УДК 007

АНАЛІЗ СУЧАСНИХ ПОШУКОВИХ СИСТЕМ НА ПРЕДМЕТ ЇХ ПРИДАТНОСТІ ДЛЯ ПОШУКУ ТА ВИЛУЧЕННЯ ІНФОРМАЦІЇ ПРО ОДНОТИПНІ ОБ'ЄКТИ З WEB-ПРОСТОРУ

М.С. Бурматова, М.В. Оленін

Національний авіаційний університет, 03680, Київ, проспект Космонавта Комарова, 1,
тел. 067 3210723, mburmatova@aidoss.com

Проаналізовані основні сучасні світові та українські пошукові системи на предмет їх придатності для вилучення інформації про однотипні об'єкти з простору Web. Системи проаналізовані за наступними критеріями: ступінь автоматизованості наповнення бази даних, кількість джерел, що охоплюються системами, можливості розширеного пошуку, якість вмісту та представлення результату.

The article contains the analyzes of modern international and Ukrainian search engines on their suitability for information about listed objects retrieval. The engines are analyzed based on the following criteria: the level of databases filling automation, the amount of sources used for data retrieval, the possibilities of advanced search, the quality of the results' content and the quality of the results' presentation.

Вступ

Протягом останніх десятиліть інформаційні технологій (IT) пройшли шлях розвитку від великих лампових обчислювальних машин для наукових установ і аналітичних центрів, до доступних пересічній людині портативних комп'ютерів із надзвичайною для таких маленьких пристроїв обчислювальною потужністю. Ці технології зробили доступною для користувачів величезну кількість джерел інформації для пошуку та вилучення з них інформації. Чи не найкраще динаміку розвитку IT ми можемо спостерігати через розвиток Web-простору: якщо у 2000 році у світі нараховувалося близько 300 мільйонів користувачів Інтернет, то в 2004 їх кількість становила 700 мільйонів, а у 2008 вона зросла до 1 мільярда. Завдяки розвитку новітніх технологій WEB простір став найбільшим з відкритих джерел інформації сучасності. За статистикою 2009 року 93 % нової інформації світу зберігається в електронному вигляді і є частково доступною для користувачів через Web-простір.

Зі зростанням обсягів інформації, доступної через Web, зросла необхідність розвитку та вдосконалення засобів вилучення й обробки цієї інформації. Web-простір за своїм характером є мало організованою розподіленою системою та не має чіткої організованої структури, тому централізованої системи для пошуку й обробки інформації у Web також не існує. На даний момент для пошуку та вилучення інформації з Web використовуються автоматизовані пошукові системи, які розрізняються методами обробки пошукового запиту, методами обробки джерел інформації та галузями застосування.

Далі в роботі різні типи пошукових систем та аспекти їх застосування для пошуку однотипних об'єктів розглянуті детальніше. Однотипними об'єктами вважаються такі, що мають визначений набір атрибутів і обмежений певними величинами набір значень цих атрибутів. Завдяки цьому ці об'єкти можна віднести до певних класів. До класів однотипних об'єктів можна віднести такі: літаки, автомобілі, об'єкти нерухомості, комп'ютери, mp3-плеєри і т.д. Всі такі об'єкти мають певний унікальний визначений обмежений набір характеристик, що є обумовленим класом об'єкта. Задача пошуку однотипних об'єктів за характеристиками, а не за назвою об'єкта, є доволі поширеною задачею у Web. Протягом останніх десяти років надзвичайного зріс відсоток електронної комерції, дуже великого поширення набули Internet-магазини, електронні агенції нерухомості, окрім того часто виробники публікують характеристики свої товарів на власних Web-сайтах.

Огляд сучасних пошукових систем

Розглянемо відкриті та частково відкриті пошукові системи [1], що в різних ступенях дозволяють виконувати задачу пошуку та вилучення інформації про однотипні об'єкти з Web-простору.

Всі пошукові системи, обрані для аналізу, оцінюються за трьома наступними основними характеристиками:

1) ступінь автоматизованості наповнення бази даних пошукової системи – система є повністю автоматизованою, частково автоматизованою чи вимагає ручного постачання інформації в базу даних (БД) системи;

2) якість пошуку однотипних об'єктів – наскільки система дозволяє задавати детальний пошуковий запит за значеннями атрибутів об'єктів, як реагує на зміну запиту, наскільки результат пошуку відповідає очікуванням користувача;

3) метод і якість представлення інформації – наскільки наглядно представлена інформація.

Повнотекстові пошукові системи такі як Google, Yahoo, Ask.com, Meta.ua, Yandex, Rambler, Bing, Alta vista – це он-лайн служби, які дозволяють здійснювати пошук інформації на Web-сторінках, відповідної пошуковому запиту користувача, що представлений природномовним текстом.

Наповнення БД цих систем відбувається автоматизованим шляхом за допомогою індексації сторінок пошуковими роботами. Системи спеціалізуються на пошуку саме Web-сторінок, що містять найбільш релевантну інформацію [2]. Скажімо користувач шукає " 3 кімнатна квартира з балконом у житомирі", то найбільш релевантними сторінками будуть сторінки, що містять найбільшу кількість зі слів "кімнатна", "квартира", "балконом", "житомир" або всі слова, які розташовані найближче одне до одного з певними морфологічними правилами, індивідуальними для кожної мови та кожної пошукової системи. Для пошукових систем усі перераховані слова мають коефіцієнт релевантності, який відповідає їх морфологічному розташуванню у самому запиті. Важливість слова в семантичному сенсі може бути визначена неправильно, бо залежить від семантичного контексту запиту, відповідно пошукові системи надають як результат на вищезгаданий пошуковий запит, посилання на сторінки з описом об'єктів нерухомості з балконом, але, наприклад, не в Житомирі, квартир, але не трикімнатних, або на сторінки з описом декількох об'єктів нерухомості, один з яких у Житомирі, інший є трикімнатним будинком, третій є квартирою з балконом.

Повнотекстові пошукові системи представляють результати пошуку у вигляді посилання на сторінку, її заголовка та певного узагальнення сторінки (часто з демонстрацією відповідності сторінки пошуковому запиту шляхом відображення того, як слова з пошукового запиту вживаються на сторінці). Ці системи мають свої методи відсіювання інформаційного спаму [3].

У зв'язку зі своєю універсальністю та намаганням охопити якомога більше інформаційне поле повнотекстові пошукові системи мають наступні риси:

1) повністю автоматизований метод заповнення БД;
2) низька якість результатів при пошуку однотипних об'єктів;
3) відсутність прямої можливості вказати значення атрибутів пошукового об'єкта;
4) часте ігнорування пошуковою системою частини атрибутів, частини значень атрибутів або змішування атрибутів і їх значень при завданні пошукового запиту природномовним текстом;
5) представлення результатів у вигляді посилання на Web-сторінки, які можуть містити пошуковий однотипний об'єкт або декілька однотипних об'єктів.

Мета-пошукові системи. Мета-пошукова система – це пошуковий інструмент, який надсилає запит користувача одночасно на декілька пошукових систем (ПС), каталогів і іноді в невидиме (приховане) павутиння – збірку он-лайн інформації, не проіндексованої традиційними пошуковими системами.

Зібравши результати мета-пошукова система (МПС) видаляє дубльовані посилання та відповідно до свого алгоритму об'єднує / ранжує результати в загальному списку.

Розглянемо детальніше основні мета-пошукові системи.

Clusty. Мета-пошукова система, що має кластеризуючий двигун, який використовує технологію Vivisimo.

Пошук здійснюється за повнотекстовим користувацьким запитом і кластеризує результати пошуку за пов'язаними з пошуковим запитом кластерами.

Заповнення бази даних здійснюється під час індексування сторінок (відповідно є автоматизованим).

Пошуковий запит є аналогічним звичайній повнотекстовій пошуковій системі. Результати видаються у вигляді Веб-сторінки з посиланнями на знайдені сторінки, та кластерами, які можна обрати для звуження результатів пошуку [4].

Незважаючи на те, що система надає кращі результати в порівнянні зі стандартними повнотекстовими пошуковими системами та надає можливість користувачу «звузити» результати, вона все ще представляє результати пошуку у вигляді списку посилань на Web-сторінки, які можуть містити більш ніж один об'єкт. Крім того серед кластерів можуть бути далеко не всі ті, за якими має бути здійснене відповідне звуження запиту.

Dogpile. Мета-пошукова система, що здійснює пошук водночас декількома пошуковими системами, фільтрує результати, знаходить дублікати й об'єднує їх для представлення користувачеві. Використовує Google, Yahoo! Search, Live Search, Ask.com, About, MIVA, LookSmart та інші. Має ті ж самі недоліки що й звичайні повнотекстові пошукові системи [5].

Excite. Інтернет-портал дозволяє здійснювати як звичайний повнотекстовий пошук, так і спеціалізований пошук. У галузі вилучення інформації про однотипні об'єкти діє Excite Shopping, що дозволяє шукати однотипні об'єкти (в даному разі товари) за допомогою повнотекстового пошукового запиту. Директорія об'єктів збудована вручну, індексація нових товарів здійснюється лише у невеликому списку акредитованих продавців (Web-сайтів) за допомогою методів наданих самими продавцями. Результатом пошуку є список посилань на об'єкти в акредитованих продавців [6].

Excite Shopping надає якісну інформацію, але водночас не охоплює навіть задовільної кількості джерел інформації та не може здійснювати пошук у сирих даних, якими є більшість даних у Web.

Info.com. Мета-пошукова система здійснює пошук водночас декількома пошуковими системами, фільтрує результати, знаходить дублікати й об'єднує їх для представлення користувачеві. Використовує Google, Yahoo!, Bing.com, Ask.com, About, Open Directory, LookSmart та інші. Має ті ж самі недоліки що й звичайні повнотекстові пошукові системи.

Ixquick. Мета-пошукова система здійснює пошук водночас декількома пошуковими системами, фільтрує результати, знаходить дублікати й об'єднує їх для представлення користувачеві. Використовує 11 основних пошукових систем. Має ті ж самі недоліки що й звичайні повнотекстові пошукові системи.

Mamma. Мета-пошукова система здійснює пошук водночас декількома пошуковими системами, фільтрує результати, знаходить дублікати й об'єднує їх для представлення користувачеві. Має ті ж самі недоліки що й звичайні повнотекстові пошукові системи.

MetaCrawler. Мета-пошукова система здійснює пошук водночас декількома пошуковими системами, фільтрує результати, знаходить дублікати й об'єднує їх для представлення користувачеві. Використовує Google, Yahoo!, Live Search, Ask.com, About.com, MIVA, LookSmart та інші. Має ті ж самі недоліки що й звичайні повнотекстові пошукові системи.

MetaLib. Мета-пошукова система здійснює пошук у прихованому павутинні (бібліотечних каталогах, журнальних статтях, електронних і паперових виданнях). Пошук здійснюється стандартним пошуковим запитом у відповідному ресурсі. Система має ті ж самі недоліки що й звичайні повнотекстові пошукові системи.

Myriad Search. Мета-пошукова система здійснює пошук водночас декількома пошуковими системами, фільтрує результати, знаходить дублікати й об'єднує їх для представлення користувачеві. Використовує Ask Jeeves, Google, MSN, and Yahoo!. Має ті ж самі недоліки що ц звичайні повнотекстові пошукові системи.

Surfwax. Набір інструментів для пошуку, збереження та розповсюдження інформації в мережі Internet.

Для вилучення інформації використовує семантичний метод пошуку з використанням семантично-значущих ключів. як вхідна інформація використовується повнотекстовий пошуковий запит, що обробляється системою для вилучення семантичних ключів, за якими ведеться пошук [12].

Результати пошуку надаються у вигляді списку заголовків Web сторінок.

З усіх попередньо розглянутих систем ця система завдяки семантичному механізму надає найкращі результати пошуку через попередню обробку користувацького запиту і своєрідного вилучення атрибутів та їх значень, однак результати містять багато помилок і є посиланнями на Web-сторінки, які можуть містити більше одного об'єкта.

Turbo10.com. Це пошукова система в більше ніж 800 базах даних оголошень.

Пошуковий запит задається у вигляді природномовного тексту, результати надаються у вигляді посилань на Web сторінки та короткого кешу. Система має ті самі недоліки що й універсальні повнотекстові пошукові системи.

WebCrawler. Мета-пошукова система здійснює пошук водночас декількома пошуковими системами, фільтрує результати, знаходить дублікати й об'єднує їх для представлення користувачеві. Використовує Google, Yahoo!, Windows Live, Ask Smart та інші. Має ті ж самі недоліки що й звичайні повнотекстові пошукові системи.

Спеціалізовані пошукові системи. Спеціалізована пошукова система здійснює пошук інформації у певній предметній галузі або у певному сегменті Web-простору. Спеціалізована пошукова система надає можливість задати значно більш детальний пошуковий запит ніж звичайна повнотекстова пошукова система завдяки наявності заповнюваних атрибутів шукомих об'єктів. Атрибути зазвичай є дефолтними й їх кількість не розширюється. Розглянемо детальніше основні спеціалізовані пошукові системи.

ForSaleByOwner.com, Realtor.com, Rightmove, Zillow, Trulia. Пошукові системи об'єктів нерухомості. Надають широкі можливості пошуку з зазначенням пошукового запиту та значень основних атрибутів однотипних об'єктів (у даному випадку об'єктів нерухомості).

Заповнення бази даних системи здійснюється вручну (шляхом імпорту інформації для кожного окремого об'єкта нерухомості або списка об'єктів нерухомості).

Результати надаються у вигляді списка посилань на детальний опис кожного зі знайдених за зазначеними критеріями пошуку об'єктів [13].

Home.co.uk. Пошукова система об'єктів нерухомості. Надає широкі можливості пошуку з зазначенням пошукового запиту та значень основних атрибутів однотипних об'єктів (в даному разі об'єктів нерухомості).

Заповнення бази даних системи здійснюється шляхом імпорту з сертифікованих сайтів.

Результати надаються у вигляді списку посилань на детальний опис кожного зі знайдених за зазначеними критеріями пошуку об'єктів.

Enormo. Найбільша пошукова система нерухомості світу. Надає широкі можливості пошуку з зазначенням пошукового запиту та значень основних атрибутів однотипних об'єктів (у даному випадку об'єктів нерухомості).

Система має власний кроулер та систему вилучення інформації з Web. Має власні синтаксичні методи розпізнавання об'єктів нерухомості, які потребують доповнення у випадку зміни структури представлення інформації на сайті, який попередньо оброблявся, створення нових методів при обробці нових сайтів.

Результати надаються у вигляді списку посилань на детальний опис кожного зі знайдених за зазначеними критеріями пошуку об'єктів.

Google Product Search, Kelkoo, MSN Shopping, PriceGrabber.com, PriceRunner, Shopzilla, TheFind, Hotline.ua. Пошукові системи товарів. Надають широкі можливості пошуку з зазначенням пошукового запиту та значень основних атрибутів однотипних об'єктів (у даному випадку товарів).

Заповнення бази даних системи здійснюється шляхом імпорту з сертифікованих сайтів.

Результати надаються у вигляді списку посилань на детальний опис кожного зі знайдених за зазначеними критеріями пошуку об'єктів.

MySimon, Nextag, Shopping.com, Shopwiki, Price.ua, Nadavi.com.ua, Ekatalog.com.ua. Пошукові системи товарів. Надають певні можливості пошуку з зазначенням пошукового запиту та значень невеликої кількості атрибутів однотипних об'єктів (у даному випадку товарів).

Заповнення бази даних системи здійснюється шляхом імпорту з сертифікованих сайтів.

Результати надаються у вигляді списку посилань на детальний опис кожного зі знайдених за вказаними критеріями пошуку об'єктів.

Ebay. Он-лайн майданчик для проведення аукціонів і торгівельний Web-сайт, на якому приватні та юридичні особи здійснюють продаж та купівлю різноманітних товарів та послуг.

Ebay містить власну пошукову систему, що дозволяє здійснювати пошук з зазначенням пошукового запиту та значень невеликої кількості атрибутів пошукових об'єктів [14].

Заповнення бази даних системи здійснюється вручну (шляхом імпорту інформації для кожного окремого об'єкта нерухомості або списка об'єктів нерухомості).

Результати надаються у вигляді списка посилань на детальний опис кожного зі знайдених за вказаними критеріями пошуку об'єктів.

Автоматизовані пошукові системи та дослідження в галузі вилучення інформації про однотипні об'єкти з використанням технологій data mining

Dimo. Набір засобів для вилучення інформації про об'єкти нерухомості з Web. Являє собою стенделоун програму, що встановлюється на комп'ютері людини-оператора з можливістю завантаження результатів пошуку на Web-сайт dimo.com.ua.

Пошук об'єктів здійснюється шляхом надання розширеного атрибутами об'єктів та їх значеннями пошукового запиту. Результати надаються у вигляді списку посилань на детальний опис кожного зі знайдених за зазначеними критеріями пошуку об'єктів.

Заповнення бази даних є частково-автоматизованим процесом і вимагає попереднього створення набору правил вилучення інформації для кожного джерела інформації (Web-сайту), з якого відбудеться вилучення інформації. Набір правил базується на подібності html-структури опису кожного окремого об'єкта на Web-сайті або у розділі Web-сайту. Правила вилучення інформації базуються на тому, що оголошення окремого атрибуту об'єкту завжди розпочинається і закінчується певними послідовностями символів або наборами послідовностей, опис кожного окремого об'єкта також розпочинається певною послідовністю символів або набором послідовностей, наступна сторінка опису об'єктів і посилання на детальний опис об'єкта визначаються таким же чином. Наприклад, опис кожного окремого об'єкта нерухомості на сторінці може розпочинатися тегом <TR>, тип нерухомості розпочинається тегом <TD> і закінчується тегом </TD>, посилання на детальний опис міститься між <A href=' та '>, а місце розташування об'єкта визначається послідовністю двох тегів <TD> і закінчується тегом </TD>. Всі подібні послідовності (спочатку їх наявність \ відстуність) має визначити людина-оператор системи для кожного окремого Web-сайту, що є достатньо трудомісткою та монотонною роботою. Зазвичай у разі навіть невеликої зміни дизайну Web-сайту означення правил для нього доводиться переробляти.

Sunny. Спеціалізована автоматизована пошукова система розроблена автором статті і призначена для вилучення інформації про однотипні об'єкти з простору Web. Застосовується у предметній області об'єктів нерухомості.

Система є модульною і складається з трьох наступних основних компонентів:

1) мета-пошукова система для знаходження Web-сторінок з інформацією про однотипні об'єкти;
2) багатокрокова система вилучення інформації про однотипні об'єкти в бази даних;
3) інтерфейси для доступу (адміністратора, користувача).

Процес вилучення інформації в системі є автоматизованим і здійснюється мета-пошуковою системою, що здійснює пошук за допомогою таких пошукових систем як Google, Yahoo! Search, Live Search, Ask.com, About, Look Smart, Meta.ua, Nextag, Shopping.com, Shopwiki, Price.ua, Nadavi.com.ua, Ekatalog.com.ua. Набір пошукових систем, які використовуються для пошуку, є розширюваним і крім побажань користувача залежить від мови та природи запиту. Після знаходження необхідних посилань на Web-сторінки (одразу після знаходження або за розкладом, в залежності від налаштувань системи) система вилучає інформацію про об'єкти (значення їх атрибутів) в Базу даних системи за допомогою послідовного алгоритма перетворення Web-сторінки (трикрокове перетворення документа, що включає структурне, синтаксичне і семантичне перетворення) з застосуванням методів Data Mining (дерева рішень Random Forest). Кінцевий користувач має можливість пошуку однотипних об'єктів за їх класами та значеннями атрибутів об'єктів через користувацький інтерфейс.

Висновки

Результати аналітичного огляду інструментів вилучення інформації про однотипні об'єкти з Web представлені в таблиці.

Таблиця. Порівняння рис різноманітних пошукових систем при їх застосуванні для вилучення інформації про однотипні об'єкти

Назва пошукової системи / дослідження	Автоматизоване вилучення інформації в БД системи	Необмежена кількість джерел (закриті не враховуються)	Можливість задати атрибути об'єктів у пошуковому запиті	Результат переважно семантично відповідає запиту	Структурований результат	Коментар
Повнотекстові пошукові системи	Так	Так	Ні	Частково	Ні	
Clusty	Так	Так	Ні	Частково	Ні	
Dogpile, Info.com, Ixquick, Mamma, MetaCrawler, Myriad Search, WebCrawler	Так	Так	Ні	Частково	Ні	
Excite Shopping	Так	Ні	Ні	Так	Так	
MetaLib	Так	Ні	Ні	Частково	Частково	
Surfwax	Так	Так	Так	Частково	Ні	
Turbo10.com	Так	Ні	Ні	Частково	Ні	
ForSaleByOwner.com, Realtor.com, Rightmove, Zillow, Trulia	Так	Ні	Так	Так	Так	
Home.co.uk	Так	Ні	Так	Так	Так	
Enormo	Частково	Так	Так	Так	Так	
Google Product Search, Kelkoo, MSN Shopping, PriceGrabber.com, PriceRunner, Shopzilla, TheFind, Hotline.ua	Так	Ні	Так	Так	Так	
MySimon, Nextag, Shopping.com, Shopwiki, Price.ua, Nadavi.com.ua, Ekatalog.com.ua	Так	Ні	Частково	Так	Так	
Ebay	Так	Ні	Частково	Так	Так	
Dimo	Частково	Так	Так	Так	Так	
Sunny	Так	Так	Так	Так	Так	

Таким чином, можна зробити наступні висновки: на даний момент у Web-просторі не існує жодної автоматизованої пошукової системи, що є альтернативою системі Sunny та повністю відповідає поставленим вимогам автоматизованості наповнення БД з динамічно змінюваних джерел, не фіксованості набору джерел інформації, можливості завдання атрибутів об'єктів у пошуковому запиті, семантичній відповідності між результатом пошуку і поставленим запитом, структурованості результату.

1. Список основних пошукових систем світу [Електронний ресурс]. URL: http://thesearchenginelist.com/index.html
2. Пошукова система Google [Електронний ресурс]. URL: http://www.google.com/
3. Пошукова система Rambler [Електронний ресурс]. URL: http://www.rambler.ru/
4. Мета-пошукова система Clusty [Електронний ресурс]. URL: http://www.clusty.com/
5. Мета-пошукова система Dogpile [Електронний ресурс]. URL: http://www.dogpile.com/
6. Мета-пошукова система Excite [Електронний ресурс]. URL: http://www.excite.com/
7. Мета-пошукова система Ixquick [Електронний ресурс]. URL: http://www.ixquick.com/
8. Мета-пошукова система Mamma [Електронний ресурс]. URL: http://www.mamma.com/
9. Мета-пошукова система MetaCrawler [Електронний ресурс]. URL: http://www.metacrawler.com/
10. Мета-пошукова система MetaLib [Електронний ресурс]. URL: http://www.exlibrisgroup.com/category/MetaLibOverview
11. Мета-пошукова система [Електронний ресурс]. URL: http://www.myriadsearch.com/
12. Мета-пошукова система [Електронний ресурс]. URL: http://www.surfwax.com/
13. Спеціалізована пошукова система Trulia [Електронний ресурс]. URL: http://www.trulia.com/
14. Міжнародний аукціон eBay [Електронний ресурс]. URL: http://ebay.co.uk/