УДК 681.3

ПІДХІД ДО ОРГАНІЗАЦІЇ ПОШУКУ ІНФОРМАЦІЇ В РІЗНОРІДНИХ КОРПОРАТИВНИХ ДЖЕРЕЛАХ

В. Дерецький, М. Богданова, С. Ремарович

Інститут програмних систем НАН України,
03187, Київ-187, проспект Академіка Глушкова, 40.
Тел.: +38 044526 4342, тел: +38 0445262 49;
{dva, rem, bmm}@isofts.kiev.ua

Приведено огляд найбільш поширених інформаційно-пошукових систем. Увагу надано пошуку інформації в корпоративних сховищах. Розглядається новий підхід до поглибленого пошуку інформації в корпоративних джерелах, в основу якого покладено дворівневий пошук: тематичний пошук та відбір інформації в повнотекстову базу даних і ітераційний пошук в повнотекстовій базі даних.

Resulted most used information-searching systems. Attention is paid to the information searching in corporate storages. Considered new approach of deep information searching in corporate information sources, which is based on the two-level search system: thematic search, selecting of information in full-text database and iterative search in a full-text database.

Вступ

Пошукові системи в корпоративній мережі призначені для роботи з масивами текстових документів підприємства, що мають обсяги від декількох гігабайт до декількох десятків гігабайт. Такі програми мають бути реалізовані в мережному варіанті, при якому доступ до бази даних на сервері локальної мережі, здійснюється з робочих станцій співробітників.

Джерелами інформації в корпоративних мережах виступають не тільки сайти і сторінки, що наповнюють ці сайти і які створені за єдиними стандартами, але і безліч інших баз даних і різних репозиторіїв структурованих і неструктурованих даних. Різнорідність джерел робить неможливим просте перенесення в корпоративне середовище звичайних пошукових машин.

Існує кілька варіантів мережного виконання пошукової програми.

Перший і найпростіший – це можливість пошуку в мережному оточенні. Така програма може індексувати файли, що розташовані не тільки на локальному комп'ютері, але і на дисках інших робочих станцій, з'єднаних в локальну мережу. При цьому пошук може здійснюватися тільки з комп'ютера, на якому встановлена система і розташована база даних, включаючи пошуковий індекс. Багатокористувацький режим, при якому користувачі з своїх робочих місць можуть звернутися до бази даних за інформацією, не забезпечується.

Другий варіант – це пошукові системи, що працюють по Інтернет протоколу. В цьому випадку база даних і основна програма розташована на центральному сервері локальної мережі, а всі користувачі мають доступ до інформації з своїх комп'ютерів через стандартний Інтернет браузер. Тобто все відбувається так само, як і при пошуку в глобальному Інтернеті. Користувач, працюючи в Інтранет мережі, для доступу до бази даних вводить адреси баз даних і далі шукає інформацію по стандартній схемі, із стандартним інтерфейсом пошукової системи. Природно, що пошукові програми, створені на основі пошукових Інтернет систем, в основному використовують Інтернет протоколи у випадку роботи у режимі багатокористувацький.

Наступний рівень – це програмні системи, що мають клієнт-серверну архітектуру з власною клієнтською частиною програми. Програма клієнт встановлюється на всіх робочих станціях мережі, а програма сервер забезпечує індексування інформації всієї мережі, створення бази даних на сервері та доступ до неї всіх користувачів. Такі системи складніші в розробці, але мають більше функціональних можливостей, чим системи, які використовують стандартний браузер. Наприклад, для розмежування доступу користувачів до різних видів корпоративної інформації використовуються системні засоби, а інтерфейс призначений для користувача можна зробити більш функціональним і зручним.

1. Комерційні системи пошуку інформації в корпоративних джерелах

Серед багатьох популярних пошукових систем слід визначити найбільш відомі ситеми: dtSearch, Greenstone, Google Search Appliance, Google Custom Search, Google Desktop, Autonomy, RetrievalWare,

Яndex.ServerStandard 3.0, PolyAnalyst, МЕТАТЕКА. Далі розглянемо їх основні можливості.

Пошукова система dtSearch. Основне призначення програми dtSearch 7.0. [1] – пошук інформації в локальному і мережному оточенні. Система має англійський інтерфейс і працює під керуванням операційних систем Windows 9x/Me/NT/2000/XP/2003. Вона складається з наступних модулів: dtSearch Desktop 7.0 – головний інтерфейс програми, dtSearch Indexer – індексатор документів, dtSearch Index Library Manager – менеджер бібліотек індексів, dtSearch CD Wizard – індексатор даних, що знаходяться на CD. DtSearch 7.0 дозволяє створювати один загальний індекс для кількох комп'ютерів в локальній мережі.

Система забезпечує пошук інформації різних типів, і на різних мовах, включаючи zip, rtf, pdf, html, xml, документи Microsoft Office (Word, Excel, PowerPoint) і WordPerfect. Підтримується кодування Unicode. Допускаються декілька видів пошуку, а саме морфологічний, фонетичний пошук, а також пошук синонімів і пошук в словах з орфографічними помилками. Для лінгвістичного опрацювання текстів dtSearch використовує засоби WordNet. Розроблені додаткові засоби підтримки української мови.

Система Greenstone [2] є Open Source-рішенням для створення "цифрових бібліотек". Природно, вона включає пошук з попередним індексуванням документів різних форматів, і перш за все doc і pdf, які можуть бути представлені і у вигляді архівних файлів. Система створює каталог документів, конвертує їх в html-формат, а потім забезпечує видалений доступ до бібліотек та інших ресурсів за допомогою браузера.

Програмно-апаратний комплекс Google Search Appliance [3] забезпечує пошук документів в рамках корпоративної мережі. Пошуковий механізм комплексу забезпечує роботу більш ніж з 200-ми типами файлів (включаючи html, pdf, doc). При цьому враховуються синоніми при повнотекстовому пошуку за запитами і можлива робота більш з 50-ю природними мовами.

Google Search Appliance підтримує функції пошуку захищеної інформації, що знаходиться на закритих серверах. Але користувач може звернутися до захищеного документа лише за наявності у нього відповідних повноважень доступу.

Google Desktop Search (GDE). Безкоштовна локальна версія відомої пошукової системи Google. На жаль, як сам Google Desktop Search, так і ряд інших безкоштовних зарубіжних пошукових систем поки малопридатні для текстових масивів на російській і українській мовах. Вони не працюють з російською та українською морфологією, погано індексують російськомовні ті українськомовні текстові масиви.

Граничний обсяг тексту в пошуковому індексі також на два порядки нижче, ніж у пошукових систем вищого класу. Наприклад, для GDE максимальний обсяг тексту, що індексується 2 Гб.

Технологія **компанії Autonomy** для корпоративних систем [4] є інструментарієм для автоматизованого керування інформаційними потоками. Основні наукові принципи Autonomy базуються на інформаційній теорії Клода Шеннона, байесовских ймовірностей і нейронних мережах. Концепція адаптивного моделювання ймовірності дозволяє системі Autonomy ідентифікувати шаблони в тексті документа і автоматично визначати подібні шаблони в масиві інших документів.

Обробляючи шаблони рядків у документах, система Autonomy визначає кореляцію образів і виявляє закономірності серед великих масивів документів. При цьому не враховуються ніякі специфічні правила (у тому числі і лінгвістичні). Оскільки система не базується на визначених раніш ключових словах, вона може працювати з будь-якими мовами.

Одним з програмних продуктів Autonomy є пакет Portal-in-a-box, який крім традиційних функцій агрегації інформації з різнорідних джерел має засоби для вирішення такої проблеми як систематизація неструктурованих даних, що виникає при побудові порталів. Очевидно, що угрупування документів за категоріями і створення їх метаописів вимагає чималих редакторських зусиль. Portal-in-a-box в цьому випадку повністю автоматизує процеси категоризації інформації, її реферування і розстановки гіперпосилань.

Інформаційно-пошукова система RetrievalWare [5] є засобом повнотекстового і атрибутивного пошуку. До документів, з якими RetrievalWare здатна працювати, належать тексти в різних форматах і кодуваннях, електронні таблиці, бази даних, поштові повідомлення і т. д., – всього більше двохсот форматів. Система володіє додатковим інструментарієм, який дозволяє налаштуватися на підтримку документів специфічних форматів. Обсяг архіву при необхідності може вимірюватися терабайтами.

Архітектура RetrievalWare дозволяє працювати з системою, як через корпоративну локальну мережу, так і через Інтернет. Серверна частина системи підтримує всі поширені серверні платформи, а клієнтським місцем може бути будь-який комп'ютер, який має графічний Web-браузер. Система володіє можливістю роботи в різних багатопроцесорних і розподілених багатосерверних конфігураціях.

Джерелом інформації може бути файлова система, системи управління базами даних (MS SQL, ORACLE, Sybase та інші СУБД), поштові системи (Microsoft Exchange, Lotus Notes і т.д.), системи управління документами (Documentum EDMS, FileNET Panagon і т.д.), вузли корпоративної мережі й Інтернет, а також електронний архів Excalibur File Room – засіб організації доступу до паперових документів. В основі системи

покладена технологія адаптивного розпізнавання образів, яка базується на нейронних мережах для обробки інформації і діє як система, яка виділяє з масиву збережену інформацію і індексує бінарні образи, що самоорганізовуються. До переваг застосування цієї технології для пошуку текстової інформації можна віднести здійснення нечіткого пошуку, мовну незалежність, малі обсяги індексних файлів.

Основою технології семантичного пошуку є використовування семантичних мереж, які описують значення слів природної мови і зв'язки між поняттями, що визначаються ними. Реалізована також підтримка російської морфології. Семантична мережа словника цієї мови включає близько 40 тисяч семантичних груп в базовому варіанті. Це дозволяє користувачу вводити запит на природній мові і система сама шукає всі документи, контекст яких збігається з контекстом запиту. Застосування семантики дозволяє враховувати загальний контекст документа.

Система Яndex.ServerStandard 3.0 [6] є системним сервісом для організації повнотекстового пошуку інформації у заданій колекції документів. Він призначений для роботи з текстами, як в локальній, так і в глобальній мережах. Система не містить ліцензійних обмежень на число документів, що індексуються, їх розмір або сумарний розмір індексу і дозволяє індексувати документи як через http-з'єднання, так і з локальної файлової системи.

Система Яndex.Server 3.0 складається з двох основних логічних частин: індексатора і пошукового серверу. Індексатор аналізує документи, серед яких має здійснюватися пошук, і зберігає інформацію про них у спеціальних індексних файлах. Звичайно використовується режим роботи, при якому індексні файли не створюються знов, а відпрацьовується інформація тільки за документами, що змінилися, новим і видаленим.

Пошуковий сервер після запуску знаходиться в постійному очікуванні запитів, які можуть бути представлені на природній мові. Пошук може здійснюватися з урахуванням морфології мови, в одній або декількох колекціях документів.

Яndex.Server 3.0 підтримує формати html, xml, rtf, pdf, doc, mp3 і багато інших. Вміст документів, що індексуються, також може бути отриманий при зверненні до довільної бази даних, зокрема, MySQL і MS SQL. Система надає можливість кластеризації результатів пошуку (групує знайдені документи відповідно до зовнішніх атрибутів), а також ранжирує результати (сортує документи за ступенем релевантності запиту).

Рішення PolyAnalyst російської компанії "Мегапьютор" [7] – це система, яка призначена для автоматичного і напівавтоматичного аналізу числових і текстових баз даних з метою виявлення в них раніше невідомих, нетривіальних, практично корисних і доступних закономірностей, які необхідні для ухвалення оптимальних рішень в бізнесі та інших областях людської діяльності.

За своєю природою, PolyAnalyst є клієнт/серверним застосуванням. Користувач працює з клієнтською програмою PolyAnalyst Workplace. Математичні модулі виділені в серверну частину – PolyAnalyst Knowledge Server. Така архітектура надає природну можливість для масштабування системи від розрахованого на одного користувача варіанта, і до корпоративного рішення з декількома серверами.

PolyAnalyst працює з різними типами даних. Це – числа, логічні змінні, категоріальні змінні, текстові рядки, дати, а також вільний текст. PolyAnalyst може обробляти початкові дані з різних джерел, таких як файли Microsoft Excel 97/2000, будь-яка ODBC-сумісна СУБД, SAS data files, Oracle Express, IBM Visual Warehouse.

Модулі PolyAnalyst використовують різні алгоритми Data і Text Mining, у тому числі, модуль Text Categorizer – каталогізатор текстів, який дозволяє автоматично створити ієрархічний деревовидний каталог наявних текстів і відмітити кожний вузол цієї деревовидної структури який є самий індикативний для текстів, що належать до нього.

Модуль Link Terms забезпечує зв'язок понять. Він дозволяє виявляти зв'язки між поняттями, що зустрічаються в текстових полях бази даних, що розглядається, і представляти їх у вигляді графа, який може бути використаний для виділення записів, що реалізовують вибраний зв'язок. Модуль Link Analysis виявляє кореляційні і антикореляційні зв'язки між значеннями категоріальних і булевих полів.

Завдяки унікальній технології "еволюційного програмування" й іншим інтелектуальним алгоритмам, PolyAnalyst з успіхом застосовується в різних типах бізнес-задач, в соціологічних дослідженнях, в прикладних наукових і інженерних задачах, в банківській справі, в страхуванні та медицині.

PolyAnalyst отримав широке розповсюдження в світі, серед її користувачів Boeing, 3M, Chase Manhattan Bank, Dupont, Siemens та інші.

Ядром механізму обробки контента InfoStream [8] є повнотекстова інформаційно-пошукова система InfoReS. Технологія InfoStream дозволяє створювати повнотекстові бази даних і здійснювати пошук інформації, формувати тематичні інформаційні канали, автоматично створювати рубрики інформації, формувати дайджести, таблиці взаємозв'язків понять (як вони зустрічаються в мережних публікаціях), гістограми розподілу вагових значень окремих понять, а також динаміки їх зустрічі за часом. За допомогою InfoStream можна обробляти дані у форматах Microsoft WORD (версії 2000, 97, 6), rtf, pdf, і всіх текстових форматах (простий текст, html, xml). Системи на основі InfoStream у даний час функціонують під керуванням таких операційних систем, як FreeBDS, Linux, Solaris.

Технології InfoStream дозволяють створити комплекс підтримки документального інформаційного сховища, в якому реалізується інтегроване інформаційно-пошукове середовище на основі веб-рішень. За її допомогою забезпечується доступ до електронних документів, розміщених на комп'ютерах в корпоративній мережі, в режимах пошуку, навігація по комп'ютерах/каталогам, перегляд як оригіналів документів, так і їх текстових образів. Комплекс забезпечує інтерактивний повнотекстовий пошук інформації по складних запитах, що складаються з ключових слів, логічних і контекстних операторів, різноманітне ранжирування результатів пошуку. Надається можливість уточнення результатів пошуку за допомогою механізму "інформаційних портретів" [9].

SearchInform Server є ідеальним інструментом для пошуку в корпоративній мережі. Пошук здійснюється не тільки за файлами, які знаходяться в локальній мережі, але і за архівами поштових клієнтів, базах даних і т.д. Забезпечується швидкий пошук інформації у всій мережі, навіть при величезній кількості одночасних клієнтських запитів. Унікальна технологія пошуку документів, схожих за змістом на текст запиту дозволяє легко знайти документи-дублі, а також скоротити час пошуку, максимально конкретизувавши пошуковий запит. Упровадження SearchInform не вимагає зміни існуючих бізнес-процесів і дозволяє максимально зберегти інвестиції компанії, вкладені в існуючу інформаційну інфраструктуру [10].

Пошукову систему "МЕТАТЕКА" компанії "Мета" [11] слід виділити серед пошукових застосувань українського виробництва. "МЕТАТЕКА" доступна як для Windows NT/2000/ XP/2003, так і для Linux/FreeBSD, а для ресурсів, що індексуються, використовується універсальний синтаксис. Структурно "МЕТАТЕКА" виглядає достатньо традиційно, трьома основними компонентами є пошуково-індексуючий механізм, оформлений у вигляді системного сервісу, і два CGI-скрипти для адміністрування і виконання пошуку. Підключені фільтри для обробки файлів в спеціальних форматах (DOC, RTF, XLS, PDF, а в перспективі, ймовірно, і інші), а також словники для англійської, української і російської мов. Слід зазначити, що "МЕТАТЕКА" – один з небагатьох продуктів даного класу, в яких реалізована повноцінна словарна морфологічна підтримка (особливо для української мови), доповнена алгоритмами без словників для невідомих слів, що в сукупності з мовою запитів гарантує дуже ефективний пошук. У системі, окрім звичних плюса, мінуса і лапок, також підтримуються круглі дужки – для групування слів і керування черговістю виконання логічних операторів – і фігурні – для пошуку виразів з урахуванням словозміни. Ядро "МЕТАТЕКИ" призначено для роботи з великими масивами інформації, зокрема воно може одночасно обновляти індекс і обслуговувати призначені для користувача запити.

2. Існуючі технології пошуку інформації

Основною задачею перелічених систем – це пошук інформації у великих повнотекстових масивах [12]. В базі даних таких систем можуть закачуватися будь-які текстові джерела інформації, у тому числі великого обсягу: енциклопедії, довідники, архіви періодичних видань, цілі бібліотеки спеціальної літератури, архіви документів корпорацій, спеціалізовані архіви типу історичних, патентних, судових, розшифровки розмов, протоколи і багато що інше. У відповідь на конкретний запит система видає множину посилань. Далі система має обробити кожне посилання і видати всі відповідні тексти, тобто система має шукати не просто документи, а інформацію, що міститься в них.

Існуючі технології пошуку недостатньо ефективні, щоб знайти у великій кількості різнорідній інформації глобальної мережі відомості, що відповідають запиту. Тому розвивається процес дворівневого пошуку: перший – тематичний пошук і відбір даних в повнотекстову базу даних, другий – пошук в повнотекстовій базі даних.

Всі пошукові системи WWW побудовані за принципом Single Shot Relevancy, тобто «релевантність з першого попадання»: ми робимо один запит і одержуємо потрібні результати. Ми можемо якось змінити запит, але спеціальних механізмів, що забезпечують зворотний зв'язок, не передбачається.

Підхід **Single Shot Relevancy** відповідає ідеї пошуку, вираженій терміном search, що припускає **приблизність**, він забезпечує достовірність, а в корпоративних умовах не менше важлива точність, тут важливо find, виявлення саме того документа, що потрібен і, про існування якого користувач знає наперед.

Для того, щоб підвищити точність пошуку по запитах (drill down analysis – «аналіз з підвищеним рівнем деталізації») необхідно зробити пошук «інтерактивним». Для цього користувач одержує можливість, використовуючи зворотний зв'язок, коректувати запити і поступово добиватися необхідної йому точності.

Для ефективності такої процедури потрібні нові методи представлення результатів пошуку, простого списку сторінок з короткою анотацією, як це роблять звичайні пошукові машини, недостатньо. Результати можуть бути одержані в текстовій формі, з вказівкою зв'язків між окремими компонентами. Існують великі перспективи у графічних формах представлення результатів пошуку. Особливо слід виділити роботу з

мультимедійними даними, жодна пошукова система не працює з такими даними. Рішення цієї задачі зажадає якісно нові підходи. Сучасні пошукові машини побудовані на декількох основних принципах: дані зберігаються у вигляді окремих документів, тому для підвищення продуктивності можна використовувати розподілену архітектуру, використовуючи метадані, можна якимось чином визначати значення документа і, спираючись на моделі, добиватися необхідної релевантності.

3. Функціональність та інтерфейс системи e-content

Пошукова система e-Content розроблена в Інституті програмних систем НАН України і призначена для об'єднання наявних або створюваних цифрових ресурсів наукових установ (корпорації) і забезпечення інтегрованого представлення, пошуку та доступу за уніфікованим Web-інтерфейсом до інформаційних ресурсів. Система забезпечує збір, нормалізацію, індексування і актуалізацію корпоративних цифрових ресурсів.

Система включає три основних компоненти:

визначення та специфікація джерел інформаційних ресурсів;

спрощений пошук типу **Single Shot Relevancy;**

поглиблений пошук типу **Drill Down Analysis.**

3.1. Основою для визначення та специфікації джерел електронних ресурсів є **реєстр електронних ресурсів**. Інформаційні джерела специфікуються за наступними основними показниками:

- **засобами представлення та зберігання**: Web-сайти та портали, бази даних, FTP – сервери тощо;
- **типом інформації**: періодичні видання, звіти, дисертації, автореферати, бібліотеки, окремі статті, адміністративна, нормативна та законодавча інформація тощо;
- **рубриками**: наукові напрямки, відповідно до державного рубрикатора наукових інформаційних ресурсів України.

Створення реєстру електронних інформаційних ресурсів засновано на єдиних принципах і загальноприйнятих стандартах, що дозволяє створювати засоби інтеграції інформаційних ресурсів різних галузей; впорядкувати і стандартизувати пошук та доступ до електронних ресурсів, що мають розподілений характер; організувати пошук та доступ до ресурсів з боку «зовнішніх (за відношенням до ресурсу) клієнтів».

Основними типами електронних ресурсів є: наукові матеріали, електронні періодичні видання, електронні бібліотеки, сайти, учбові матеріали, навчально-методичні матеріали, довідкові матеріали, ілюстративні та демонстраційні матеріали, нормативні документи. До реєстру включаються електронні ресурси наукових установ, бібліотек, освітницьких закладів, інших юридичних осіб, доступ до яких здійснюється через телекомунікаційні засоби загального користування і які мають наукову, історичну, виробничо-комерційну, культурну цінність та становлять інтерес або можуть бути корисними для наукових установ України і окремих науковців.

Для включення електронного ресурсу до реєстру (рис. 1) власник електронного ресурсу надає такі відомості: найменування ресурсу, його доменне ім'я та адреса електронної пошти, умови доступу, анотація, ключові слова, мова, обсяг інформації (у мегабайтах), основні дані (реквізити) власника. Інші відомості про інформаційний ресурс можуть вноситися за пропозицією його власника.

Власники інформаційних ресурсів, фізичні та юридичні особи, що надають послуги із використанням ресурсів, зобов'язані дотримуватися принципів законності, достовірності, загальнодоступності.

3.2. Засоби спрощеного пошук типу **Single Shot Relevancy.** Для забезпечення можливості індексування та пошуку інформації в межах системи були реалізовані базові сервіси інформаційно-управляючого ядра системи (служба захисту, індексування, пошуку та доступу до корпоративних електронних ресурсів). Для реалізації засобів індексування та пошуку в спрощеному режимі були використані сервіси GOOGLE Custom Search (рис. 2). Після попередніх перевірок зареєстрованого ресурсу, ресурс специфікується засобами пошукового сервісу GOOGLE – Custom Search. Далі він автоматично індексується і стає доступним через пошукові сервіси системи GOOGLE корпоративним користувачам. Для пошуку ресурсу в системі пропонується пошуковий сервіс (рис. 3). Через пошукову сторінку користувачі можуть формувати пошуковий контекст за правилами, які визначені в системі GOOGLE. Ресурси, що проіндексовані сервісом Custom Search не доступні для загального пошуку в системі GOOGLE.

В системі підтримуються комерційні формати даних html, txt, doc, rtf, pdf, chm та інші. Реалізовано індексування корпоративних електронних ресурсів, забезпечується пошук ресурсів та доступ до знайдених ресурсів через посилання.

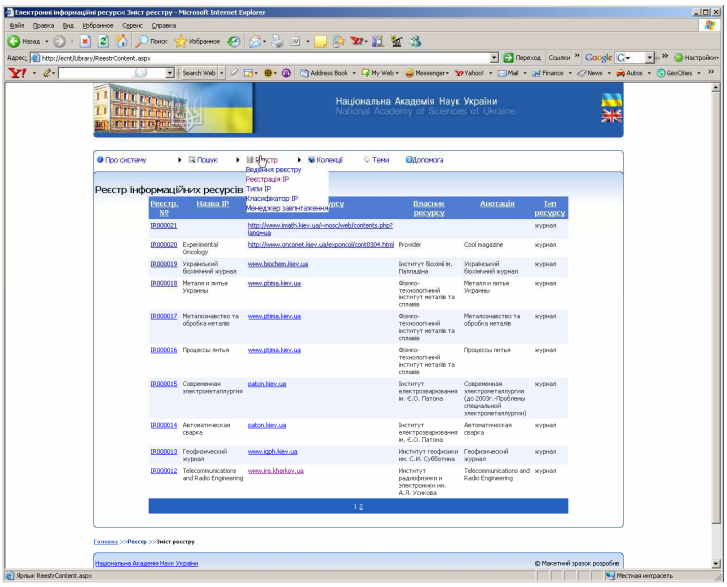

Рис. 1. Інтерфейс ведення реєстру джерел інформаційних ресурсів

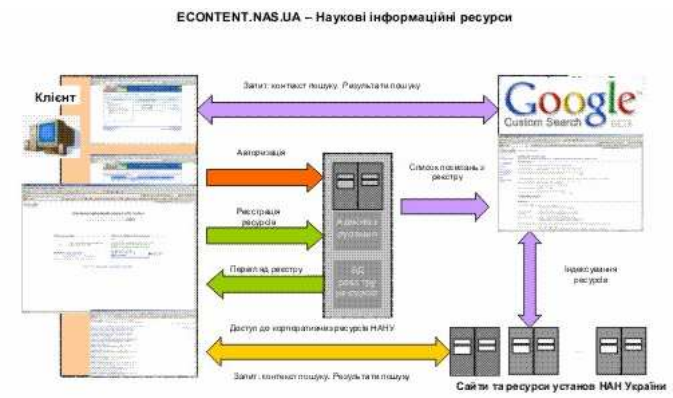

Рис. 2. Схема взаємодії з сервісом GOOGLE – Custom Search

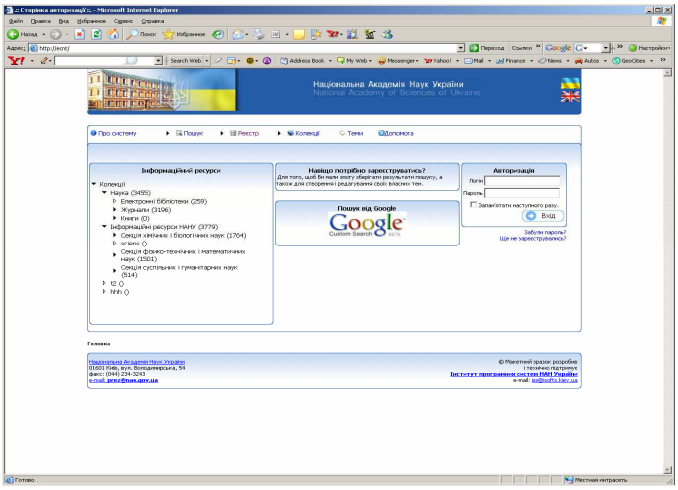

Рис. 3. Спрощений пошуковий сервіс

3.2. Поглиблений пошук в системі e-Content. Сервіс поглибленого пошуку (рис. 4) надає користувачеві такі основні можливості:

- визначити джерела, в яких здійснюється пошук інформації. Користувач має змогу визначити пошукові інформаційні джерела за інтересами з тих, які задані в реєстрі;
- визначити пошукові тематичні категорії, які обмежують контекст пошуку та забезпечують отримання більш релевантної до запиту інформації;
- створювати пошуковий контекст за заданим текстом (документом). Для формування запиту користувач має змогу визначити текстовий документ або фрагмент документу, за яким формується запит. Цей сервіс заснований на методі латентного семантичного аналізу [13]. Його можна розглядати як ітераційний пошуковий сервіс. Документами, що використовуються для наступного пошуку є результати попереднього пошуку.

Додаткові пошукові сервіси забезпечуються засобами dtSearch TextRetieval Engine. Користувачі системи одержують релевантні відповіді на пошукові запити, мають можливість використовувати єдиний інтерфейс для формування пошукових запитів, такий, що не вимагає спеціальної підготовки та виконувати гнучке підключення нових ресурсів до системи. Служба підтримує повнотекстовий пошук, використовує засоби категоризації даних, використовує можливості єдиної реєстрації ресурсів, підтримує аутентифікацію та авторизацію користувачів.

Система e-Content використовується як пошуковий сервіс у корпоративному порталі НАН України www.nas.gov.ua.

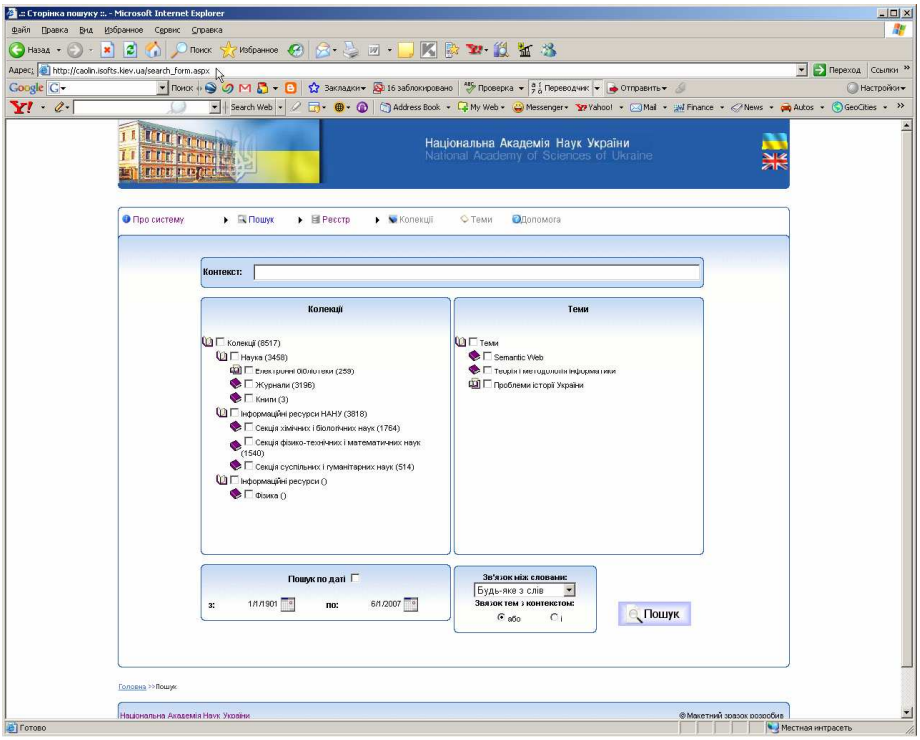

Рис. 4. Інтерфейс поглибленого пошукового сервісу

Висновки

На сучасному етапі основним джерелом інформації поступово стає Інтернет, де експерти і аналітики користуються загальнодоступними пошуковими машинами і періодично проглядають найбільш інтересні з їхньої точки зору сайти у пошуках новин за тематиками, що цікавлять їх. При цьому основний час експертів витрачається на непродуктивну рутинну роботу − перегляд і відбір найбільш релевантної інформації.

Збір інформації в Інтернет сьогодні неефективний з наступних причин.

Неповнота вибірки. Традиційні пошукові системи, засновані на ключовому пошуку, орієнтовані на людину, яка точно знає, що вона шукає. При знайомстві з новим для себе питанням експерти знаходяться в абсолютно протилежній ситуації. Вони заздалегідь не знають релевантних даному питанню термінів і словосполучень, так що маса інформації випадає з їх поля зору.

Нерелевантність документів. Ключовий пошук без урахування потрібного контексту приносить дуже багато нерелевантної інформації, проглянути яку, зазвичай, просто не представляється можливим. У результаті, корисні документи часто виявляються захороненні під горою інформаційного сміття.

Неуважність фактів. Документи, що представляють інтерес для експертів, зазвичай досить об'ємні, так що за умов жорсткого ліміту часу експерт просто не встигає прочитати всі знайдені з даного питання документи. Відсутні інструменти пошуку найбільш насичених тематичною інформацією фактів.

Множинність джерел. Моніторинг новин припускає перегляд експертами безлічі джерел з тематики, що цікавить експерта. Ця рутинна процедура віднімає багато часу і сил. У ідеалі всі новини мають збиратися з численних джерел автоматично і сортуватися відповідно до їх тематик. Пошук авторитетних джерел з даної тематики є важливим самостійним завданням.

Потрібні нові підходи і методи до пошуку неструктурованої інформації. Можливо дослідження, що проводяться в Semantic Web середовищі дозволять підійти до вирішення цих проблем.

1. http://www.dtsearch.com
2. http://www.greenstone.org
3. http://www.google.com
4. http://www.autonomy.com
5. http://www.convera.com
6. http://www.yandex.ru
7. http://www.megaputer.ru
8. http://infostream.com.ua
9. *Ландэ Д.В.* Поиск знаний в Internet. − Санкт-Петербург: Диалектика, 2005. − 266 с.
10. http://www.corporateinform.ru
11. http://www.meta.ua
12. *Андон П.І., Дерецкий В.А.* Процесори пошуку та аналізу природномовної текстової інформації в аналітичних системах – 2001.– № 3-4. – С. 144–163.
13. *Дерецкий В.А.* Підхід до автоматичної побудови тематичної онтології документу для удосконалення інформаційного пошуку – 2005. – № 3. – С. 76–82.