Міністерство освіти і науки України
Національний технічний університет України
“Київський політехнічний інститут імені Ігоря Сікорського”
Факультет інформатики та обчислювальної техніки
Кафедра інформаційних систем та технологій

Лабораторна робота №2

Технології розроблення програмного забезпечення

«ДІАГРАМА ВАРІАНТІВ ВИКОРИСТАННЯ. СЦЕНАРІЇ ВАРІАНТІВ ВИКОРИСТАННЯ. ДІАГРАМИ UML. ДІАГРАМИ КЛАСІВ. КОНЦЕПТУАЛЬНА МОДЕЛЬ СИСТЕМИ»

Варіант 11

Виконала
студентка групи ІА-14
Літвін Юлія Олесандрівна

Перевірив:
Драган Михайло
Сергійович

Київ 2023

Завдання:

1. Ознайомитися з короткими теоретичними відомостями.

2. Проаналізуйте тему та намалюйте схему прецеденту, що відповідає обраній темі лабораторній.

3. Намалюйте діаграму класів для реалізованої частини системи.

4. Виберіть 3 прецеденти і напишіть на їх основі прецеденти.

5. Розробити основні класи і структуру системи баз даних.

6. Класи даних повинні реалізувати шаблон Репозиторію для взаємодії з базою даних.

7. Підготувати звіт про хід виконання лабораторних робіт. Звіт, що подається повинен містити: діаграму прецедентів, діаграму класів системи, вихідні коди класів системи, а також зображення структури бази даних.

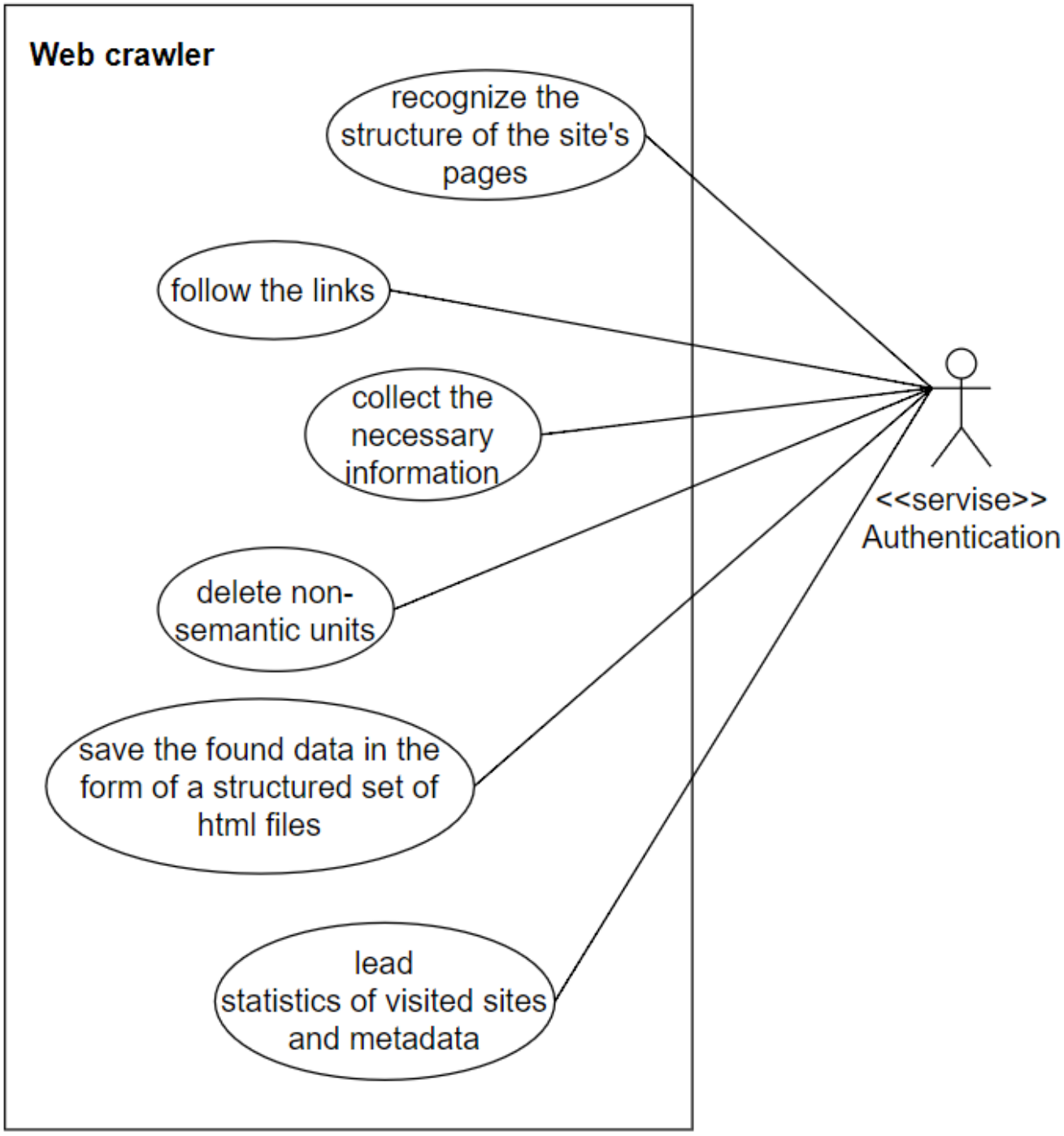
Варіант:

..11 Web crawler (proxy, chain of responsibility, memento, template method, composite, p2p)

Веб-сканер повинен вміти розпізнавати структуру сторінок сайту, переходити за посиланнями, збирати необхідну інформацію про зазначений термін, видаляти не семантичні одиниці (рекламу, об'єкти javascript і т.д.), зберігати знайдені дані у вигляді структурованого набору html файлів вести статистику відвіданих сайтів і метадані.

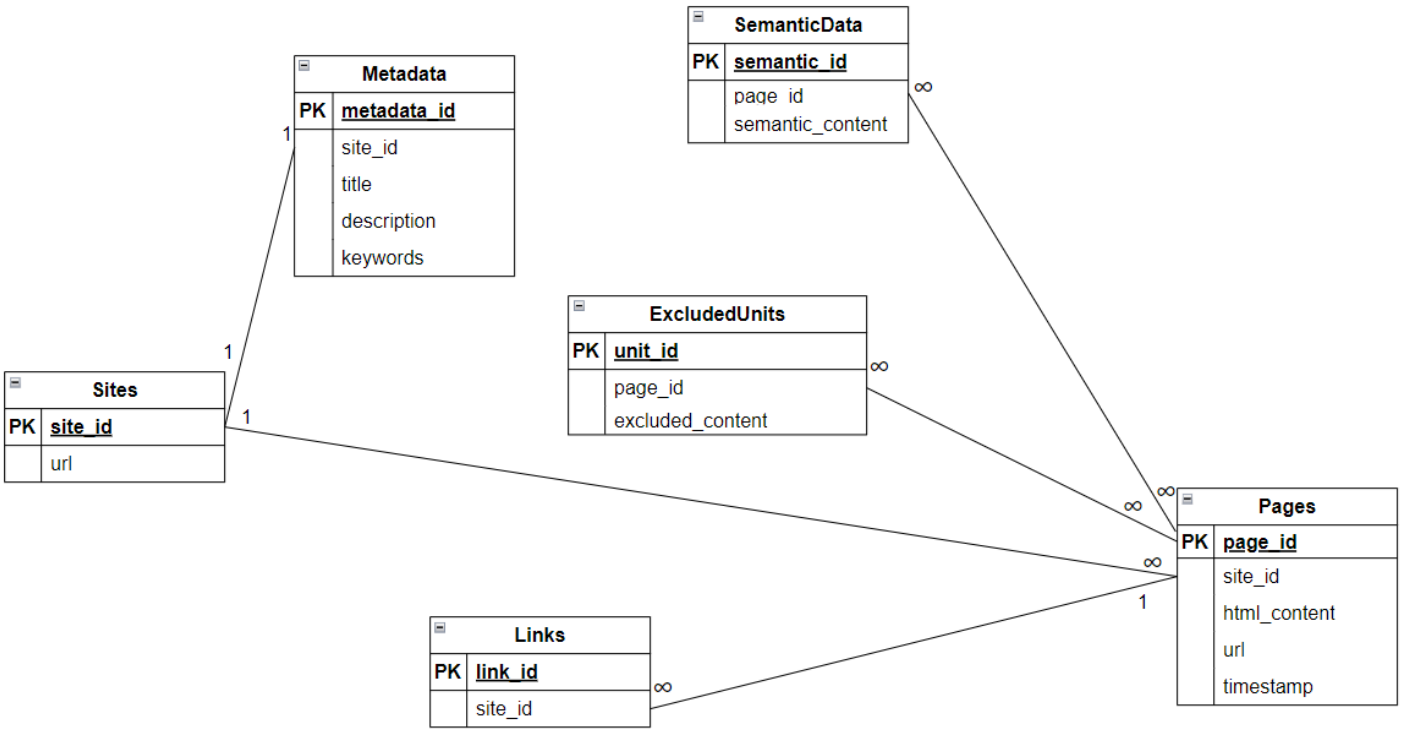
Хід роботи

Use-case діаграма



Веб-сканер розпізнає структуру сторінок сайту, переходить за посиланнями, збирає необхідну інформацію про зазначений термін, видаляє не семантичні одиниці (рекламу, об'єкти javascript і т.д.), зберігає знайдені дані у вигляді структурованого набору html файлів і веде статистику відвіданих сайтів та метадані.

Діаграма класів



Таблиця "Сайти" (Sites):

site_id (ідентифікатор сайту, PRIMARY KEY)

url (URL сайту)

Таблиця "Сторінки" (Pages):

page_id (ідентифікатор сторінки, PRIMARY KEY)

site_id (зовнішній ключ, посилання на таблицю "Сайти")

url (URL сторінки)

html_content (вміст сторінки у вигляді HTML)

timestamp (час сканування)

Таблиця "Посилання" (Links):

link_id (ідентифікатор посилання, PRIMARY KEY)

page_id (зовнішній ключ, посилання на таблицю "Сторінки")

Таблиця "Семантичні Дані" (SemanticData):

semantic_id (ідентифікатор семантичних даних, PRIMARY KEY)

page_id (зовнішній ключ, посилання на таблицю "Сторінки")

semantic_content (семантична інформація про зазначений термін)

Таблиця "Вилучені Одиниці" (ExcludedUnits):

unit_id (ідентифікатор вилученої одиниці, PRIMARY KEY)

page_id (зовнішній ключ, посилання на таблицю "Сторінки")

excluded_content (вміст вилученої несемантичної одиниці)

Таблиця "Метадані" (Metadata):

metadata_id (ідентифікатор метаданих, PRIMARY KEY)

site_id (зовнішній ключ, посилання на таблицю "Сайти")

title (заголовок сайту)

description (опис сайту)

keywords (ключові слова сайту)

Зв’язки:

Відношення "1 до багатьох" між таблицями "Сайти" і "Сторінки":

Один сайт може мати багато сторінок, тому це відношення "1 до багатьох".

Відношення "1 до багатьох" між таблицею "Сторінки" і таблицею "Посилання":

Одна сторінка може мати багато посилань, наприклад, посилання на інші сторінки на сайті, тому це відношення "1 до багатьох".

Відношення "багато до багатьох" між таблицями "Сторінки" і "Семантичні Дані":

Багато сторінок можуть мати багато семантичних даних, оскільки кожна сторінка може містити інформацію про різні терміни.

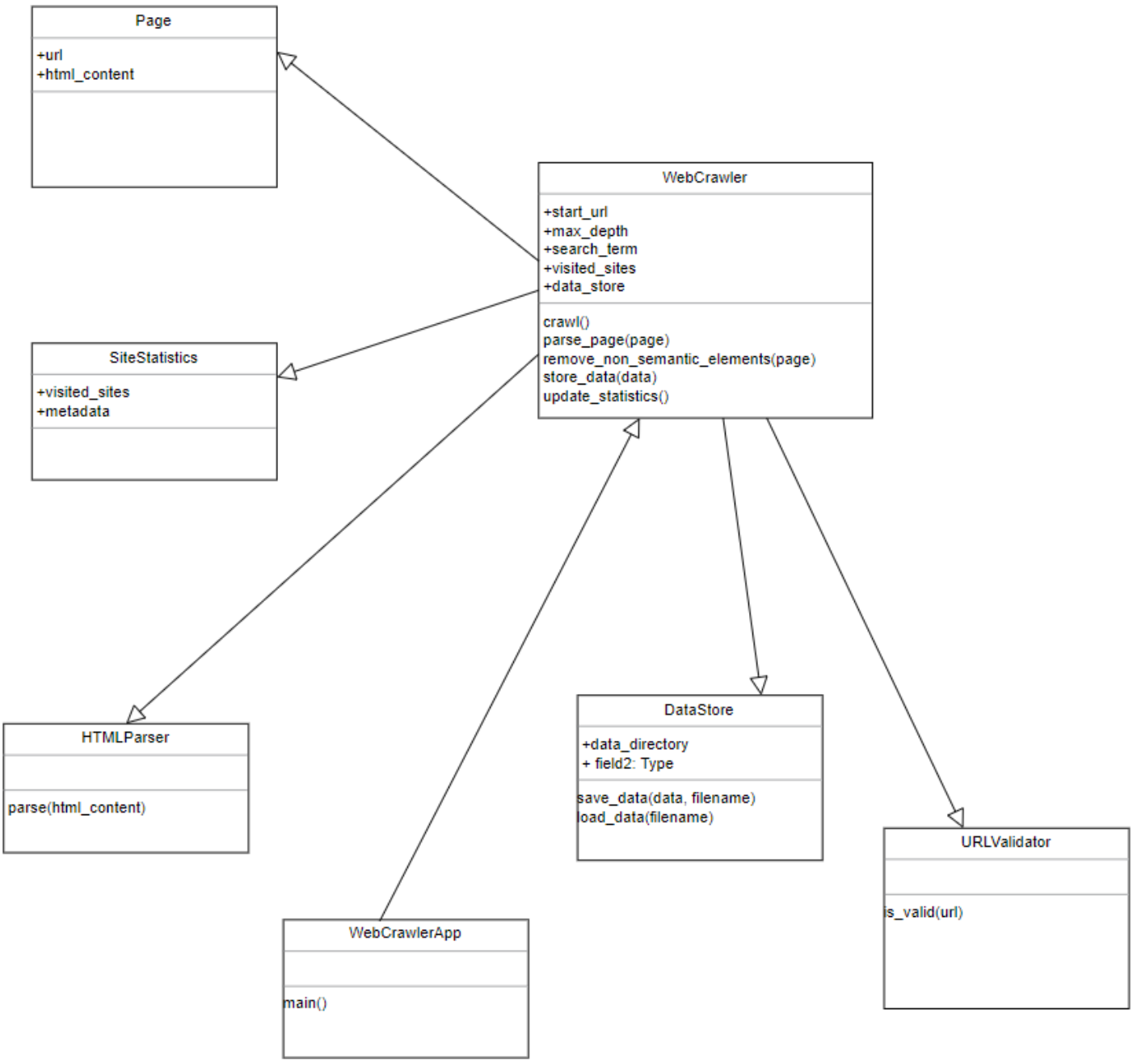
Відношення "багато до багатьох" між таблицею "Сторінки" і таблицею "Вилучені Одиниці":

Багато сторінок можуть мати багато вилучених несемантичних одиниць, оскільки різні сторінки можуть містити різні несемантичні елементи, які потрібно вилучити.

Відношення "1 до 1" між таблицями "Сайти" і "Метадані":

Кожен сайт має одну запис метаданих, і кожен запис метаданих відноситься лише до одного сайту, тому це відношення "1 до 1".

Структура бази данних



WebCrawler (Веб-сканер):

Атрибути:

start_url: URL-адреса, з якої починається сканування.

max_depth: Максимальна глибина сканування.

search_term: Термін, інформацію про який потрібно збирати.

visited_sites: Список відвіданих сайтів.

data_store: Об'єкт для зберігання зібраної інформації.

Методи:

crawl(): Основний метод для початку сканування. Він рекурсивно переходить за посиланнями, обробляє сторінки та збирає інформацію.

parse_page(page): Метод для аналізу HTML-сторінки та вилучення необхідної інформації.

remove_non_semantic_elements(page): Метод для видалення несемантичних елементів, таких як реклама та JavaScript-об'єкти.

store_data(data): Метод для зберігання знайденої інформації у вигляді структурованих HTML-файлів.

update_statistics(): Метод для оновлення статистики відвіданих сайтів та метаданих.

Page (Сторінка):

Атрибути:

url: URL-адреса сторінки.

html_content: HTML-вміст сторінки.

DataStore (Сховище даних):

Атрибути:

data_directory: Директорія для зберігання HTML-файлів та метаданих.

Методи:

save_data(data, filename): Метод для збереження даних у файл.

load_data(filename): Метод для завантаження даних з файлу.

SiteStatistics (Статистика сайту):

Атрибути:

visited_sites: Кількість відвіданих сайтів.

metadata: Метадані про сканування (дата, час, автор і т.д.).

HTMLParser (HTML-парсер):

Методи:

parse(html_content): Метод для парсингу HTML-вмісту та виділення необхідної інформації.

URLValidator (Валідатор URL):

Методи:

is_valid(url): Метод для перевірки валідності URL-адреси.

WebCrawlerApp (Додаток веб-сканера):

Метод main(): Основний метод додатку для налаштування та запуску веб-сканера.

Зв'язки між класами:

WebCrawler використовує Page для представлення сторінок, DataStore для зберігання даних та SiteStatistics для ведення статистики.

WebCrawler також використовує HTMLParser для парсингу HTML-сторінок і URLValidator для перевірки валідності URL.

DataStore використовується WebCrawler для збереження даних.

SiteStatistics веде статистику для WebCrawler.

Висновок: у ході виконання лабораторної роботи було проведено ознайомлення з теоретичними відомостями та розроблено прецеденти та діаграми класів для системи керування завданнями. Окрім того, підготовлений звіт включає всі необхідні компоненти, що відображають структуру розробленої системи.