується, щоб показати зв'язки між різними термінами та концептами. Концептуальні вузли поєднані з термінологічними, які непрямо пов'язані з іншими концептуальними. Термінологічні вузли зв'язані з іншими термінологічними вузлами тільки шляхом зв'язування з вузлом загального концепту.

Онтологічний граф будується з матриці *U* та списку назв термінів. З вектора *ui* видаляються терміни низької кореляції шляхом встановлення нуля в *ui*, для тих термінів, які нижче певного порогу найбільш корелятивних термінів в *ui*.

В результаті отримано модифікований вектор *ui*, який стає зразком для побудови концептуального вузла. Концептуальний вузол з'єднується зі всіма термінами, що містять ненульові елементи в модифікованому *ui* векторі, а термінологічні вузли будуються тільки для термінів, які з'єднуються з концептуальними вузлами.

Всім вузлам надаються імена. Термінологічним вузлам даються назви, що відповідають назві терміну, що використовується. Концептуальні вузли будуються автоматично. Назва концепту складається із п'яти найбільш високо корелятивних термінів у векторі концепту, написаних через дефіс. Оскільки це не приводить до надмірно інтуїтивних концептуальних назв, користувач повинен мати можливість змінити назву кон-

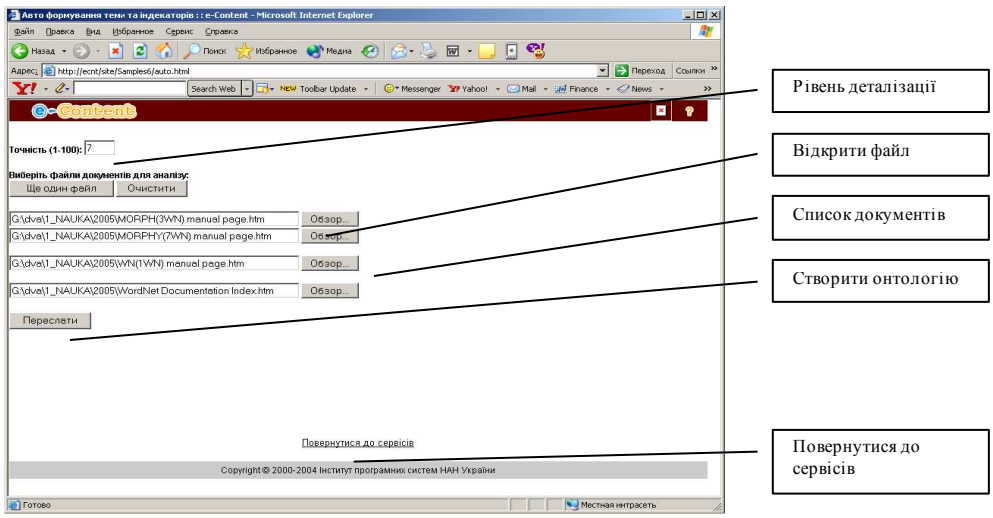

Рис. 2. Графічний інтерфейс користувача для побудови онтології документа

цепту до більш придатної, використовуючи графічний інтерфейс користувача (GUI).

Графічний інтерфейс користувача

Графічний інтерфейс користувача (далі – інтерфейс) дозволяє створювати, перевіряти та маніпулювати онтологією. Інтерфейс включає компоненти відображення графа, ієрархію концептів, список термінів, документів і також функції створення та зміни, що дозволяють користувачеві створювати та редагувати онтологію [7, 8]. Екранні форми графічного інтерфейсу представлені на рис. 2, рис.3.

Користувач може створювати групи шляхом вибору концептів та термінів із списку концептів та списку термінів відповідно.

На екрані відображається ієрархії концептів, з'єднані між собою та відповідними термінами (рис. 3).

Напрямки подальшого дослідження

Внесення декількох удосконалень до системи зробить її більш універсальною. Однією з проблем є поліпшення опрацювання документів, які змінюють

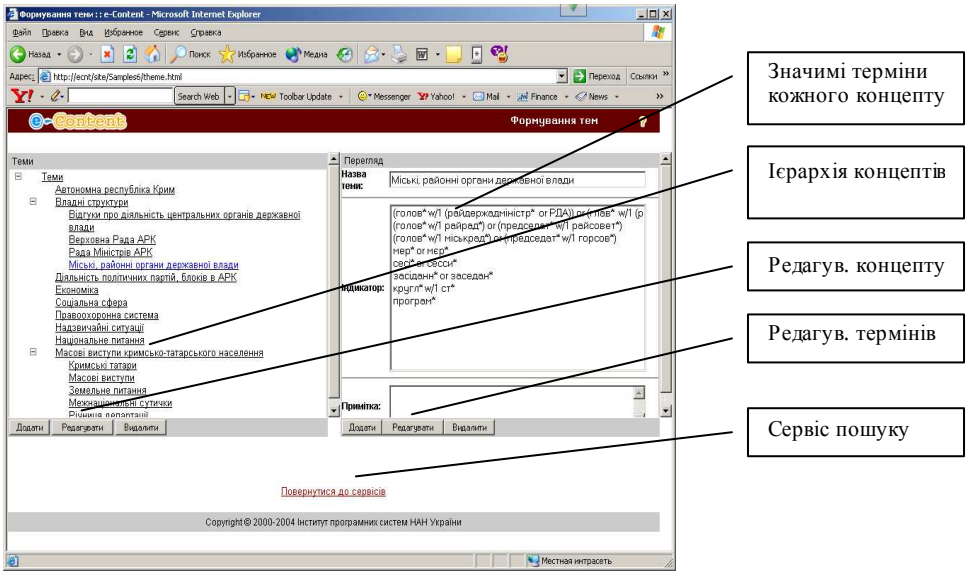

Рис. 3. Графічний інтерфейс користувача. Онтологія документа

ються, наприклад оновлені Web-сторінки, друга полягає у визначенні семантики зв'язків концептів та в використанні зв'язків між концептами в якості запиту.

Висновки

У статті запропоновані підхід та система створення онтології з текстових документів з використанням методу LSI, який будує предметно-орієнтовану онтологію із збірки текстових документів. Метод, у цій конструкції є статистичним. Він застосовує добре відому матричну декомпозицію та надає результати, дійсність яких підтверджується теоретично. Система забезпечує швидке формування предметно-орієнтованої онтології для використання її в якості запиту в інформаційно-пошукових системах.

1. RussianWordNet. - http://www.pgups.ru/webwn/
2. *Berry M.W., Dumais S.T., O'Brein G.W.* Using linear algebra intelligent information retrieval // *SIAM Review.* –1995. –37(4). – P. 573-595.
3. *Miller G.A., Beckwith R., Fellbaum Ch., Gross D., Miller K.* Introduction to WordNet: An On-line Lexical Database. http://www.isi.edu/isd/kr/5papers.pdf
4. WordNet. - http://www.cogsci.princeton.edu/~wn/
5. *Bassu D., Behrens C.* Applied Research Distributed LSI: Scalable Concept-based Information Retrieval with High Semantic Resolution. - http://research.telcordia.com
6. *Chen C., Stoffel N., Post M.* Telcordia LSI Engine: Implementation and Scalability Issues. http://citeseer.ist.psu.edu/chen01telcordia.html
7. *Андон П.І., Дерецкий В.А.* Процесори пошуку та аналізу природномовної текстової інформації в аналітичних системах // Проблемы программирования. – 2001. – N3-4. – С.144-165.
8. *Дерецкий В.А.* Об одном подходе к обработке естественно-языковых данных на основе анализа семантических сетей // Первая Всерос. науч. конф. "Электронные библиотеки: Перспективные методы и технологии, электронные коллекции", 18-22 октября 1999 г., Санкт-Петербург. – С.100-103.

Отримано 19.04.05

Про автора

Дерецький Валентин Олександрович,
канд.фіз.-мат.наук,
провідний науковий співробітник

Місце роботи автора:
Інститут програмних систем
НАН України,
Просп. Академіка Глушкова, 40
Київ-187, 03680, Україна
Тел. (044) 526 4342
E-mail:dva@isofts.kiev.ua