УДК 004.4’414

ВИКОРИСТАННЯ ПРЕДМЕТНИХ ОНТОЛОГІЙ В УПРАВЛІННІ ЗНАННЯМИ ДЛЯ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ СЕМАНТИЧНОГО ПОШУКУ

Д. Є. Костенко, Б. І. Мороз, В. В. Костенко
Університет митної справи та фінансів
вул. Дзержинського, 2/4, м. Дніпропетровськ, 49000, Україна. E-mail: denis_and_ko@ukr.net

Проведено аналіз та дослідження можливості підвищення ефективності пошуку інформації в масивах неструктурованих документів. Розглянуто методи класифікації текстів, які використовуються не тільки для пошуку, але і для рішення виникаючих у процесі пошуку проблем, оскільки класифікація текстів є одним з найбільш важливих напрямків у дослідженнях. Розроблено метод і алгоритм семантичного пошуку з урахуванням особливостей обробки масивів неструктурованих текстів у документах. Розглядається питання використання онтологій, у тому числі і для вирішення проблеми старіння інформації. Описано процес побудови онтологій. Виявлено можливості застосування онтологій для організації процесу пошуку. Для розв’язання задач інформаційного пошуку (в тому числі – для автоматичної обробки текстів) поняттям онтології зіставляється набір слів і словосполучень, якими поняття можуть виражатися в текстах. Розроблені основні модулі системи семантичного пошуку. Пропонуються практичні підходи для реалізації процесу пошуку в неструктурованих масивах даних.

Ключові слова: дані, запит з джокером, інформація, онтології, стеммінг, пошук документів, семантичний пошук.

ИСПОЛЬЗОВАНИЕ ПРЕДМЕТНЫХ ОНТОЛОГИЙ В УПРАВЛЕНИИ ЗНАНИЯМИ ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ СЕМАНТИЧЕСКОГО ПОИСКА

Д. Е. Костенко, Б. И. Мороз, В. В. Костенко
Университет таможенного дела и финансов
ул. Дзержинского, 2/4, г. Днепропетровск, 49000, Украина. E-mail: denis_and_ko@ukr.net

Проведены анализ и исследование возможности улучшения эффективности поиска информации в массивах неструктурированных документов. Рассмотрены методы классификации текстов, используемые не только для поиска, но и для решения возникающих в процессе поиска проблем, поскольку классификация текстов является одним из наиболее важных направлений в исследованиях. Разработан метод и алгоритм семантического поиска с учетом особенностей обработки массивов неструктурированных текстов в документах. Рассматривается вопрос использования онтологий, в том числе и для решения проблемы старения информации. Описан процесс построения онтологий. Выявлены возможности применения онтологий для организации процесса поиска. Для решения задач информационного поиска (в том числе – для автоматической обработки текстов) понятием онтологии сопоставляется набор слов и словосочетаний, которыми понятия могут выражаться в текстах. Разработаны основные модули системы семантического поиска. Предлагаются практические подходы для реализации процесса поиска в неструктурированных массивах данных.

Ключевые слова: данные, запрос с джокером, информация, онтологии, стемминг, поиск документов, семантический поиск.

АКТУАЛЬНІСТЬ РОБОТИ. На сьогоднішній день інформаційний пошук швидко стає основною формою доступу до інформації у інформаційному суспільстві.

З кожним роком збільшується об’єм доступних користувачу масивів текстової інформації, що сприяє більшій актуалізації задачі пошуку необхідних користувачу документів в таких масивах. Для вирішення подібних задач дуже часто використовуються різноманітні дуже елементарні програмні засоби - тематичні класифікатори, рубрикатори і т.д., які дозволяють шукати (автоматично або вручну) документи в невеличкій підмножині документної бази, яка відповідає зацікавленості користувача [1]. Але не завжди результат пошуку може відповідати поставленим завданням.

Сучасні моделі інформаційного пошуку не використовують знань, описаних у тезаурусах і онтологіях, а базуються на моделях тексту як набору слів, пропонуючи методи урахування частоти появи слів у реченні, тексті, наборі документів. Нерідко враховується кількість спільної появи слів [2].

До інформаційного пошуку можна віднести деякі завдання, які не підпадають під базове визначення. Коли ми говоримо “неструктуровані дані”, ми маємо на увазі дані, які не мають ясної, семантично очевидної й легко реалізованої на комп'ютері структури. Вони являють собою протилежність структурованим даним, канонічним прикладом яких є реляційні БД на зразок тих, які звичайно використовуються підприємствами для зберігання реєстрів продукції й персональних даних співробітників. У реальності зовсім “неструктурованих даних” практично не існує. Наприклад, звичайні текстові дані мають сховану структуру, характерну для природних мов.

Навіть якщо вимагати явної наявності структури, то більшість текстів її очевидно мають, оскільки в них є заголовки, абзаци й виноски, які звичайно представлені у тексті у вигляді явної розмітки (наприклад, у коді WEB-сторінок). Тому методи інформаційного пошуку використовуються також для “напівструктурованого” пошуку, наприклад для знаходження документа, у заголовку якого втримується слово Java, а в тілі – слово threading [3].

До інформаційного пошуку відносяться й такі завдання, як навігація користувачів по колекції документів і фільтрація документів, а також подальша обробка знайдених документів. Якщо є набір документів, то виникає завдання кластеризації, що полягає у визначенні найкращої сукупності документів за їх змістом. Це нагадує розміщення книг на полицях по темах.

Процес пошуку може відбуватися в досить великому документаційному сховищі. Буває, що навіть на конкретно сформований запит(саме користувачами), може бути досить неточний результат пошуку.

Це залежить від структурованості або неструктурованості даних. Неструктуровані дані становлять більшу частину інформації, з якою мають справу користувачі. Ці дані становлять не менш 90% всієї інформації, а 10 % – це структуровані, впорядковані дані.

Звідси завжди існує дилема: чому віддати перевагу – потужним обчислювальним процедурам, що опираються на відносно невеликі словникові системи з багатою граматичною й семантичною інформацією, або потужному декларативному компоненту при відносно простих процедурних засобах? Питання залишається відкритим.

Кожному зрозуміло, що навіть у величезних масивах інформації повинні працювати пошукові системи, які забезпечували б користувачу результат – швидкий і точний. Сучасне інформаційне суспільство використовує ряд аналітичних підходів до подання інформації для забезпечення її наступного пошуку [4]. Деякі базуються на теорії множин, інші – на елементах векторної алгебри. Обидва підходи ефективно реалізовуються в умовах практики [4].

Інформація достовірна, якщо вона показує реальне положення справ. Об'єктивна інформація завжди достовірна, але достовірна інформація може бути як об'єктивною, так і суб'єктивною. Достовірна інформація допомагає прийняти нам правильне рішення.

На жаль, достовірна інформація, як і будь-яка інша, піддається процесу старіння. Процес старіння додає купу проблем при процесах пошуку необхідних документів.

Старіння інформації полягає в зменшенні її цінності із часом. Старіння інформації забезпечує не сам час, а поява нової інформації, яка уточнює, доповнює або відкидає повністю або частково більш ранню. Фактом є те, що науково-технічна інформація старіє швидше, естетична (твори мистецтва) – повільніше. Із часом кількість інформації зростає, інформація накопичується, відбувається її систематизація, оцінка й узагальнення. На жаль, обсяги ростуть дуже швидко. Сьогодні - це терабайти текстових даних.

Моделі пошуку визначають міру відповідності між запитом та результатом. Взагалі, можна вивести загальну ідею: більший відсоток збігу між запитом та документом дає зрозуміти, що документ є більш відповідним до запиту користувача. Поняття “збіг” є багатоаспектним: включається термінологічна база, ключові поняття, співпадання предметних областей.

Одним з можливих варіантів вирішення проблем пошуку інформації можна назвати семантичний пошук. Він надає результат не тільки за заданими словами з запиту, але й за еквівалентними за сенсом [5].

Для ефективного семантичного пошуку необхідна інформація про предметну область, про властиві їй поняття та відношення між ними, а також про обмеження, які існують між відношеннями [5].

Таку інформацію прийнято називати онтологією. Онтологічна модель може бути використана як для повнотекстового пошуку, так і для окремої класифікації документів [5].

Мета роботи полягає в аналізі, дослідженні, покращенні деяких можливостей ефективного пошуку інформації у сукупності неструктурованих документів та у формулюванні вимог до моделі предметно-орієнтованої системи для ефективного семантичного пошуку.

МАТЕРІАЛ ТА РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ. Поняття онтології, яке було запозичене з філософії, зараз активно застосовується в штучному інтелекті й інформатиці. У філософії онтологія вивчає категорії буття, які існують або можуть існувати. У штучному інтелекті онтології згадуються в контексті з такими поняттями як концептуалізація, знання, подання знань, засновані на знаннях системи.

Було помічено, що при дослідженнях одні намагаються дати неформальні визначення, а інші описують онтології на основі понять і конструкцій логіки й математики. Але, незважаючи на те, що побудовано безліч різних онтологій і збільшується область їхнього застосування, дотепер немає точного визначення цього поняття стосовно до області штучного інтелекту [6].

Будь-яка база знань, система, заснована на знаннях, фіксується явно або неявно деякою концептуалізацією. Безліч об'єктів і відносини між ними відбиваються в словнику, у якому система, заснована на знаннях, представляє свої знання. Таким чином, вважається, що основу онтології становлять безлічі представлених у ній термінів. Втім, не тільки термінів. Як уже говорилося, в онтологічну сукупність включаються також відомості про предметні області, про області визначень і т.д.

Х.Такеда ставить онтології в центр проблеми організації знань, тому що в кожній області можуть існувати різні розуміння тих самих термінів. У цьому випадку онтологія використовується для структурування інформації, залишаючись посередником між людино-машинно-орієнтованим і машинно-машинно-орієнтованим рівнем подання інформації. Тоді онтологія визначається як “угода”, “контракт” про деяку область інтересів для досягнення певних цілей [7].

Трохи інший підхід декларує Н.Гуаріано. Для встановлення взаєморозуміння про знання, які представлені на деякій мові, зокрема логічній, на думку Н.Гуаріано, онтологія повинна характеризувати концептуалізацію, обмежуючи можливі значення предикатів і функцій. У цьому розумінні, онтологія - це логічна теорія, аксіоми якої обмежують інтерпретації нелогічних символів мови [8].

На думку Т.А.Гаврилової, онтологія – це структурна специфікація деякої предметної області, її формалізоване подання, яке включає словник (або імена) покажчиків на терміни предметної області та логічні зв'язки, які описують як терміни співвідносяться один з одним. Онтології забезпечують словник для представлення та обміну знаннями про деяку предметну область і безліч зв'язків, установлених між термінами в цьому словнику [9].

Важливість підходу, пов'язаного з онтологіями, обумовлена також тим, що знання, яке не описано і не тиражоване, в кінцевому рахунку стає застарілим і непотрібним. Навпаки, знання, яке поширюється є генератором нових знань [9].

Тепер спробуємо мислити неформально. Онтологія являє собою деякий опис погляду на світ стосовно до конкретної області інтересів. Цей опис складається з термінів і правил використання цих термінів, що обмежують їхні значення в рамках конкретної області. На формальному ж рівні, онтологія це система, що складається з набору понять і набору тверджень про ці поняття, на основі яких можна будувати класи, об'єкти, відносини, функції й теорії. Онтологія, як приклад загальної угоди про семантику області, сприяє встановленню коректних зв'язків між значеннями елементів області, тим самим, створюючи умови для їхнього спільного використання.

У загальному вигляді формальна модель онтології може бути описана таким кортежем [10]:

$$O = \{L, C, F, G, H, R, A\}, \quad (1)$$

де $L = L^C \cup L^R$ – це словник онтології, що містить набір лексичних одиниць (знаків) для понять L^C і набір знаків для відносин L^R;

C – набір понять онтології, причому для кожного поняття ($c \in C$) в онтології існує принаймні одне твердження;

F і G – функції посилань такі, що $F: F^{LC} \rightarrow 2^C$ та $G: F^{LR} \rightarrow 2^R$. Тобто F і G пов'язують набори лексичних одиниць $\{L_i\} \subset L$ з наборами понять і відносин, на які вони відповідно посилаються в даній онтології. При цьому одна лексична одиниця може посилатися на кілька понять або відносин і одне поняття чи відношення може посилатися на кілька лексичних одиниць. Інверсіями функцій посилань є F^{-1} та G^{-1};

H – фіксує таксономічний характер відносин (зв'язків), при якому поняття онтології пов'язані нерефлексивними, ациклічними, транзитивними відносинами $H \subset C \times C$. Вираз $H(C_1, C_2)$ означає, що поняття C_1 є підпоняттям C_2;

R – означає бінарний характер відносин між поняттями онтології, які фіксують пари $(D\ R)$ з $D, R \in C$ (область застосування/область значень);

A – набір аксіом онтології.

Більш просту модель предметної онтології O^d можна представити як впорядковану сукупність компонентів:

$$O^d = <C,R,F>, \quad (2)$$

де C – кінцева множина концептів (понять) предметної області, яку представляє предметна онтологія; R – кінцева множина відносин між концептами (поняттями) конкретної (заданої) прикладної сфери; F – кінцева множина функцій інтерпретації (аксіоматизація), заданих на концептах та/або відношеннях предметної онтології.

Істотними обмеженнями, які накладаються на множину C, є скінченність та умова відсутності порожності. Множини R та F можуть бути порожніми, що відповідає окремим типам предметної онтології.

Проведений аналіз існуючих підходів до використання онтологій дозволяє зрозуміти, що онтології можна застосовувати:

1) Як будівельні блоки компонентів баз знань;

2) Як окремі блоки впливу на процес пошуку;

3) Як схеми об'єктів в об'єктно-орієнтованих системах,

4) Як структурований глосарій взаємодіючих блоків системи;

5) Як словник для зв'язку між елементами,

6) Як спосіб визначення класів для програмних систем.

Процес побудування онтологій включає в себе наступні етапи [6]:

1. Фіксація знань стосовно предметної області, яка включає в себе:

– визначення основних понять та їх взаємовідносин у обраній предметній області;

– створення точних визначень (які не конфліктують між собою) для понять і відношень;

– визначення термінів, які пов'язані між собою відповідними відношеннями;

– остаточне погодження всіх вищеназваних етапів.

2. Кодування:

– процес поділу сукупності окремих термінів (які, звісно використовуються в онтології) на окремі класи понять;

– вибір або розробка спеціальної мови для представлення онтології;

– створення фіксованої концептуалізації на обраній мові представлення даних.

Онтології необхідно застосувати у ролі посередника між користувачем і процесом пошуку, між процесом пошуку і пошуковою системою. Подання про проектування онтологічної системи представимо за допомогою діаграм UML (рис. 1 та рис. 2).

Для побудови онтології потрібно формальне декларативне подання чітко організованих конструкцій, які містять у собі словник термінів тематичної області, опис визначень цих термінів, існуючі взаємозв'язки між ними, і взагалі – теоретично можливі й неможливі взаємозв'язки.

Було з'ясовано, що взаємодія з онтологією відбувається на наступних етапах:

1) Обробка та аналіз запиту;

2) Розкладання запиту на складові;

3) Перевірки інформації на старіння та відповідність.

4) Перевірка на старіння інформації, яка була обрана зі сховища даних.

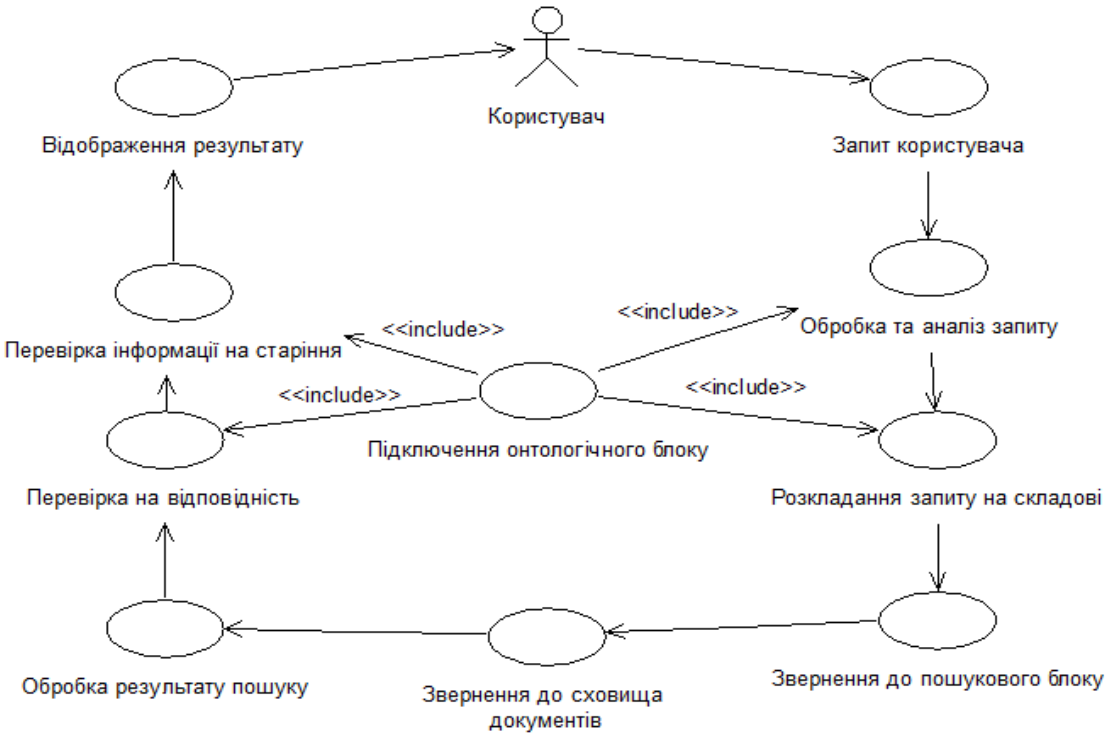

Рисунок 1 – Діаграма варіантів використання системи онтологічного пошуку

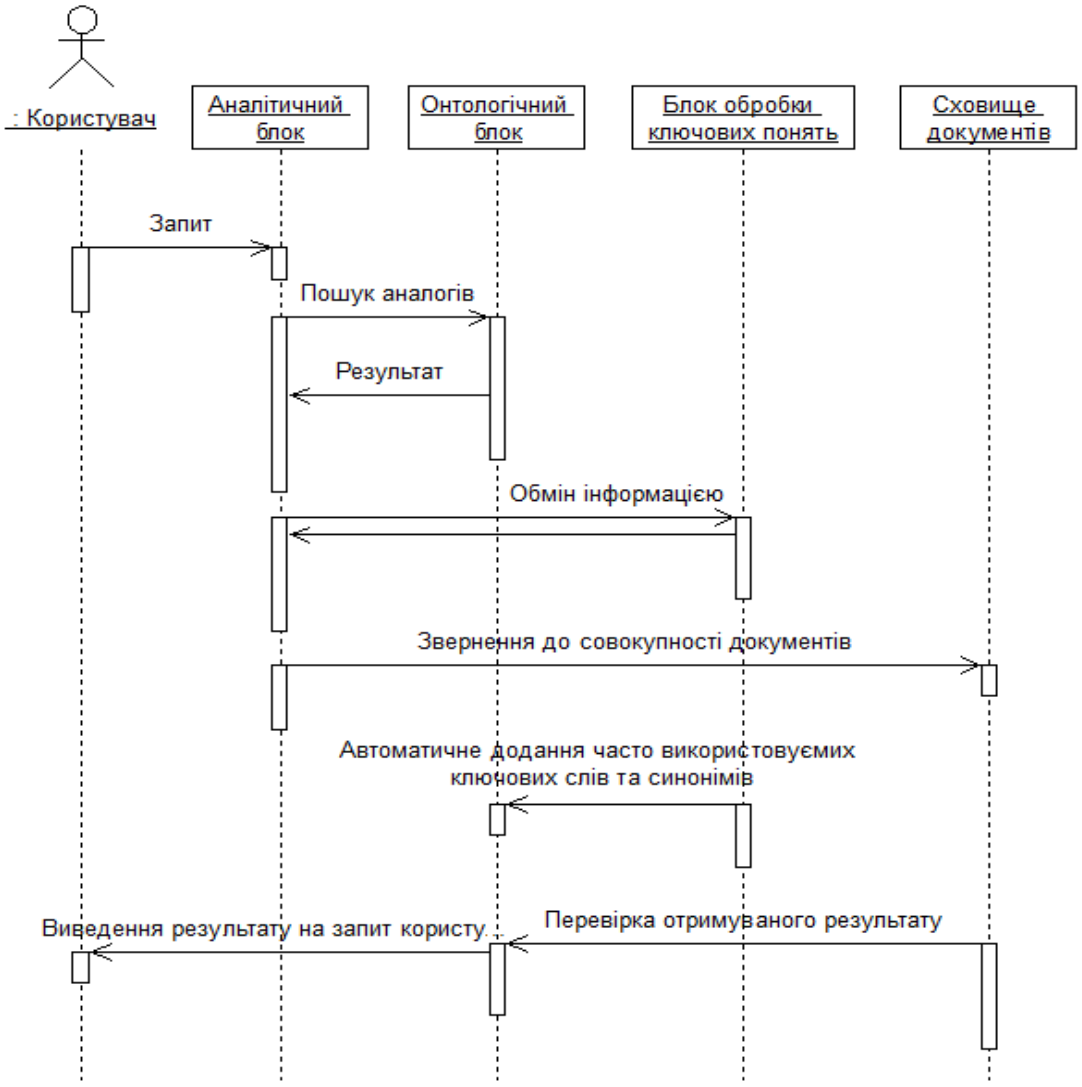

Рисунок 2 – Діаграма послідовності для варіанту використання “Запит”

Проблема полягає в тому, щоб зробити пошук динамічним і зручним для користувача. Для будь-якого типу запиту, що виникає в практичній діяльності, повинні бути знайдені адекватні знання в інформаційному просторі. При цьому мова для формулювання пошукової вимоги не повинна бути занадто складною.

Зокрема, спілкування користувача з пошуковою системою можна зробити більш простим, наблизивши мову запиту до природньої мови (але не "сленгової").

При такій організації пошуку на етапі формування образу релевантного документа з користувальницького запиту виділяються значеннєві структури: значимі слова й терміни предметної області.

Ці значеннєві структури потім використовуються для формування окремого пошукового образу під впливом онтологічної складової. На етапі побудови запиту до пошукової системи здійснюється виведення на онтології.

При цьому виконується перетворення запиту користувача в з'єднаний логічними зв'язками набір термінів і понять, які будуть використовуватися пошуковою системою.

Розглянемо детальніше деякі окремі компоненти цієї діаграми. При роботі з пошуковими системами користувач може зіштовхнутися з різними нестандартними ситуаціями.

Спробуємо змоделювати деякі з них:

1) Користувач може не знати як точно пишеться потрібний йому термін. Змоделюємо ситуації:

"Sydney" або "Sidney"

"Kyiv" або "Kiev"

"Dnipropetrovsk" або "Dniepropetrovsk"

У даному випадку треба зазначити, що подібна проблема характерна для документів, які заповнені англомовними термінами. Стосовно тих документів, які заповнені україномовними термінами, то проблема подібного роду в ідеальних умовах все-таки є більш рідкісною. Проблема може полягати тільки в рівні мовної підготовки працівника.

2) Користувач знає про існування декількох варіантів написання терміна й свідомо шукає документи, у яких є наявність хоча б одного із цих варіантів.

3) Користувач шукає документи, що містять варіант терміна, уніфікований у результаті стеммінга - процесу знаходження основи слова для заданого вихідного слова, але не впевнений, що пошукова система виконує цей стеммінг (наприклад, варіанти "критика" і "критичний" можна об'єднати в шаблоновому запиті – крити*).

4) Користувач не має впевненості у правильності відтворення слів або фраз іноземного походження.

Пропонується наступний підхід, який можна реалізувати на практиці. У онтологічному блоці майбутньої пошукової системи можна було б використати досить відомий принцип – запит з джокером.

Звісно, що можуть бути ситуації наступного виду:

word* або *word

У цьому випадку у онтологічному блоці реалізуються два типи запиту з джокером – з замикаючим джокером та провідним джокером.

Використання запиту з замикаючим джокером зручно у тих випадках, коли користувачу невідомо, як слово завершується. Наприклад:

color і colour

Запит буде мати вигляд типу colo*.

Введемо поняття множини термінів словника (з конкретним префіксом) та позначимо її як *W*. Помітимо, що символ * використаний тільки один раз у самому кінці пошукового рядка. Для обробки таких запитів пропонується використати дерево пошуку над словником: рухаючись по дереву словника вниз і по черзі переходячи по галузях, що відповідають буквам c, o, l, o, можна перебрати всю множину *W*-термінів словника із префіксом colo. На закінчення після проходження всіх етапів пошуку (від обробки та аналізу запиту і до перевірки інформації на старіння) та після $|W|$ переглядів стандартного інвертованого індексу (саме у онтологічному блоці), можна витягти з бази всі документи, які містять хоча б один термін з множини *W*.

Якщо ж символ * знаходиться на початку запиту, рекомендується застосування запиту з провідним джокером. Як приклад наведу одне з моїх власних спостережень. Іноземне ім'я Остін у різних базах даних на англійській мові пишеться по-різному.

Ostin або Austin.

Запит, який розглядається системою пошуку, буде мати наступний вигляд: *stin. У подальшому доцільно використати принцип роботи зі звичайними та зворотніми бінарними деревами (*B*-дерево). Кожен шлях від кореня до листа в *B*-дереві відповідає терміну в словнику, записаному у зворотньому порядку. Отже, термін у звичайному *B*-дереві буде виглядати наступним чином:

n-i-t-s-u-A або n-i-t-s-O

Проходження вниз по зворотньому *B*-дереву дозволить перерахувати у лексиконі словника всі терміни, які мають суфікс stin та витягти зі сховища даних всі документи, які містять хоча б один термін з множини *W*-термінів словника з суфіксом stin.

Однак не можна обмежуватися таким розумінням, що символ * буде стояти на початку або у кінці запиту. Приклад:

Dnipropetrovsk або Dniepropetrovsk

Запит, який розглядається системою пошуку, буде мати наступний вигляд:

Dn*propetrovsk

Фактично за допомогою звичайного бінарного дерева в сполученні зі зворотним *B*-деревом можна вирішити ще більш загальне (можна говорити не тільки про загальність, але також про ускладненість та нечіткість запиту) завдання: обробку запитів, у яких є один символ * (мається на увазі, що цей символ стоїть не в кінці і не на початку).

При цьому для перерахування введемо множину W-термінів, що містять префікс Dn і непустий суфікс, можна використати звичайне *B*-дерево, а для перерахування введеної множини *R*-термінів, що закінчуються суфіксом propetrovsk, можна використати зворотне *B*-дерево. Далі пропонується використати операцію перетину множин *W* та *R*.

$$W \cap R = A, \qquad (3)$$

де множина A – нехай це буде результат наступного типу: всі терміни, які починаються з префіксу Dn та закінчуються суфіксом propetrovsk.

Треба звернути увагу на виключення, при яких подібний підхід може не спрацювати. Як приклад, запит для слів "казка", "заноза" може надати при подібному аналізі неправильний результат. На закінчення за допомогою стандартного інвертованого індексу можна одержати всі документи, що містять хоча б один термін з перетинання. Отже, за допомогою нормального й зворотнього B-дерев можна обробляти запити, що містять один символ *.

У блоці на перевірку старіння інформації повинен виконуватися наступний принцип. Старіння інформації в різні моменту часу формально можна представити у вигляді:

$$S_{t'}(t) = M_t - M_{t'} \quad (4)$$

де $S_{t'}(t)$ – старіння інформації (даних) зареєстрованих в момент часу t' на момент t; $M_{t'}, M_t$ – значення даних по відношенню до досягнення мети в момент часу t, t'.

Додатково введемо ще одну перевірку:

$$d = |t' - t| \quad (5)$$

де d - діапазон часу старіння; t, t' - початкова та кінцева дати.

Після цього можна порівнювати отримане значення з кількістю років або місяців (мається на увазі значення, при якому інформація вважається застарілою) і на основі цього робити висновок про те, чи сильно застарілою є інформація або документ.

Обов'язковим у системі пошуку повинен бути процес "самонавчання" системи. Вважається, що цей процес дозволить ліквідувати ситуації з термінами або назвами, які записуються некоректно у базі даних. Подібні приклади були наведені раніше: Ostin або Austin, Kyiv або Kiev, Dnipropetrovsk або Dniepropetrovsk.

Передбачається, що у процесі "самонавчання" повинно проводитися побудування правил або функцій, диференційованих по ситуаціям, якими система повинна користуватися при виникненні незнайомих або нестандартних ситуацій. З узагальнених правил автоматично буде формуватися словник термінів, правил та умов.

ВИСНОВКИ. 1) Онтології:

– можуть і могли б вирішувати проблему подання знань для виводу інформації, що релевантна запиту користувача;

– дозволили б займатися фільтрацією й класифікацією інформації;

– дозволили б займатися створенням загальної термінології для програмних агентів і користувачів;

– допомогли б захистити сховища інформації від тотального переповнення й виникнення помилок;

– вважаються одним з засобів вирішення питання старіння інформації.

2) Були запропоновані деякі підходи для вирішення нестандартних ситуацій, які виникають при виконанні процедур пошуку інформації. Робота у цьому напрямку продовжується.

3) Передбачається, що модель системи повинна мати наступні властивості:

– цілісність (система буде розглядається як єдине ціле, що складається з взаємодіючих модулів, можливо неоднорідних, але одночасно та "точково" сумісних між собою);

– зв'язність (наявність істотних стійких зв'язків між елементами та їх властивостями, причому з системних позицій значення мають не будь-які, а лише істотні зв'язки, які визначають інтегративні властивості системи);

– організованість (наявність певної структурної й функціональної організації, що забезпечує зниження ентропії системи в порівнянні з ентропією системоутворюючих факторів, що визначають можливість створення системи, до яких можуть відноситися: число елементів системи, число істотних зв'язків, якими може володіти кожний елемент, і т.д.);

– інтегративність (наявність якостей, властивих системі в цілому, але не властивих жодному з її елементів окремо, тобто властивості системи хоча й залежать від властивостей елементів, але не визначаються ними абсолютно повністю);

– мобільність (у даному випадку було складно підібрати термін, поки що під цим будемо розуміти можливість швидкої перебудови моделі і системи під виникаючі обставини; обов'язковою умовою можна додати процес "самонавчання" системи).

4) Продовжується підбір алгоритмів та методів для окремих компонентів системи, що розробляється. Продовжується аналіз та підбір кращих методів та способів для реалізації в блоках системи пошуку.

5) Розроблено спрощений тестовий варіант програмного модуля, який виконує функції онтологічного блоку та пошуку в текстових документах.

ЛІТЕРАТУРА

1. Шатовская Т., Каменева И. Интегрированный подход текстовой кластеризации для неструктурированных документов // "Internet – Education – Science": материалы 6-й Международной конференции, 7–11 октября. – Винница, Украина. – 2008. – С. 504–506.

2. Лукашевич Н.В. Тезаурусы в задачах информационного поиска – М.: Издательство Московского университета, 2011. – 512 с.

3. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск: пер. с англ. – М: ООО "И.Д. Вильямс", 2011. – 528 с.

4. Аникин В.М. Аналитические модели детерминированного хаоса – М.: Физматлит, 2007. – 328 с.

5. Захарова И.В. Об одном подходе к реализации семантического поиска документов в электронных библиотеках // Вестник УГАТУ. – 2009. – Т.12, №1(30): Серия "Управление, вычислительная техника, информатика". – С. 133–138.

6. Россеева О., Загорулько Ю. Организация эффективного поиска на основе онтологий // Труды международного семинара "Диалог'2001" по компьютерной лингвистике и ее приложениям. – Т.2. – Аксаково. – 2001. – С. 333–342.

7. Takeda H., Takaai M., Nishida T. Collaborative development and Use of Ontologies for Design // Proceedings of the Tenth International IFIP WG 5.2/5.3 Conference Prolamat 98, September 9–12. – Trento, Italy. – 1998. – pp. 77-89.

8. Guarino N., Masolo C., Vetere G. OntoSeek: Content-BasedAccess to the Web. // IEEE Intelligent Systems, May/June. – Italy. – 1999. – pp.70–80.

9. Гаврилова Т.А. Онтологический подход к управлению знаниями при разработке корпоративных информационных систем // Новости искусственного интеллекта. – №2. – 2003. – С. 24–30.

10. Тузовский А.Ф., Чириков С.В., Ямпольский В.З. Системы управления знаниями (методы и технологии) – Томск: Изд-во НТЛ, 2007. – 260 с.

DOMAIN ONTOLOGIES APPLICATION IN KNOWLEDGE MANAGEMENT TO IMPROVE THE SEMANTIC SEARCH EFFECTIVENESS

D. Kostenko, B. Moroz, V. Kostenko

University of Customs and Financy

vul. Dzerzhinskogo, 2/4 Dnipropetrovsk, 49000, Ukraine. E-mail: denis_and_ko@ukr.net

Purposes. The analysis of this research is the improvement of several possibilities of effective information search in combined unstructured documents and the formulation of requirements to the model of object-oriented system for effective semantic search. **Methodology** is based on the approach of creating systems, which is connected with ontologies for forming the particular image search influenced the ontological component. **Originality.** Firstly, some methods of improvement of effectiveness in interaction between the user and search engine with using ontological method in view of non-standard situations, which can be appeared in time of processing unstructured documents, were offered. **Practical value** is in the working-out of the quality search engines for solving the problem of improvement some possibilities of effective information search in total document files of unstructured documents. **Results.** Several methods for solving non-standard situations, which can be appeared in carrying out the procedures of search, are offered. The properties and functions, which this model must have, are formulated. The work-out of semantic search system is being realized. The question of ontology's using, including the decision of problem of information aging, is being considered. References 10, figures 2.

Key words: data, joker request, information, ontologies, stemming, documents searching, semantic searching.

REFERENCES

1. Shatovskaya, T., Cameneva, I. (2008), "Computer-integrated approach of text clusterization for unstructured documents", *"Internet – Education – Science": materialy 6 Mezhdunarodnoi conferencii* ["Internet – Education – Science": materials of the 6th International conference], Vinnitca, Ukraine, October 7-11, pp. 504–506.

2. Lukashevich, N.V. (2011), *Thezaurusy v zadachah informazionnogo poiska* [Thesauri in information retrieval], Moscow University publishing house, Moscow, Russia.

3. Manning, C.D., Raghavan, P., Schütze, H. (2011), *Vvedenie v informazionnyi poisk* [Introduction to Information Retrieval], Williams, Moscow, Russia.

4. Anikin, V.M. (2007), *Analiticheskie modeli determinirovannogo haosa* [Analytical models of determined chaos], Phismatlit, Moscow, Russia.

5. Zakharova, I.V. (2009), "About one method to realization of semantic search of documents in electronic libraries", *Vestnik USATU*, Vol.12, no.1(30), pp. 133–138.

6. Rosseeva, O., Zagorulko, Y. (2001), "Effective search organization on the basis of ontologies", *Trudy mezhdunarodnogo seminara Dialog'2001 po komputernoi lingvistike i prilogheniam* [Proceedings of the international seminar "Dialogue'2001" of computer linguistics and its applications], Aksakovo,2001, Vol. 2, pp. 333–342.

7. Takeda, H., Takaai, M., Nishida, T. (1998), "Collaborative development and Use of Ontologies for Design", *Proceedings of the Tenth International IFIP WG 5.2/5.3 Conference PROLAMAT 98*, Trento, Italy, September 9–12, pp. 77–89.

8. Guarino, N., Masolo, C., Vetere, G. (1999), "OntoSeek: Content-Based Access to the Web", *IEEE Intelligent Systems*, May/June, Italy, pp. 70–80.

9. Gavrilova ,T.A. (2003), "Ontological approach to knowledge management in the development of corporate information systems", *Novosti iskusstvennogo intellekta*, no.2, pp. 24–30.

10. Tuzovskiy, A.F., Chirikov, S.V., Yampolskiy, V.Z. (2007), *Sistemy upravlenia znaniyami (metody i technologii)* [Knowledge management system (methods, technologies)], NTL, Tomsk, Russia.

Стаття надійшла 17.12.2015.