УДК 004.9

Кебкало О.С., Михайлюк А.Ю., Тарасенко В.П.

ПОШУКОВА СИСТЕМА ДЛЯ ЕЛЕКТРОННИХ ІНФОРМАЦІЙНИХ РЕСУРСІВ ВИЩОГО НАВЧАЛЬНОГО ЗАКЛАДУ

У статті розглядається спосіб створення пошукової системи для отримання даних з усіх ресурсів інформаційного простору навчального закладу.

Ключові слова інформаційно-пошукова система, ІПС, інформаційне середовище навчального закладу, інформаційні ресурси навчального закладу.

Вступ

Інформаційний ресурс сучасного вищого навчального закладу включає велику кількість даних та знань, як науково-освітнього характеру, так і інших спрямувань, які забезпечують та супроводжують діяльність закладу. Ці дані та знання утворюють інформаційний простір, а їх різноманіття за змістом і за формою подання обумовлює його гетерогенний характер. Дані та знання знаходяться на різних ресурсах (різних фізично та різних за типом): web-сайти, ftp-сервери, спільні папки тощо, а також бази даних та бази знань інформаційних систем, які впроваджені в закладі. Природним бажанням користувачів (якими є як працівники закладу, так і студенти) є оперативне отримання необхідних даних з інформаційного простору, для чого, як правило, використовуються засоби пошуку для окремого типу ресурсу або навіть для окремого ресурсу: пошук серед web-сторінок, пошук по ftp-сайтам, пошук в спільних папках, пошук у конкретній інформаційній системі. Різноманітність ресурсів та пошукових засобів, якими доводиться користуватись, часто призводить до складності та незручності знаходження інформації, тому актуальним завданням є створення єдиного пошукового механізму з єдиним інтерфейсом для пошуку серед усіх або більшості ресурсів навчального закладу.

Розділ 1 Вимоги до пошукової системи в інформаційному середовищі навчального закладу

Більшість інформаційно-пошукових систем (ІПС), які існують сьогодні, мають справу з відносно однорідною інформацією, яка найчастіше зберігається в одній базі даних. Але, у випадку з ресурсами вищого навчального закладу, виникає необхідність взаємодіяти з багатьма джерелами інформації та з багатьма базами даних, кожна з яких має свою структуру та зберігає свій тип інформації. Одночасна робота інформаційно-пошукової системи з декількома базами даних не може не внести свій відбиток при проектуванні такої системи [1,2].

Виходячи з аналізу можливостей та функцій існуючих мультибазових пошукових систем, а також враховуючи специфіку системи для використання в навчальному закладі, виділимо наступні вимоги до пошукових систем для навчальних закладів, орієнтованих на роботу з усім інформаційним простором закладу:

- можливість роботи з розподіленими даними – система повинна забезпечувати можливість роботи з даними, що знаходяться на різних

фізичних серверах, різноманітних апаратно-програмних платформах та тими, що зберігаються у різних внутрішніх форматах;

– логічне угрупування даних – система повинна дозволяти обробляти усі запити на логічних групах баз даних, повністю приховуючи тим самим фізичне розташування останніх;

– абстрактна модель даних – система має бути побудована на основі абстрактної схеми даних, на яку повинні бути відображені конкретні БД; це дозволяє поєднати дані з різнорідних систем у одній логічній групі;

– абстрактна система запитів – система повинна оперувати не конкретним синтаксисом запитів, а його логічною суттю на основі абстрактних атрибутів;

– метаінформація – система повинна надавати повну інформацію про себе та про всі свої ресурси;

– розмежування доступу – система повинна мати змогу надавати різні рівні привілеїв для користувачів по доступу до інформації;

– облік та контроль – система повинна вміти збирати статистичні дані за запитами користувачів та вести їх бюджети;

– відкритість – система повинна допускати розширення та має бути заснована на відкритих стандартах та протоколах;

– зв'язок з іншими системами – можливість одночасного використання власних ресурсів з ресурсами інших інформаційних систем;

– демократичність у спілкуванні – система повинна надавати як прості та зрозумілі для непідготованого користувача інтерфейси, так і професійні інтерфейси для доступу до інформації;

– зв'язок з Інтернет – система повинна бути доступною для використання через Інтернет.

Реалізація цих вимог у реальній інформаційній системі дозволила б максимально задовольнити сьогоднішні потреби мережного інформаційного сервісу для доступу до інформаційних ресурсів навчального закладу.

Розділ 2 Прототип пошукової системи для єдиного інформаційного середовища

Виходячи з вимог, які були розглянуті, пропонується підхід до створення універсальної документоорієнтованої інформаційно-пошукової системи, яка забезпечує пошук різнотипних даних на різнотипних ресурсах (web-ресурси, ftp-ресурси, інші ресурси зі стандартними протоколами доступу, ресурси спеціалізованих інформаційних систем). Даний підхід було реалізовано у вигляді прототипу пошукової системи, призначеної для роботи в інформаційному середовищі вищих навчальних закладів. Документоорієнтованість системи визначається тим, що основними об'єктами, з якими працює система, є, відповідно, документи, під якими розуміємо текстомісткі природо мовні інформаційні об'єкти. Для документів передбачена значна кількість можливостей системи, хоча в цілому можна здійснювати пошук довільних інформаційних об'єктів.

Пошукова система може бути реалізованою окремо від інших інформаційних систем навчального закладу для виконання своєї основної функції (оперативного надання інформації користувачу системи), але можна вести мову про пошукову систему як частину єдиної інформаційної системи або єдиного інформаційного середовища (ЄІС), яке охоплює всі або переважну більшість систем та ресурсів, що використовуються в учбовому закладі. Місце

пошукової системи в такому середовищі може виглядати наступним чином (рис.1).

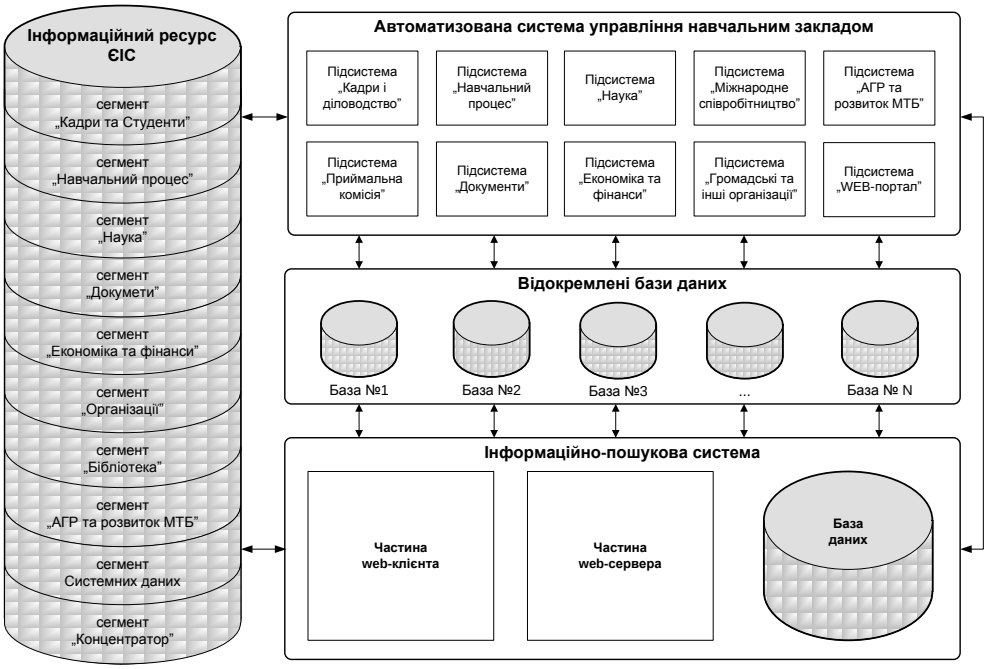

Рис. 1 – Пошукова система в складі єдиного інформаційного середовища вищого навчального закладу

Як видно з рисунка, ЄІС включає в себе наступні компоненти:

- інформаційний ресурс ЄІС;
- відокремлені бази даних;
- автоматизована система управління навчальним закладом;
- інформаційно-пошукова система.

Тобто ЄІС включає інформаційні ресурси навчального закладу та системи, які працюють з цими ресурсами.

Задача «максимум» пошукової системи – це здійснення пошуку серед наступних інформаційних зрізів:

- інформація про кадри та студентів;
- інформація щодо підтримки навчального процесу;
- наукова інформація;
- економічно-фінансова інформація;
- інформація бібліотеки навчального закладу;
- інформація щодо матеріально-технічної бази;
- інша інформація.

В розробленому прототипі були створені механізми для пошуку серед сегментів навчально-методичної інформації, наукової інформації, бібліотечної інформації та інформації системи документообігу.

Прототип орієнтований на використання в web-середовищі. Це визначає необхідність вибору трирівневої архітектури системи, яка передбачає наступні рівні:

– частина web-клієнта, тобто програмне забезпечення, реалізоване за допомогою мови JavaScript, яке забезпечує інтерактивний інтерфейс користувача на основі функціональності web-браузера;

– частина web-сервера. Це програмне забезпечення, реалізоване мовою C#, що працює під управлінням .NET Framework, та забезпечує трансляцію клієнтських запитів у запити до менеджера бази даних;

– рівень бази даних, який містить дані системи і форми інтерфейсу мовою HTML, а також включає до себе ряд процедур, тригерів і функцій мовою SQL, які виконують ряд специфічних функцій обробки даних.

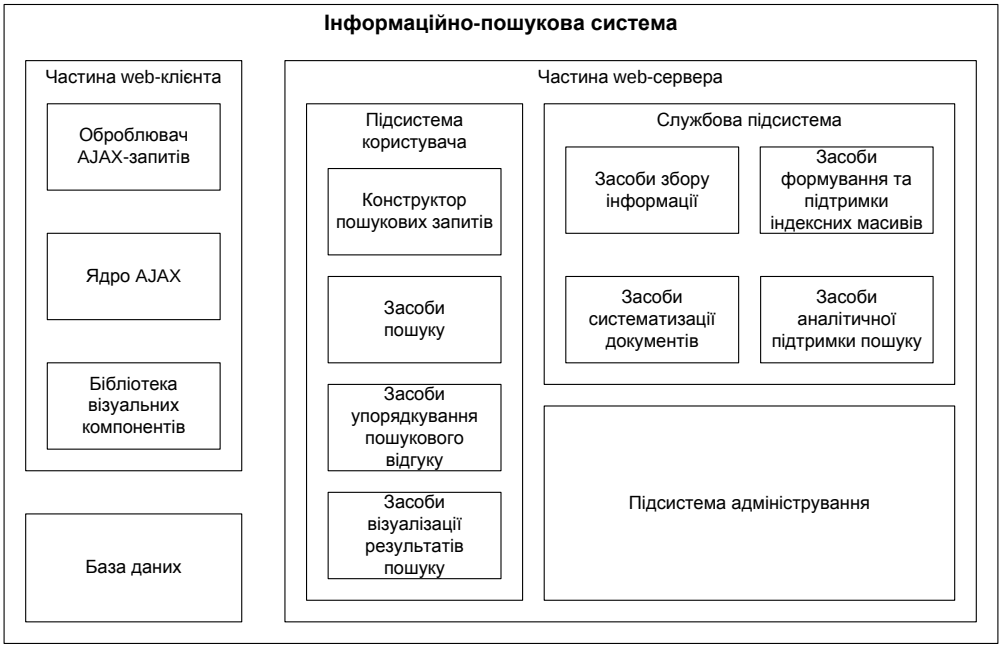

Рис. 2 – Узагальнена структура пошукової системи

Для реалізації обміну даними між web-клієнтом і web-сервером вибрана технологія AJAX, як така, котра оптимально розподіляє навантаження між клієнтом і сервером, а також є найбільш перспективною з боку подальшої інтеграції з додатками на основі Web 2.0.

Клієнтська частина управляє станом елементів управління сторінки і забезпечує інтерактивний інтерфейс користувача. До її задач відноситься реагування на дії користувача і відправлення запитів до серверної підсистеми у випадку, якщо даних не вистачає. Склад клієнтської підсистеми:

– оброблювач AJAX-запитів, що відправляє на сервер XML-запити, одержує від нього результати обробки даних і передає їх у відповідні процедури;

– ядро AJAX, що одержує повідомлення від всіх елементів сторінки, обробляє їх і, якщо буде потреба відправляє запити на до оброблювача запитів;

– бібліотека візуальних компонентів, яка містить у собі набір складних елементів управління, призначених для візуалізації і зміни стану певних об'єктів бази даних; до таких компонентів відносяться інтерактивна

таблиця, дерево, елемент введення типу календар, список з динамічним завантаженням і т.п..

Серверну частину системи умовно можна поділити на три підсистеми: підсистему користувача, підсистему адміністрування та службову підсистему.

До підсистеми користувача відносяться:

- конструктор пошукового запиту;
- засоби пошуку;
- засоби упорядкування пошукового відгуку;
- засоби візуалізації результатів пошуку.

До підсистеми адміністрування входять відповідно засоби адміністрування [3].

Службова підсистема включає:

- засоби збору інформації;
- засоби формування та підтримки індексних масивів;
- засоби систематизації документів;
- засоби аналітичної підтримки пошуку.

Розділ 3 Модель даних пошукової системи

На рис.1 показано інформаційні ресурс навчального закладу, серед яких в принципі можна здійснювати пошук. З їх переліку зрозуміло, що ресурси є різними як за змістом, так і за формою доступу, тому створення для них єдиного пошукового механізму вимагає їх приведення до деякого єдиного формату подання даних. В розробленій системі таким форматом стала модифікована модель «сутність-атрибут-значення» [4], згідно з якою кожному елементу ресурсів (ми їх називаємо документами) відповідає деяка множина атрибутів, котрі описують даний елемент. Наприклад, якщо елементом ресурсу є студентський реферат, то його атрибутами можуть бути назва, ПІБ автора або авторів, група, номер залікової книжки, дисципліна, ПІБ викладача, рік, класифікація, безпосередньо текст реферату тощо. В найбільш простому випадку атрибут може бути один – текст документа. Використання множини атрибутів надає можливість упорядкування документів та значно розширює пошукові можливості, що, у кінцевому результаті, призводить до підвищення ефективності пошуку для користувача за рахунок зменшення часу отримання необхідних документів.

Таким чином для кожного документа формується перелік атрибутів, які його описують. Документи об’єднуються за типами, документи одного типу мають однакові базові набори атрибутів, які визначають тип . Пошук документів здійснюється на основі пошуку значень атрибутів. Основним атрибутом текстомістких документів є текст документа, для якого надається можливість повнотекстового пошуку.

Формування типів документів та переліку атрибутів, з якими має справу пошукова система, виконується вручну адміністратором системи.

Для забезпечення можливості знаходження певного документу, його потрібно попередньо проіндексувати для пошуку, а для цього система повинна отримати інформацію про існування документа та мати можливість отримати сам документ. Документи, серед яких здійснюється пошук, можуть зберігатися:

- на web-серверах;
- на ftp-серверах;
- на інших ресурсах зі стандартними протоколами доступу;
- в різноманітних інформаційних системах.

В прототипі системи реалізовано пошук документів, які розміщені на web- або ftp-серверах. Отримання документів здійснюється засобами збору інформації, до яких відносяться т.з. програми-«павуки». Кожна з таких програм сканує свій тип ресурсів (web, ftp, спільні папки) та, при знаходженні документа відповідного типу, запускають механізми індексування. В процесі індексування в системі по суті створюється копія документа, яка забезпечить доступ до документа навіть у разі непрацездатності самого ресурсу, на якому він знаходився, або навіть його видалення з ресурсу. Також надається можливість додавати в систему документи в пакетному режимі з жорсткого диску користувача з відповідними правами.

Для документів, які зберігаються в інформаційних системах, необхідні спеціальні програмні модулі, які будуть взаємодіяти або з самою системою, або з базою даних системи та отримувати відповідні документи. В прототипі була реалізована можливість пошуку в ресурсі системи документообігу.

В узагальненому вигляді схема отримання документів пошуковою системою виглядає наступним чином (рис. 3).

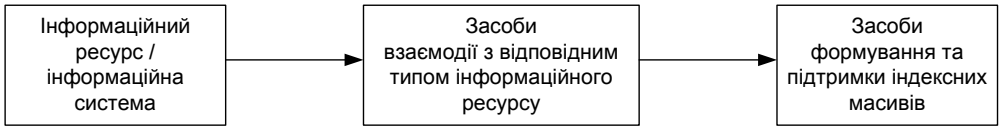

Рис. 3 – Схема отримання документів пошуковою системою

Засоби індексування документів [5] на рис. 3 не залежать від ресурсу, з якого ці документи отримуються, та типу доступу до нього. Взаємодія з інформаційними ресурсами або з інформаційними системами відбувається виключно за допомогою засобів збору інформації. Для того щоб залучити ресурси нового типу, наприклад, документи якоїсь інформаційної системи, достатньо додати в систему модуль для взаємодії з цим типом ресурсів. При цьому ні засоби індексування, ні пошукові механізми змінювати не треба.

Індексування документів виконується за допомогою засобів формування та підтримки індексних масивів за наступною схемою (рис. 4).

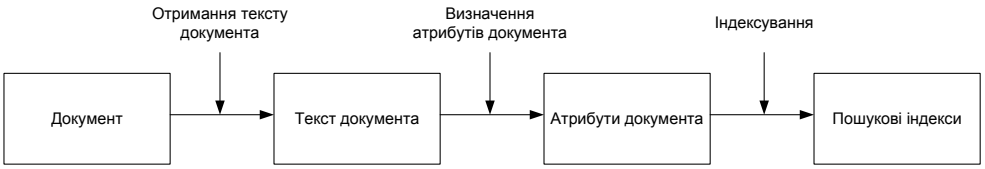

Рис. 4 – Схема отримання документів пошуковою системою

Оскільки система орієнтована перш за все на текстомісткі документи, першим кроком при виконанні індексування є отримання безпосередньо тексту документа. Прототип системи здатний працювати з найбільш поширеними форматами (.doc, .txt, .rtf, .pdf, .djvu, html та інші). Після отримання тексту виконується процедура визначення атрибутів документа. Як правило, доцільно формувати перелік атрибутів таким чином, щоб була можливість визначити їх автоматично. Частину атрибутів можна визначити з атрибутів файлу, в якому зберігається документ, але більшість атрибутів визначається на основі аналізу тексту документа. Якщо тип документа заздалегідь не відомий, попередньо виконується спроба визначення типу документа. Якщо якісь атрибути для

даного типу документа не вдалось визначити автоматично, то їх значення необхідно ввести вручну.

Після визначення атрибутів виконується безпосередньо індексування документа. Воно включає індексування для здійснення повнотекстового пошуку (для чого виконується також морфологічний аналіз із залученням відповідного словника) та індексування для здійснення аналітичних операцій над документом.

Розділ 4 Пошукові засоби

Інтерфейс пошуку надає конструктор пошукового запиту для виконання наступних видів пошуку:

- пошук за атрибутами;
- простий повнотекстовий пошук;
- розширений повнотекстовий пошук;
- інтелектуалізований повнотекстовий пошук за шаблоном мовної конструкції;
- комбінований пошук.

Режимом пошуку за замовчуванням є повнотекстовий пошук за шаблоном мовної конструкції [6,7], який враховує порядок слів у запиті та відстань між словами у шуканих документах (результати пошуку будуть відсортовані за ступенем відповідності пошуковому запиту).

За допомогою розширеного повнотекстового пошуку є можливість шукати такі документи, в яких немає введених слів, в яких є хоча б одне або кілька введених слів та в яких є всі введені слова.

При здійсненні пошуку також передбачена можливість залучення тезауруса, за рахунок чого відбувається розширення запиту словами-синонімами.

Результати пошуку сортуються за релевантністю, що призводить до того, що першими в списку знайдених документів знаходяться документи, які найбільше відповідають пошуковому запиту.

Також для результатів пошуку передбачені засоби кластеризації, які розбивають множину результатів на підмножини документів, споріднених за змістом, що полегшує користувачеві подальшу роботу у випадку, коли знайдених документів велика кількість.

Окрім безпосередньо пошуку необхідні документи можна отримати за допомогою засобів систематизації [8,9], до яких відноситься адаптивний каталог, що також включає в себе класифікатор. Засоби каталогізації надають можливість будувати каталоги шляхом групування документів за значеннями атрибутів, які обираються користувачем. Побудова каталогу відбувається інтерактивно. Користувач може обирати атрибути, за якими слід виконати групування, в довільній послідовності.

Режим каталогу включає можливості для роботи з деревовидними класифікаторами, для чого використовується атрибут класифікації. Класифікація документа відбувається автоматично при додаванні документа в систему.

Висновки

Таким чином описаний підхід до створення інформаційно-пошукової системи для інформаційного середовища вищого навчального закладу дозволяє реалізувати універсальні потужні пошукові механізми, орієнтовані на оперативне отримання даних з інформаційних ресурсів закладу.

Подальші дослідження будуть спрямовані на розширення переліку типів ресурсів, що охоплюються пошуком, зокрема на забезпечення взаємодії не тільки з системою документообігу, а й з ресурсами інших спеціалізованих інформаційних систем, які використовуються в навчальному закладі; на поглиблення аналізу, ідентифікації та визначення структури документів, отриманих засобами збору інформації; на нарощення переліку аналітичних можливостей системи.

Література

1. Управління інформаційними технологіями вищих навчальних закладів: навчальний посібник. Видання третє, доповнене / За ред. О.В. Співаковського. — Херсон: Айлант, 2010. — 302 с.
2. Саак А.Э., Пахомов Е.В., Тюшняков В.Н. Информационные технологии управления: Учебник для вузов. 2-е изд. — СПб.: Питер, 2008. — 320 с.
3. Петрашенко А.В., Замятин Д.С. Методы разграничения доступа в неструктурированных текстовых хранилищах // Тези доповідей Міжнародної науково-практичної конференції "Інформаційні технології та комп'ютерна інженерія". — Вінниця. — 2010. — С. 217—218.
4. Клюско В.В., Петрашенко А.В Модель композитних атрибутів текстових документів // Збірник тез доповідей II наукової конференції магістрантів та аспірантів присвяченої 20-річчю факультету прикладної математики "Прикладна математика ти комп'ютинг". — Київ. — 2010. — С. 261—264.
5. Замятін Д.С., Михайлюк В.А., Петрашенко А.В. Підвищення продуктивності повнотекстового пошуку шляхом реорганізації подання інвертованих індексів // Науковий вісник чернівецького університету. Збірник наукових праць. Випуск 426. — Чернівці. — 2008. — С. 63—67.
6. Mykhailyuk A., Zamiatin D., Petrashenko A. Unstructured Data Warehouse Processing System Based on an Uniform Set of Functions // Proceedings of the 4-th International Conference ACSN-2009 "Advanced Computer Systems and Networks: Design and Application". - Lviv. - 2009. - P. 117-119.
7. D.S. Zamyatin, V.A. Mykhaylyuk, A.V. Petrashenko, O.S. Mykhaylyuk Method of Full-text Search Accelerating by Cache-friendliness optimization // Збірник праць Ювілейної міжнародної науково-практичної конференції, що присвячена 50-річчю створення першої на Україні кафедри обчислювальної техніки РКС-2010 "Розподіленні комп'ютерні системи". — Київ. — 2010. — С. 77—80.
8. A. Mykhaylyuk, A. Petrashenko, D. Zamiatin Method of Unstructured Text-Based Warehouses Systematization Based on Composite Attributes // Збірник праць Ювілейної міжнародної науково-практичної конференції, що присвячена 50-річчю створення першої на Україні кафедри обчислювальної техніки РКС-2010 "Розподіленні комп'ютерні системи". — Київ. — 2010. — С. 86—88.
9. Михайлюк А.Ю., Замятін Д.С., Петрашенко А.В. Проблеми систематизації даних у неструктурованих текстових сховищах // Вісник університету "Україна". Серія "Інформатика, обчислювальна техніка та кібернетика". №8. — Київ. —2010. — С. 88—91.

Кебкало А.С., Михайлюк А.Ю., Тарасенко В.П. Поисковая система для электронный информационных ресурсов высшего учебного заведения

В статье рассматривается способ создания поисковой системы для получения данных из всех ресурсов информационного пространства учебного заведения
Ключевые слова: информационно-поисковая система, ИПС, информационная среда учебного заведения, информационные ресурсы учебного заведения

Kebkalo A., Mykhaylyuk A., Tarasenko V. Search engine for electronic information resources of a higher educational institution

In article the search engine creation way for data acquisition from all resources of educational institution information field is considered
Keywords: information retrieval system, educational institution information environment, educational institution information resources