

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

**ДВНЗ «ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ
УНІВЕРСИТЕТ»**

**ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
(М. ВІННИЦЯ, УКРАЇНА)**

**ШТУТГАРТСЬКИЙ УНІВЕРСИТЕТ
(М. ШТУТГАРТ, ФРН)**

**ТЕХНІЧНИЙ УНІВЕРСИТЕТ ГАМБУРГ-ХАРБУРГ (М. ГАМБУРГ,
ФРН)**

**ІНСТИТУТ ПРОБЛЕМ МОДЕЛЮВАННЯ В ЕНЕРГЕТИЦІ ім.
Г.Є.ПУХОВА НАН УКРАЇНИ**

М А Т Е Р І А Л И

Сьомої міжнародної науково-технічної конференції

"МОДЕЛЮВАННЯ І КОМП'ЮТЕРНА ГРАФІКА"

**18-24 вересня 2017 року
м. Покровськ, м. Київ**

УДК 004.3+004.9+004.2+51.7+519.6+519.7

Публікується згідно з рішенням Вченої ради ДВНЗ «Донецький національний технічний університет» (протокол № 1 від 07.09.2017).

Збірник містить наукові статті співробітників ДонНТУ та інших навчальних і наукових закладів України, які взяли участь у роботі Сьомої міжнародної науково-технічної конференції "Моделювання і комп'ютерна графіка", що проводилась 18-24 вересня 2017 року у ДВНЗ «Донецький національний технічний університет» у м. Покровську. Публікації висвітлюють результати наукових досліджень і розробок в таких напрямках, як інформатика, чисельні методи, паралельні обчислення, програмування, розробка засобів обчислювальної техніки, дослідження комп'ютерних мереж, машинна графіка і обробка зображень, математичне моделювання в різних галузях. Матеріали збірника призначені для наукових співробітників, викладачів, інженерно-технічних працівників, аспірантів та студентів.

Видавець – Донецький національний технічний університет (ДонНТУ)

Конференція проводилась за підтримки:

- гранту Erasmus+ CBHE action 561728-EPP-1-2015-1-ES-EPPKA2-CBHE-JP «GameHub: University-enterprises cooperation in game industry in Ukraine»,
- гранту Tempus 2013–4587–001–001- 544010-TEMPUS-DE-TEMPUS-JPHES TATU – «Trainings in Automation Technologies for Ukraine»,
- компанії QuartSoft,
- компанії PHOENIX CONTACT.

Адреса редакції: 85300, м. Покровськ, пл. Шибанкова, 2, ДонНТУ.

Тел.: (06239) 2-09-38. E-mail: natalia.kostiukova@donntu.edu.ua

Масюк А.Л., Мирошкин А.Н. МУЛЬТИПЛАТФОРМЕННЫЙ РЕДАКТОР ДВУМЕРНЫХ ГРАФОВ ОБЩЕГО НАЗНАЧЕНИЯ В КОНТЕКСТЕ МОДЕЛЬНОЙ ПРЕДМЕТНОЙ ОБЛАСТИ ГРАФОВ ШАХТНЫХ ВЕНТИЛЯЦИОННЫХ СЕТЕЙ	78
Фельдман Л.П., Назарова І.А., Гризадубова Я.А., Костін В.І. ЕФЕКТИВНІСТЬ ОБЧИСЛЕНЬ ПРИ ПАРАЛЕЛЬНОМУ РОЗВ'ЯЗАННІ ЗАДАЧІ КОШІ НА ГРАФІЧНИХ ПРОЦЕСОРАХ ЕКСТРАПОЛЯЦІЙНИМИ МЕТОДАМИ	84
Ануфрієв П. О. ПРОГРАМНА СИСТЕМА ДЛЯ ВИВЧЕННЯ ЛЕКСИКИ АНГЛІЙСЬКОЇ МОВИ	92
Федоров Є.Є., Патрушев В.О., Патрушева О.І. МЕТОД ПРОГНОЗУ ПРИБУТКУ ІНТЕРНЕТ МАГАЗИНУ НА ОСНОВІ ПОВНОЗВ'ЯЗНОЇ ДОВГОКОРОТКОЧАСНОЇ ПАМ'ЯТІ	95
Псьол В. О., Маслова Н. О. АЛГЕБРАЇЧНИЙ АНАЛІЗ ШИФРІВ, ОСНОВАНИХ НА КРИПТОГРАФІЧНИХ ПЕРЕТВОРЕННЯХ В АЛГЕБРІ ПОЛІВ ГАЛУА	108
Тищенко Г. В., Святний В. А., Бищенко Ю. А. СИСТЕМИ ЗБОРУ ДАНИХ З ВЕБ-РЕСУРСІВ	112

Секція 2.

High Performance Computing (HPC) and Parallel Simulation Technology (ParSimTech)

Resch M. BRINGING TOGETHER HIGH PERFORMANCE COMPUTING AND DATA ANALYTICS	117
Mangold M., Khlopov D., Lihong Feng L. NONLINEAR MODEL REDUCTION OF POPULATION BALANCE EQUATIONS BY POD-DEIM BASIS CONSTRUCTION	117
Погорілий С.Д., Бойко Ю.В., Сальніков А.О., Слюсар Є.А., Борецький О.Ф. ВИКОРИСТАННЯ СПЕЦІАЛІЗОВАНОГО ПРИКЛАДНОГО ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ RAINBOW У НАУКОВО- ДОСЛІДНИХ ГРІД-ПРОЕКТАХ	123

Секція 3.

Комп'ютерна графіка та візуалізація в ігрових додатках

Драченко М.Ю., Зорі С.А. ДОСЛІДЖЕННЯ МОЖЛИВОСТЕЙ ПАРАЛЕЛЬНОЇ РЕАЛІЗАЦІЇ АЛГОРИТМУ БІЛАТЕРАЛЬНОЇ ФІЛЬТРАЦІЇ ДЛЯ ВИРШЕННЯ ЗАВДАНЬ РОЗПІЗНАВАННЯ ОБРАЗІВ НА ЗОБРАЖЕННЯХ ЗАСОБАМИ ТЕХНОЛОГІЇ CUDA	130
Глинська К.С. РЕАЛІЗАЦІЯ ШТУЧНОГО ІНТЕЛЕКТУ У СТРАТЕГІЧНИХ ІГРАХ НА ПРИКЛАДІ ІГРОВОГО ДОДАТКУ «ШАХИ»	136

УДК 004.4

СИСТЕМИ ЗБОРУ ДАНИХ З ВЕБ-РЕСУРСІВ

Г. В. Тищенко, В. А. Святний, Ю. А. Бищенко

ДВНЗ «ДонНТУ», м. Покровськ

Проведено аналіз зростання обсягу даних в мережі Інтернет, розглянуто існуючі системи збору даних, показано, що їх архітектура потребує доробки. Розроблено архітектуру системи збору даних, яка допоможе позбавитися існуючих недоліків з використанням фреймворку Scrappy, бази даних PostgreSQL та платформи RabbitMQ.

Ключові слова: система збору даних, фреймворк Scrappy, база даних PostgreSQL, платформа RabbitMQ

У попередні роки відзначається експоненціальне зростання обсягу даних, також прогнозується продовження зростання в майбутньому. Тому задача написання системи програмного забезпечення для збору даних, яка ефективно накопичує дані високої якості, дуже важлива.

Кожна компанія зацікавлена в інформації про ринок, на якому вона працює. Для розвитку потрібні дані, які допомагають генерувати ці знання. Вони надають незаперечну перевагу над конкурентами, допомагаючи знаходити потенційних клієнтів, аналізувати реакцію споживачів на нові продукти, розміщувати більш ефективні стратегії маркетингу, а також розробляти більш якісні продукти.

Збір цієї інформації - складне завдання, коли основний підхід полягає в опитуванні людей. Однак, завдяки сучасним системам зв'язку, а особливо, інтернет-порталам, все більше даних надається у вільному доступі. У попередні роки відзначається експоненціальне зростання обсягу даних, також прогнозується продовження зростання в майбутньому. Основною перевагою збору є те, що дані, як правило, можуть бути отримані автоматично, за допомогою комп'ютерів.

Процес вилучення структурованої корисної інформації з сайту називається парсингом (parsing), а інструменти для реалізації даного процесу - парсерами (parsers).

Як правило, сайти розробляються з урахуванням того, що зчитувати інформацію з їх сторінок буде людина. Але формат представлених даних, зрозумілий людині, найчастіше не настільки зрозумілий програмним засобам. Крім того, структура представлених даних змінюється від сайту до сайту, тому не існує універсального засобу для вилучення інформації з них. Однак, існують готові

рішення, що дозволяють отримувати інформацію з сайту після попередньої конфігурації. Ці рішення часто коштують дорого і не мають тієї гнучкості, яку можуть дати рішення, розроблені під конкретний сайт.

Можна виділити наступні методи збору інформації з сайтів:

- а) Ручний - в даному випадку роль парсеру виконує людина.
- б) Гібридний - користувач як і раніше виконує основні дії для отримання інформації, але може використовувати допоміжні програмні засоби для автоматизації збору.
- в) Автоматичний - отримання та структурування інформації виконується автоматично.

Для того, щоб один раз отримати значення декількох полів з 10-15 сторінок, писати окремий парсер недоцільно, однак в разі великого числа сторінок, полів або високої періодичності збору інформації, цей процес потребує автоматизації [1]. Тому завдання написання системи програмного забезпечення для збору даних, яка накопичує дані високої якості, дуже важлива.

Web Mining - це процес отримання даних з веб-ресурсів. Так як веб-джерела, як правило, не є текстовими даними, то і підходи до процесу вилучення даних відрізняються в цьому випадку. В першу чергу необхідно пам'ятати, що інформація в інтернеті зберігається у вигляді спеціальної мови розмітки HTML (RSS, Atom та інші), веб-сторінки можуть мати додаткову метайнформацію, а також інформацію про структуру (семантику) документа [6].

Підходи до вилучення даних:

- а) аналіз DOM дерева;
- б) використання XPath;
- в) використання регулярних виразів;
- г) візуальний підхід.

Перший підхід базується на аналізі DOM дерева. DOM (Document Object Model - «об'єктна модель документа») - це не залежний від платформи та мови програмний інтерфейс, що дозволяє програмам та скриптам отримати доступ до вмісту HTML, XHTML і XML-документів, а також змінювати структуру та оформлення таких документів.

На рис. 1 зображено модель DOM - дерева.

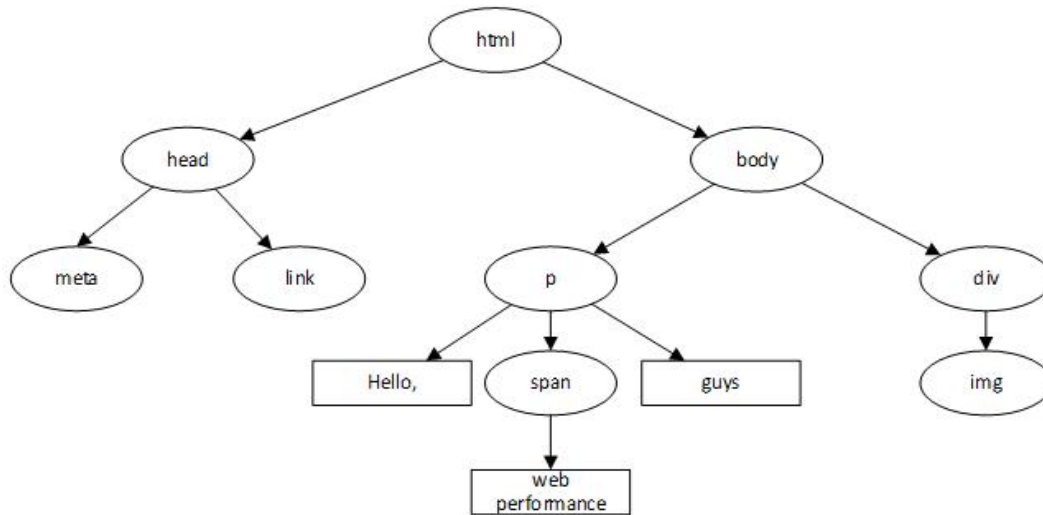


Рисунок 1 - Модель DOM – дерева

Використовуючи цей підхід, дані можна отримати безпосередньо за ідентифікатором, ім'ям або іншим атрибутом елемента дерева (таким елементом може служити параграф, таблиця, блок і т.п.).

Наступним еволюційним етапом аналізу DOM-дерева є використання XPath - це мова запитів до елементів xml- або xhtml-документа. Суть даного підходу в тому, щоб за допомогою деякого простого синтаксису описувати шлях до елемента без необхідності поступового руху вниз DOM-деревом.

Регулярні вирази – це система обробки тексту, заснована на спеціальній системі запису шаблонів для пошуку.

Візуальний підхід в даний момент знаходиться на початковій стадії розвитку. Суть підходу полягає в тому, щоб користувач міг без використання програмної мови або API «налаштувати» систему для отримання потрібних даних будь-якої складності та вкладеності [8].

У Web Mining можна виділити наступні етапи:

- а) вхідний етап (input stage) - отримання "сирих" даних з джерел;
- б) етап обробки (preprocessing stage) - дані подаються у формі, необхідній для успішної побудови тієї чи іншої моделі;
- в) етап моделювання (pattern discovery stage);
- г) етап аналізу моделі (pattern analysis stage) - інтерпретація отриманих результатів.

Існує багато систем для вилучення даних з інтернет-джерел. Найпопулярніші парсери - Datacol, Content Downloader, X-Parser Light, FDE Grabber, WP Uniparser.

Всі вони мають такі недоліки:

- а) працюють повільно;
- б) багато разів програма звертається в базу даних або в файл;

- в) багато часу займають налаштування;
- г) збирається менше 95% даних з порталів;
- ґ) іноді на виході отримуємо помилковий контент;
- д) виникають помилки при роботі з великими обсягами даних.

В результаті аналізу цих недоліків було розроблено архітектуру, зображену на рис.2.

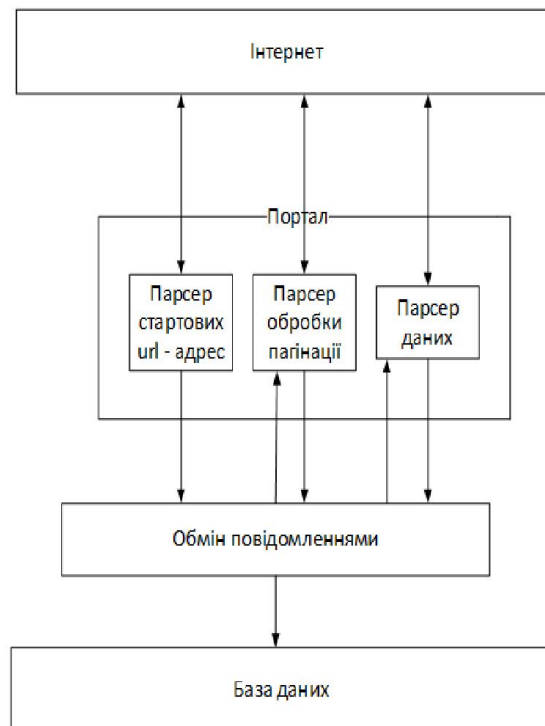


Рисунок 2 - Архітектура системи збору даних з веб-ресурсів

У результаті дослідження було виявлено, що обсяг даних в мережі Інтернет зростає за експоненціальною залежністю, тому необхідно розробляти та оптимізувати системи збору даних з веб-ресурсів. Розглянувши існуючі системи збору даних, можна зробити висновок про те, що їх архітектура потребує доробки. Тому, було розроблено архітектуру системи збору даних, яка допоможе позбавитися існуючих недоліків. Для реалізації такої системи було використано фреймворк Scrapy, база даних – PostgreSQL та платформа RabbitMQ.

ЛІТЕРАТУРА

1. Web parsing: задачи, проблемы, инструменты [Електронний ресурс]: «Inostudio». – Режим доступу: <https://inostudio.com/ru/article/web-parsing.html> – Дата доступу: березень 2017. – Загол. з екрану.

2. Data Mining [Електронний ресурс]: «UserGroup». – Режим доступу: <http://msugvnua000.web710.discountasp.net/Posts/Details/3316> – Дата доступу: березень 2017. – Загол. з екрану.
3. An introduction to big data [Електронний ресурс]: «Opensource». – Режим доступу: <https://opensource.com/resources/big-data> – Дата доступу: березень 2017. – Загол. з екрану.
4. Big Data and Data Mining [Електронний ресурс]: «Habrahabr». – Режим доступу: <https://habrahabr.ru/post/267827/> – Дата доступу: березень 2017. – Загол. з екрану.
5. Data Mining [Електронний ресурс]: «Habrahabr». – Режим доступу: <https://habrahabr.ru/post/95209/> – Дата доступу: березень 2017. – Загол. з екрану.
6. Document Object Model [Електронний ресурс]: «W3C». – Режим доступу: <https://www.w3.org/DOM/> – Дата доступу: березень 2017. – Загол. з екрану.
7. Подходы к извлечению данных с веб-ресурсов [Електронний ресурс]: «Habrahabr». – Режим доступу: <https://habrahabr.ru/post/99918/> – Дата доступу: березень 2017. – Загол. з екрану.
8. Подходы к извлечению данных с веб-ресурсов [Електронний ресурс]: «Habrahabr». – Режим доступу: <https://habrahabr.ru/company/dataart/blog/262817/> – Дата доступу: березень 2017. – Загол. з екрану.
9. Web Mining: интеллектуальный анализ данных в сети Internet [Електронний ресурс]: «Google». – Режим доступу: <https://sites.google.com/site/upravlenieznaniami/tehnologii-upravlenia-znaniami/text-mining-web-mining/web-mining> – Дата доступу: березень 2017. – Загол. з екрану.
10. Datacol – универсальный парсер с визуализацией сбора данных [Електронний ресурс]: «Profithunter». – Режим доступу: <https://www.profithunter.ru/obzory/datacol-universalnyi-parser-s-vizualizatsiei-sbora-dannyih/> – Дата доступу: березень 2017. – Загол. з екрану.
11. Datacol – парсер для сбора информации с сайтов [Електронний ресурс]: «Vlada-rykova». – Режим доступу: <https://vlada-rykova.com/parser-sajtov/> – Дата доступу: березень 2017. – Загол. з екрану.
12. Универсальный парсер контента – программа Content Downloader [Електронний ресурс]: «Seobook». – Режим доступу: <http://seobook.info/universalnyy-parser-kontenta/> – Дата доступу: березень 2017. – Загол. з екрану.
13. Перелинковка сайта [Електронний ресурс]: «Pr-cy». – Режим доступу: <http://pr-cy.ru/lib/saytostroenie/Perelinkovka-sayta-Kak-eto-sdelat> – Дата доступу: березень 2017. – Загол. з екрану.

Отримано 20.06.2017