УДК 061.62:004.42

ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ДЛЯ НАУКОВО-ДОСЛІДНОЇ ЛАБОРАТОРІЇ З ФУНКЦІЄЮ АВТОМАТИЗОВАНОГО АНАЛІЗУ ТА ПОПОВНЕННЯ КОНТЕНТУ

Студ. К. О. Несін, гр. МгІТ-2-17
Науковий керівник доц. В.Г. Резанова
Київський національний університет технологій та дизайну

Мета і завдання.

Мета – створення науково-дослідної лабораторії з функцією автоматичного поповнення контенту. Контент повинен попередньо аналізуватись за допомогою алгоритму, який забезпечить якість відібраного матеріалу.

Завдання розподіляються на декілька етапів: створення лабораторії з розподілом на різні категорії, підбір початкових даних, що допоможуть класифікувати тематику контенту та отримувати джерела нових матеріалів, створення та тестування алгоритму відбору та публікації текстів.

Об'єкт та предмет дослідження.

Щоденно в всесвітній мережі Інтернет зникають тисячі сайтів, деякі з них містять корисну інформацію, яку можна відновити.

Методи та засоби дослідження.

Для реалізації завдання використовувались базові .NET технології, ASP.NET MVC, MS SQL та Html Agility Pack, також варто розробити власні алгоритми поповнення контенту.

Наукова новизна та практичне значення отриманих результатів.

Робота забезпечує автоматичне поповнення контенту науково-дослідної лабораторії на основі інформації, що зникла з пошукових мереж та може бути цікавою користувачу.

Результати дослідження.

На першому етапі роботи варто підготувати теоретичну основу: визначити основні категорії науково-дослідної лабораторії, ключові слова, які допоможуть визначати тематику контенту, алгоритм та джерела відбору текстів. Потрібно порівняти різні джерела та методики відбору, щоб забезпечити найкращий результат.

Початкові дослідження показали, що ми можемо щоденно отримувати і аналізувати близько 5000 можливих джерел матеріалів. Джерелами виступають сайти, які зникли з видачі пошукових мереж. Розроблена система має систематизувати отримані джерела, щоб не аналізувати їх повторно, зберігати статистику про розмір джерела (кількість сторінок видаленого сайту, що доступні для перегляду) та кількість отриманих тематичних текстів.

На наступному етапі варто створити алгоритм відбору серед можливих джерел матеріалів ті, що відповідають нашій тематиці. Попереднє дослідження показало, що відбір по ключовому слову в домені надає занадто мало результатів, тому реалізований алгоритм має основуватись на аналізі стартової сторінки. Щоб отримати матеріал видалених сайтів можна використовувати публічні архіви мережі Інтернет.

Алгоритм відбору контенту має забезпечувати отримання текстових матеріалів, що об'єднані одною тематикою та відповідають критеріям якості, визначеним попередньо. Узагальнений вигляд алгоритму представлений на рисунку 1. Базові критерії якості мають чітко та однозначно відрізняти тематичну статтю від будь-якого іншого текстового матеріалу. Варто забезпечити якість виділення і очищення тексту

статті від інших текстових фрагментів та зайвої html-розмітки. Ми будемо зберігати основні елементи html-розмітки, щоб текст зберіг свою логічну структуру, яка забезпечує його зрозумілість для читача. На цьому етапі доцільно буде використання бібліотеки Html Agility Pack, яка забезпечує зручну роботу з html-документами.

Третій етап полягає в створенні науково-дослідної лабораторії зі зрозумілим користувачу інтерфейсом, вона буде візуалізувати результати роботи та відображати всі знайдені тексти з логічним розподілом їх на категорії. Для цього буде використано технології ASP.NET MVC, які забезпечать простоту і масштабованість веб-інтерфейсу.

Алгоритм підбору джерел та виділення текстів може бути реалізований як в консольній програмі, так і за допомогою WindowsForm додатку. Науково-дослідна лабораторія буде створена за допомогою ASP.NET MVC. Для поєднання окремих частин в одну систему будемо використовувати роботу з базами даних MS SQL.

Рисунок 1 − Узагальнений вигляд алгоритму аналізу тексту

Робота базується на сучасних веб-технологіях та знаннях про принцип роботи мережі Інтернет, що забезпечують автоматичне поповнення контенту науково-дослідної лабораторії на основні інформації з видалених сайтів.

Ключові слова: *ASP NET MVC, алгоритм виділення тексту на html-сторінці, автоматичне поповнення контенту.*

ЛІТЕРАТУРА

1. http://html-agility-pack.net/api
2. http://www.interface.ru/home.asp?artId=27160
3. J. Gibson, B. Wellner, and S. Lubar. Adaptive web-page content identication.
4. Hung-Yu Kao, Jan-Ming Ho. WISDOM: Web Intrapage Informative Structure Mining Based on Document Object Model
5. Christian Kohlschütter, Peter Fankhauser, Wolfgang Nejdl. Boilerplate Detection using Shallow Text Features