УДК 004

ПРОГРАМНА СИСТЕМА ДЛЯ АВТОМАТИЗОВАНОГО СТВОРЕННЯ БАЗИ ДАНИХ «ФУТБОЛ» НА ОСНОВІ АНАЛІЗУ HTML – СТОРІНОК МЕРЕЖІ ІНТЕРНЕТ

Тітов М.М., Коломойцева І.О.
Донецький національний технічний університет
кафедра комп'ютерних наук та технологій
E-mail: titov.nikolay.n@gmail.com

Анотація

Тітов М.М., Коломоцева І.О. Програмна система для автоматизованого створення бази даних «Футбол» на основі аналізу HTML – сторінок мережі Інтернет. *Розглянуті методи для аналізу текстів. Визначено мову для розробки програмного комплексу C# та СУБД для зберігання інформації MS SQL. Проаналізовані методи вилучання даних з мережі Інтернет.*

Загальна постановка проблеми

У зв'язку з бурхливим розвитком комп'ютерної техніки та телекомунікаційних технологій стало гострим завдання пошуку інформації. На сьогоднішній день в електронному вигляді зберігається величезна кількість документів, керівництв, описів, інструкцій, підручників, наукових статей та багато іншої неструктурованої інформації.

Проблема знаходження серед такого обсягу потрібної інформації стає вкрай важливою і найчастіше важко розв'язуваної без використання спеціальних.

Розвиток мережі Інтернет ще більше посилило дану проблему, оскільки кількість документів, доступних за допомогою цієї мережі, величезна і продовжує постійно зростати.

Такі засоби пошуку, як інформаційно-пошукові та метапошукові системи Інтернет, каталоги Інтернет та індивідуальні пошукові агенти, а також системи Інтернет-моніторингу, дозволяють спростити різні аспекти вирішення цієї проблеми. Однак, будучи застосовними для вирішення приватних пошукових завдань, існуючі засоби не забезпечують вирішення комплексних завдань.

Область досліджень

Областю досліджень є організація баз даних, знань та розподілених систем, розробка програмних комплексів щодо класифікації даних, створення методів аналізу текстів та пошуку інформації.

Предмет досліджень

Архітектура програмних систем, баз даних, алгоритми та програмні шляхи до класифікації текстових даних та методів пошуку інформації.

Мета досліджень

Розробка програмного комплексу для збору інформації у мережі Інтернет, яка дозволяє автоматизувати процес інформаційного пошуку. Це забезпечує можливість завдання області пошуку та знаходження необхідних документів, також зберігати засоби навігації у сформованій базі даних.

Проблема інтелектуального аналізу даних

Основною проблемою логічних методів виявлення послідовностей є проблема переліку варіантів за необхідний час. Існуючі методи або штучно обмежують перелік (алгоритми WizWhy, KOPA) або складають дерева рішень (алгоритми CART, CHAID, See5, Sipina ID3 та інші), які мають принципові межі ефективності пошуку правил if-then.

Інші проблеми пов’язані з тим, що відомі методи пошуку логічних правил не підтримують функцію узагальнення знайдених правил та функцію пошуку оптимальної

композиції таких правил. Тому необхідно шукати більш вдале рішення вказаних проблем шляхом складання нових конкуренто стійких розробок.

Аналіз існуючих методів та програмних комплексів

На сьогоднішній день основні методи ґрунтуються на добре формалізованих алгоритмах, які були отримані під час побудови математичних моделей для предметних областей. У більшості це трудомісткі розрахунки з відомих формул, або відносно прості послідовні дії, які приводять до потрібного результату упродовж довгих повторень(ітераційні алгоритми). На практиці актуальні задачі відносяться до типу, які важко формалізуються, особливо якщо це стосується природної мови, для якої немає аналітичних послідовностей або ланцюжка дій, якій би приводив до результату без інтелектуального втручання чоловіка [3].

До основи сучасної технології Data Mining покладена концепція шаблонів, які відражають фрагменти багатоаспектних взаємовідносин в даних. З їх допомогою рішаються задачі прогнозування, класифікації, розпізнання образів, сегментації бази даних, вилучення «прихованих» знань, інтерпретації даних та встановлення асоціацій в БД тощо. Результати таких алгоритмів ефективні та легко інтерпретуються [6].

Сучасні компанії та корпорації мають сильну інформаційну залежність. Розширення сфери послуг, розробка та впровадження нових технологій, а також виконання поточних бізнес-задач тісно пов’язане зі збором та обробкою різної інформації. При цьому важливі аспекти даної проблеми, як максимальна повнота інформаційного масиву, автоматизація процесу збору інформації, а також забезпечення засобів навігації в ній. Архітектури існуючих програмних засобів націлені на вирішення приватних пошукових завдань і дозволяють здійснювати вузький пошук за безпосередньої участі користувача [4].

На цей час існує потреба в опрацюванні цілого ряду аспектів щодо функціонування систем інформаційного пошуку, серед яких:

- автоматизація процесу інформаційного пошуку;
- реалізація комплексного підходу до рішення інформаційного пошуку в мережі Інтернет, який дозволяє задати усю область пошуку та забезпечити її уточнення під час роботи [3];

У даний час у вільному доступі немає аналога програми, яка витягає дані з мережі Інтернет на тематику футболу. Існує безліч програм, які вміють вилучати дані з HTML – сторінок. Однак ці програми або вилучають однотипну інформацію (номера телефонів, факсів, e-mail, ключові слова, URL адреси) або для витягання даних необхідно створювати правила сканування а це є дуже складним заняттям.

Опис програмної системи витягання знань

Розроблюваний програмний комплекс вилучає усі недоліки своїх аналогів, він має добре продуману базу даних, яка може містити основі аспекти теми футболу. Також він має паралельну обробку інформації, процес витягання даних з HTML – сторінок відбувається автоматично. Користувач має можливість переглядати інформацію з бази даних у зручному вигляді, завдяки SQL – запитам, які направлені на огляд наважливої інформації. Користувач програмного комплексу може бути адміністратором, якому дозволяється вносити інформацію до бази даних власноруч та також звичайним юзером, який може лише переглядати інформацію.

Програмний комплекс включає:

- Програмні компоненти, які послідовно рішають задачу отримання документів з мережі Інтернет відповідно етапів формування пошукових задач, запитів, пошук документів та їх завантаження, обробка і зберігання (для цього використовувався простір імен System.Net та об’єкти класу HttpWebRequest);
- Метод автоматичного формування запитів к базі даних, який дозволяє створювати потік інформації для подальшої обробки відповідно із заданою інформаційною

потребою. Для цього використовується ADO.NET [5]. ADO.NET представляє узгоджений доступ до SQL Server, XML, OLE DB, ODBC. ADO.NET розділяє доступ к даним та обробку даних на дискретні компоненти, які можуть використовуватися окремо або сумісно. Технологія ADO.NET включає постачальників даних .NET Framework для з'єднання з базою даних, виконання команд а також отримання результатів. Для зручного представлення даних користувачеві, дані поміщають до об'єкту ADO.NET DataSet, який займається обробкою інформації [2].

Опис бази даних

Оглянемо основні таблиці, у яких зберігається важлива інформація.

Таблиця чемпіонатів складається з поля ChampionshipName, яке зберігає назву чемпіонатів та поля IdChampionship – це ідентифікатор чемпіонату.

Таблиця клубів зберігає інформацію о назві клубу, цілях клубу на сезон, бюджет клубу, дату заснування, старе ім'я клубу та його веб-сайт.

У кожного клуба, який грає у чемпіонаті повинен бути стадіон на якому він буде проводити домашні матчі. Для цього є таблиця стадіонів, яка містить інформацію о назві, місткості стадіону а також його адресу. На стадіонах проходять матчі між командами, які змагаються у чемпіонаті.

Для зберігання інформації о матчах є таблиця матчів. Вона містить номер туру, ідентифікатор команд, які грали, дату матчу, ідентифікатор стадіону, результат матчу, інформацію о суддях, веб-сторінку протоколу матчу а також додаткову інформацію.

Для обслуговування матчу комітет федерації футболу України делегує команду арбітрів. Інформація о суддях зберігається у таблиці судді. Таблиця складається з імені арбітра, національності, типу судді(арбітр у полі, 1-й асистент, 2-й асистент, тощо), категорії арбітру,які матчі він може обслуговувати, вік та дата народження арбітру.

Одна з найважливіших таблиць – це таблиця гравців. Ця таблиця включає в себе інформацію о імені гравця, позиції, віку, даті народження, фізичній формі, травмах, дискваліфікаціях, національності, умовах контракту, номеру гравця та о його антропометричних даних.

Загалом це мала частина інформації, яка може зберігатися в розробленій базі даних. Окрім основних таблиць є ще багато інших допоміжних таблиць, які необхідні для того, щоб користувач мав можливість отримати усю цікаву для нього інформацію.

Реалізація вилучання даних

Для вилучення інформації з HTML – сторінок використовується парсер HTML коду. Він витягає дані з шаблону, який функція приймає на вході. Необхідні парсеру дані зберігаються у різних файлах, це дозволяє програмному комплексу бути незалежним від змін у структурі сайтів.

З початку програмний комплекс буде містити інформацію о чемпіонаті України (клубах, турах, матчах, суддях, гравцях, стадіонах, тренерах). За для розширення області пошуку інформації необхідно буде включити чемпіонат, який буде цікавий користувачеві.

Також програмний комплекс може отримувати інформацію з RSS – новин, за для цього необхідно неструктуровану інформацію структурувати та занести в необхідну таблицю. Це дуже складний процес, тому що за для його виконання необхідно використовувати методи штучного інтелекту. За для вирішення цієї проблеми буде використовуватися таблиця, у якій буде зберігатися інформація щодо співвідношення тематики новини та таблиці, у яку вона повинна бути записана. Для бистрого зростання бази даних необхідно брати новини з різних сайтів. За для скорочення часу на обробку даних з RSS – новин необхідно використовувати паралельну обробку. Кожному процесу буде визначений сайт, з якого він має отримувати дані. Потім вилучені дані необхідно звіряти, щоб уникнути повторюваної інформації. Алгоритм вилучання даних зображений нижче.

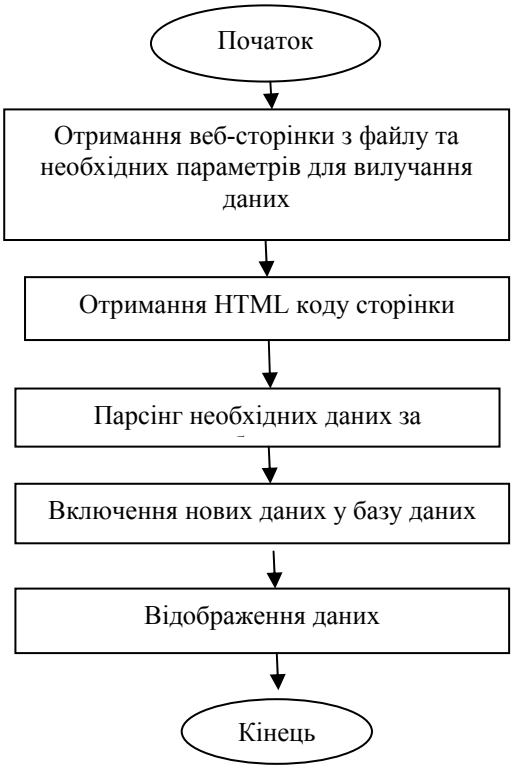

Рисунок 1 – Схема алгоритму вилучання даних

Висновки

В результаті у програмному комплексі реалізовані програмні компоненти, що дозволяють отримувати HTML – документи з мережі Інтернет, архітектура і власне сама база даних для зберігання и витягання даних. Також реалізовані методи автоматичного формування запитів до бази даних, для ефективного і швидкого доступу к даним. Програмний комплекс має інтуїтивно зрозумілий інтерфейс.

Розроблюваний програмний комплекс є дуже конкуренто стійким. Він має очевидну перевагу над своїми аналогами.

Список літератури

1. Вієйра Р. Програмування баз даних Microsoft SQL Server 2005. Базовий курс [Текст] / Вієйра Р. // Пер. з англ. – М.,2007. – 832 с.
2. ADO.NET [Electronic resourse] / Інтернет-ресурс. – Электронная документация по MS Visual Studio, 2010. – Режим доступу: http://msdn.microsoft.com/ru-ru/library/e80y5yhx(v=VS.90).aspx
3. Осіпов Г.С. Придбання знань інтелектуальними системами: Основи теорії та технологи [Текст] / Осіпов Г.С. // – М.: Наука, Физматлит, 1997.
4. Плотнікова С.В. Формальні методи оцінки ефективності систем автоматичної обробки тексту [Текст] / Плотнікова С.В., Тіхоміров В.Г. // - М.: Ера 2003. - 447 с.
5. Сеппа Д. Microsoft ADO.NET [Текст] / Сеппа Д. // Пер. з англ. – М.,2003. – 640с.
6. Кречетов Н. Продукти для інтелектуального аналізу даних [Текст] / Кречетов Н. // — Ринок програмних засобів, СПб: Пітер, 1997. - 320 с.