Система витягу та розміщення даних

Vira Shendryk[1], Kateryna Omelianenko[2], Vitaliy Hapon[3]

1. The Department Information Science, The Section of Informational Technologies of Design, Sumy State University, UKRAINE, Sumy, Rimsky-Korsakov Street 2, E-mail: ve-shen@opm.sumdu.edu.ua

2. The Department Information Science, The Section of Informational Technologies of Design, Sumy State University, UKRAINE, Sumy, Rimsky-Korsakov Street 2, E-mail: centr_kit@opm.sumdu.edu.ua

3. NetCracker Technology, UKRAINE, Sumy, E-mail: hvitalii@gmail.com

Питання про володіння інформацією стає все більш актуальним, адже підприємства та організації змушені працювати з великими обсягами інформації, джерелом якої може служити всесвітня мережа. Накоплена інформація з надзвичайною швидкістю помножується та змінюється. Використання та поширення значних масивів різноманітної інформації, спонукає до створення нових систем автоматизації щодо збору, упорядкування та подальшого аналізу необхідних даних.

Першочергове завдання такої системи полягає у зчитуванні та групуванні неструктурованої інформації, представленої у всесвітній мережі Інтернет, до реляційних баз даних.

В даній роботі запропановано універсальний алгоритм зчитування табличної інформації з Інтернет-сторінок, у вигляді програми, яка його реалізує. Зчитавши дані, дана підсистема передає керування іншому модулю – СУБД, який створює необхідні метадані. Інформацію, яку система «збирає» та заносить до бази даних, передається іншій підсистемі для подальшого аналізу.

Описані модулі створюють нову інформаційну систему, яка надає можливість користувачам полегшити збір та обробку даних.

Створена програма дозволяє виконувати оперативний моніторинг змін на Інтернет сторінках. Якщо відбулися зміни в необхідних даних, то вона самостійно зчитує їх, групує та підготовляє для аналізу, заносить до бази даних. В подальшому данні можуть бути використані аналізуючою системою, яка оперативно проводить аналіз та видає необхідні результати.

Програма стане актуальною для підприємств, які використовують в економічному аналізі дані з всесвітньої мережі.

Система витягу та розміщення даних

Віра Шендрик[1], Катерина Омеляненко[2], Віталій Гапон[3]

1. Секція інформаційних технологій проектування кафедри інформатики, Сумського державного університету, УКРАЇНА, м.Суми, вул.Римського-Корсакова, 2, E-mail: ve-shen@opm.sumdu.edu.ua

2. Секція інформаційних технологій проектування кафедри інформатики, Сумського державного університету, УКРАЇНА, м.Суми, вул.Римського-Корсакова, 2, E-mail: centr_kit@opm.sumdu.edu.ua

3. NetCracker Technology, УКРАЇНА, Суми E-mail: hvitalii@gmail.com

В статті розглянуті особливості структури Web-сторінок, запропановано метод структурування неструктурованої інформації з Інтернет сторінок та підсистему парсингу. Зчитавши дані, дана підсистема передає керування іншому модулю – СУБД, який створює необхідні метадані. Зібрана інформація передається іншій системі для подальшого аналізу. Описані модулі, об'єднані в нову інформаційну систему, яка надає можливість користувачам полегшити збір та обробку даних. Результатом проведеної роботи є програмний комплекс на основі розробленого методу, який дає змогу зчитувати табличну інформацію з Інтернет сторінок та аналізувати її Створена програма дозволяє виконувати оперативний моніторинг змін на Інтернет сторінках.

Ключові слова – HTML, DOM, ПАРСЕР, СУБД, БД.

I. Вступ

У сучасних інформаційних технологіях роль такої процедури, як витяг інформації, усе більше зростає - через стрімке збільшення кількості неструктурованої інформації, зокрема, в Інтернеті. Першочергове завдання при створенні систем обробки інформації полягає у зчитуванні та групуванні неструктурованої інформації, представленої у всесвітній мережі Інтернет, до реляційних баз даних.

Об'єкт дослідження – структура web-сторінок з неупорядкованою табличною інформацією. Предмет дослідження – метод розпарсювання неупорядкованої інформації з web-сторінок.

Мета роботи полягає у створенні нового виду універсального парсингу незгрупованої та неструктурованої HTML-інформації у вигляді таблиць для виділення необхідних даних та переведення їх у згруповану структуру баз даних, проведенні аналізу даних.

II.Аналіз структури web-сторінок

Витяг інформації полягає в скануванні набору документів, написаних на мові HTML, та заповненні баз даних виділеною корисною інформацією. Будь-який документ на мові HTML є набором елементів, які позначаються спеціальними позначками (тегами). Атрибути вказуються в відкриваючому тегу. Ім'я тегу визначає тип елемента. Також у мові гіперпосилання

спостерігається вложення елементів більш високого рівня. HTML є неструктурованимй текстом, що не дає можливість обробити документ: виконати трансформацію даних, пошук потрібних елементів документа і т.д. HTML не має чіткої ієрархічної структури, але можна виділити окремі теги та блоки, які знаходяться в чіткій ієрархії.

III.Принципи створення синтаксичного аналізатора

Синтаксичний аналізатор (парсер) — це програма або частина програми, яка виконує синтаксичний аналіз. Під час парсингу текст оформлюється у структуру даних, зазвичай — в DOM-дерево, яке відображає синтаксичну структуру вхідної послідовності, та добре підходить для подальшої обробки.

Розглянемо табличні елементи HTML. Серед табличних елементів TR визначає число рядків, тоді як TH й TD визначають число стовпців в HTML таблиці. Елемент TH використовується для завдання одного або більше заголовків. Елемент TD використовується для внесення даних в комірки таблиці. Будемо надалі називати дані в елементах TD табличними даними на відміну від даних, що знаходяться в елементах TH, які будемо називати заголовками. Типова HTML таблиця має, як мінімум, один стовпець - заголовок у верхній частині таблиці, і як мінімум, один рядок заголовок у лівій частині. Такий тип таблиць назвемо строково-стовбовим. Інший тип таблиці містить, як мінімум, один стовпець заголовок або один рядок заголовок і називається в цьому випадку стовбовим або рядковим типом таблиці. Заголовки в рядкових та стовбцевих таблицях задають схему таблиці. Для будь-яких таблиць, які не мають елементів TH, у ході аналізу було виявлено, що перший рядок використається як заголовок. Семантична ієрархія HTML таблиці визначається відповідно до нотації псевдотаблиці. Псевдотаблиця може розглядатися як особливий тип HTML таблиці та може бути використана для вираження строково-стовбцевих, рядкових і стовбцевих таблиць. До псевдотаблиці можна звертатися за індексами. Із кожною таблицею може бути зв'язаний заголовок. Рядки таблиці можуть групуватися в розділи заголовків, нижні заголовкі і тіла. При відображенні довгих таблиць інформація із заголовків може повторюватися на кожній сторінці таблиці.

Отримане значення потрібно занести до бази даних для подальшого аналізу.

IV.Система збору та розміщення інформації

Система складається з декількох окремих компонентів, що дозволяє гнучко налаштувати параметри збору інформації для баз даних різного структурного та інформаційного наповнення, а саме:

- Модуль зчитування - здійснює зчитування web-сторінок і файлів відповідно до завдання;
- Модулі структурування - перетворює данні неструктуровані в структуровані;
- База даних (БД), у якій зберігаються результати зчитування та перетворення;
- Планувальник - управляє процесом збору даних: формує завдання на зчитування і обробку відповідно до настроювань;
- Аналітичний модуль – модуль, який дає можливість проводити аналіз даних переданих для аналізу підсистемою системи управління базою даних (СУБД).

Модуль зчитування (парсер) дає доступ до сайту в режимі зчитування, після чого він розбирає елементи сторінки, знаходить вказану таблицю та зчитує дані, які передаються на обробку системі СУБД. Парсер отримує доступ до документу та кореневого елементу дерева тегів. Потім розпізхнаються елементи, які містяться в структурі під кореневим елементом. Якщо вони є, то зчитуються, звіряються чи є той елемент потрібною таблицею, якщо так, то вибираються з неї всі потрібні дані, інакше перевіряється чи має поточний елемент дочірні, якщо так, то він стає поточним елементом і запускається функція рекурсивно. Цей процес повторюється доти, доки не буде знайдено потрібну таблицю або елемент, який не має дочірніх елементів. Після цього парсер повертається до того елементу, де почалося розгалуження. Коли парсер знаходить потрібну таблицю для зчитування, він передає керування наступній функції, яка аналізує структуру таблиці, та створює псевдотаблицю. З псевдотаблиці можна звертатися до елементів за індексом рядка та стовпчика. Також парсер враховує ситуацію, коли таблиця вставлена в іншу таблицю. При цьому дані передаються на вищий рівень в основну комірку псевдотаблиці. Система має планувальник, який дає змогу проводити зчитування та аналіз в автоматичному режимі через заданий час.

Висновок

В роботі була запропонована система, яка надає можливість отримати доступ до сайту з необхідною інформацією. Ця програма містить влаштований парсер, який створює чітку ієрархію елементів, аналізує дерево елементів та вишукує необхідну інформацію, трансформує її в допустимий вигляд для зберегання в базі даних та заповнює нею таблиці. Програма працює із усіма сайтами, не залежно від їх структури та має можливість заносити інформацію в бази даних із будь-якими таблицями або створювати нові, в залежності від потреб аналізу.

Література

[1] Chawathe S., Rajaraman A., Garcia Molina H., Widom J. Change detection in hierarchically structured information // Proc. of the ACM SIGMOD Int. Conf. on Management of Data, Montreal, Quebec, 1996. V. 25. № 2. P. 493–504.