Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів Web Mining для обробки даних в освітній сфері
(тема)

Виконав:
студент 2 курсу, групи СШІм-22-3
Пархацький І.С.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва спеціалізації)
Керівник Кудрявцева М.С.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Пархацькому Іллі Сергійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів Web Mining для обробки даних в освітній сфері

затверджена наказом університету від 1 квітня 20 24 р. № 260Ст
2. Термін подання студентом роботи до екзаменаційної комісії 7 червня 20 24 р.
3. Вихідні дані до роботи Науково-технічні публікації, дані статей, результати експериментальних досліджень по технологіям, методам, моделям

4. Перелік питань, що потрібно опрацювати в роботі
1) Вступ, мета роботи та постановка задачі, визначення бізнес-логіки
2) Теоретичні дослідження
3) Аналіз технологій та засобів реалізації додатку Web Mining

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	01.04.2024	виконано
2	Анализ предметної галузі та постановка задачі	10.04.2024	виконано
3	Теоретичні дослідження	20.04.2024	виконано
4	Проектування веб-додатку	30.04.2024	виконано
5	Розробка web-додатку	10.05.2024	виконано
6	Підготовка пояснювальної записки	31.05.2024	виконано
7	Попередній захист кваліфікаційної роботи	05.06.2024	виконано
8	Захист перед ЕК	07.06.2024	

Дата видачі завдання 1 квітня 20 24 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Кудрявцева М.С.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 72 с., 19 рис., 5 табл., 2 дод., 20 джерел.

КЛЮЧОВЕ СЛОВО, НОВИНА, ОСВІТА, ТОНАЛЬНІСТЬ, DATA MINING, MASHUP, WEB-ДОДАТОК, WEBРЕСУРС,WEB MINING.

Предмет дослідження – методи Web Mining для обробки даних в освітній сфері, а також поняття mashup додатків.

Об’єкт дослідження – сфера новин вищої освіти в Україні.

Мета дослідження – застосування методів Web Mining для даних, отриманих з веб-орієнтованого простору у сфері вищої освіти в Україні в рамках web-додатку, розробленого за технологією mashup.

Методами дослідження є вивчення можливих засобів застосування Web Mining для реалізації додатка новинного типу, розробленого за технологією mashup для обробки інформації з різних освітніх порталів і сайтів.

ABSTRACT

Master’s thesis contains: 72 pp., 19 fig., 5 tabl., 2 ann., 20 references.

DATA MINING, EDUCATION, KEYWORDS, MASHUP, NEWS, SENTIMENT, WEB APPLICATION, WEB MINING.

The subject of the research is Web Mining methods for data processing in the educational field, as well as the concept of mashup applications.

The object of research is the sphere of higher education.

The purpose of the study is to apply Web Mining methods to data obtained from a web-oriented space in the field of higher education in Ukraine within the framework of a web application developed using mashup technology.

Research methods include the study of possible means of using Web Mining for the implementation of a news-type application developed using mashup technology for processing information from various educational portals and sites.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛОВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

API – Application Programming Interface – інтерфейс прикладного програмування;

IE – Information Extraction – видобуток інформації;

IR – Information Retrieval – пошук інформації;

Java EE – Java Enterprise Edition – набір специфікацій для мови Java;

WWW – World Wide Web – всесвітнє павутиння, інтернет.

ВСТУП

В даний час для аналізу поведінки відвідувачів web-сайтів широко застосовуються системи обліку статистики відвідувань, які використовують рейтингові показники та критерії, побудовані на основі підрахунку кількості звернень до веб-сторінок. Вони відстежують базові числові параметри, такі як відвідуваність сайту за період часу, глибина перегляду сайту, джерела вхідного трафіку тощо. Зростання обсягу інформації, що збирається, і збільшується складність досліджуваних об'єктів вимагає автоматизації аналізу.

Необхідність автоматичного аналізу інформації з інтернету викликана високою доступністю величезної кількості інформації, що постійно поповнюється, а також зростаючою популярністю веб-послуг серед усіх категорій користувачів. Розвиток Інтернету в глобальну інформаційну інфраструктуру дозволило звичайним користувачам бути не тільки споживачами інформації, але її творцями та розповсюджувачами. У цьому для ефективного вирішення завдань пошуку, структурування та аналізу переважно хаотично організованої інформації у мережі призначено новий напрям у методології аналізу даних – Web Mining.

Системи Web Mining можуть відповісти на багато питань, наприклад, хто з відвідувачів є потенційним клієнтом Web-магазину, яка група клієнтів Web-магазину приносить найбільший дохід, які інтереси певного відвідувача чи групи відвідувачів.

Технологія Web Mining охоплює методи, які можуть на основі даних сайту виявити нові, раніше невідомі знання і які надалі можна буде використовувати на практиці. Іншими словами, технологія Web Mining застосовує технологію Data Mining для аналізу неструктурованої, неоднорідної, розподіленої та значної обсягу інформації, що міститься на Web-вузлах.

Web Mining розвивається на перетині таких дисциплін як знання в

базах даних, ефективний пошук інформації, штучний інтелект, машинне навчання та обробка природних мов. Напрямок Web Mining, званий Web Content Mining, охоплює методи, які здатні на основі даних сайту виявити нові, раніше невідомі знання, які надалі можна буде використовувати на практиці. Іншими словами, технологія Web Content Mining застосовує технологію Data Mining для аналізу неструктурованої, неоднорідної, розподіленої та значної за обсягом інформації, що міститься на Web-вузлах.

Викладені вище положення визначили тему кваліфікаційної роботи: «Дослідження методів Web Mining для обробки даних в освітній сфері ».

Об'єкт дослідження – сфера новин вищої освіти в Україні.

Предмет дослідження – методи Web Mining для обробки даних у освітній сфері, і навіть поняття mashup.

Мета дослідження – застосування методів Web Mining для даних, отриманих із веб-орієнтованого простору у сфері вищої освіти в Україні в рамках web-додатку, розробленого у форматі mashup.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Історична еволюція Web Mining

В Інтернеті міститься безліч знань та інформації. Така велика кількість часто створює труднощі при пошуку необхідної інформації. Подібні проблеми можуть мати різний характер, наприклад:

– користувач не завжди може відразу знайти необхідні йому джерела електронної інформації, так як не всі посилання ведуть туди, куди вказано, а не проіндексовану пошуковими системами інформацію таким способом і зовсім неможливо знайти;

– знайшовши безліч інформації, користувач часто відчуває труднощі для того, щоб отримати з неї корисні знання і зрозуміти їх;

– коли йдеться про вивчення інформації про споживачів, виникає необхідність надавати їм ті відомості, які їм цікаві – наприклад, давати підказки користувачеві при виборі потрібного товару. Все це призводить до необхідності якихось спеціальних технологій для отримання корисних знань з Інтернету.

Технологія Web Mining може успішно служити цим цілям.

Методи Web Mining є результатом тривалого процесу досліджень та розробки продукції. Ця еволюція почалася, коли бізнес-дані стали вперше зберігатися на комп'ютерах і в інтернеті, продовжилася з покращенням доступу до даних і завдяки технології реального часу, такої, що користувачі в WWW (World Wide Web) можуть переміщатися за даними.

У процесі еволюції з бізнес-даних до ділової інформації кожен новий крок створено на попередньому. Наприклад, здатність зберігати великі бази даних має вирішальне значення для Web Mining. З точки зору користувача, п'ять ступенів, перелічені в таблиці 1.1, були революційними, тому що вони дозволили відповісти точно і швидко на нові питання бізнесу.

Таблиця 1.1 − Етапи в еволюції Web Mining

Еволюційний крок	Бізнес-питання	Технології	Постачальники продукту	Характеристики етапів
Збір даних (1960-ті роки)	Яким був мій загальний дохід за останні п'ять років?	Комп'ютер, стрічки, диски	IBM, CDC	Ретроспективна, статична доставка даних
Доступ до даних (1980-ті роки)	Який був обсяг продажів у Делі минулого березня.	Реляційна база даних, структурована мова запитів (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Ретроспективна, динамічна доставка даних на рівні запису
Сховища даних та підтримка прийняття рішень (1990-і роки)	Який був обсяг продажів у Делі минулого березня? По Мумбаї.	OLAP, багатовимірні бази даних, сховища даних	Pilot, Comshare, Arbor, Congnos, Microstrategy	Ретроспективна, динамічна доставка даних на декількох рівнях

Продовження табл. 1.1

Еволюційний крок	Бізнес-питання	Технології	Постачальники продукту	Характеристики етапів
Data Mining (2000-ті роки)	Що, швидше за все, станеться з продажами у Мумбаї наступного місяця? Чому?	Удосконалені алгоритми, багатопроцесорні комп'ютери, великі бази даних	Pilot, Lockheed, IBM, SGI, численні startups (nascent industry)	Перспективна, проактивна доставка інформації
Web Mining(нині)	Що, швидше за все, станеться з продажами у Мумбаї у наступних/попередніх мільйонах місяців?	WWW, Інтернет, величезні бази даних	RockWare, Apteco Ltd, Simon Fraser University, IBM, Web Trends, SPSS, Flowerfire, Angoss, Net Genesis	Потужний, доступний інструмент для видобутку великих обсягів даних із реляційних баз даних, швидке та ефективне використання кількох функцій Web Mining

В основному Web Mining використовується як метод інтелектуального аналізу даних. Web Mining є розширеною версією Data Mining. Вилучення даних робота Off-Line, в той час як Web Mining робота On-Line [1]. В інтелектуальному аналізі дані зберігаються в базах даних, а Web Mining дані зберігаються в базі даних сервера і веб-журналі.

Основні компоненти технології Web Mining були у стадії розробки протягом десятиліть, у галузі досліджень, таких як штучний інтелект та машинне навчання. Сьогодні зрілість цих методів у поєднанні з високою продуктивністю реляційних баз даних та зусиллями з інтеграції даних робить ці технології практичними для зберігання даних.

Порівняння Web Mining і Data Mining [2] представлено таблиці 1.2.

Таблиця 1.2 – Web Mining та Data Mining

Порівнян-ня	Web Mining	Data Mining
Шкала	Обробка пошуку не велика.	Обробка пошуку велика.
Доступ	Web Mining надає публічний доступ до даних. Не приховуються дані, до яких здійснюється доступ до веб-бази даних, але це дозволяє дозвіл на доступ до даних.	Data Mining надає доступ до даних лише у приватному порядку та дозволяє користувачеві доступ до даних у базі даних.

Продовження табл. 1.2

Порівняння	Web Mining	Data Mining
Структура	Web Mining отримує інформацію із структурованих, неструктурованих та напівструктурованих веб-сторінок. Web Mining отримує інформацію з великої бази даних.	Data Mining отримує інформацію із явних структур. Data Mining не працює з інформацією з великих баз даних порівняно з Web Mining.

Web Mining на відміну Data Mining працює з трьома основними джерелами інформації:

– дані про дії користувачів від журналів серверів (server log) до відстеження звернень до браузера (browser activity tracking). Аналізуючи ці відомості, можна отримати узагальнені дані, вивчити закономірності поведінки груп користувачів та інформацію про відвідувачів сайту. Крім власників браузерів, збором таких даних займаються сотні спеціалізованих компаній, утворюючи великий сегмент бізнесу;

– веб-графи, що описують прямі зв'язки між сторінками WWW. Математичний граф складається з вершин, з'єднаних ребрами (дугами в орієнтованих графах). У веб-графі вершини це сторінки WWW, а дуги гіперпосилання між ними. За графом встановлюються зв'язки між сторінками, людьми та будь-якими іншими об'єктами;

– контент веб-сторінок та пов'язаних з ними документів.

1.2 Складності аналізу даних із мережі інтернет

Всесвітня мережа зараз містить величезну кількість інформації, знань. Користувачі на різних умовах можуть переглядати всілякі документи, аудіо- та відеофайли. Однак це різноманіття даних приховує у собі проблеми, які можуть виникнути не тільки під час аналізу, але й при пошуку необхідної інформації в Інтернеті:

– проблема пошуку потрібної інформації пов'язана з тим, що користувач не завжди одразу може знайти необхідні йому електронні ресурси. Лише невеликий відсоток посилань серед запропонованих пошуковими системами призводить до необхідних документів. Також важким є пошук неіндексованої інформації такими засобами;

– проблема виявлення нових знань. Навіть якщо знайдено безліч інформації, для користувача отримання корисних знань є досить трудомістким і непростим завданням. Сюди ж можна й віднести складнощі, пов'язані з осмисленням відомостей, поняттям тих ідей, вкладених авторами;

– проблема вивчення споживачів пов'язані з наданням користувачеві інформації, яка б йому цікава. Це особливо актуально для електронних торгових порталів, які могли б підказувати користувачеві при виборі товару.

1.3 Недоліки існуючих підходів

До очевидних недоліків існуючих підходів щодо аналізу даних у мережі інтернет можна віднести такі [3]:

– час відгуку надто довгий;

– вибухове зростання Інтернет наклав великий попит на мережі;

– ресурси та веб-сервера;

– очевидним рішенням для покращення якості веб-служб буде

збільшення пропускної спроможності, але такий вибір передбачає збільшення економічних витрат;

– схема веб-кешування має істотний недолік: якщо проксі-сервер неправильно оновлюється, користувач може отримати застарілі дані;

– очевидними чинниками є обмежені системні ресурси серверів. Однак, навіть якщо кеш не обмежений, існують значні проблеми;

– головним недоліком системи є те, що деякі попередньо обрані об'єкти не можуть бути в кінцевому підсумку використані користувачами. У такому випадку попередня вибірка збільшує мережевий трафік, а також навантаження веб-серверів.

1.4 Постановка задачі

Нині в Україні величезна кількість людей, пов'язаних із сферою освіти. До них належать студенти, батьки, викладачі. Але за такої кількості людей, зацікавлених цією областю, зараз відсутня єдина точка доступу до нової інформації у сфері освіти. Так як вона подається різного роду сайтами, порталами, при цьому часом може бути спотворена авторською точкою зору або недоліком інформації. Інформаційний ресурс, який об'єднає в собі новини з різних сайтів, однозначно набуде популярності, оскільки можливість отримати всю інформацію з різних точок зору в одному місці дозволить витрачати менше часу на перевірку достовірності тих чи інших фактів та перемикання з одного сайту на пошук іншого.

Для досягнення поставленої мети необхідно вирішити такі завдання:

– вивчити теоретичні аспекти поняття Web Mining, що включає розгляд цілей Web Mining, його застосування, види, а також виявлення зв'язку Web Mining з іншими засобами аналізу даних;

– проаналізувати поняття mashup як сучасного виду програми, що набирає все більшої популярності. Аналіз сфер використання mashup, його архітектури;

– розглянути необхідність виділення ключових слів із тексту, для аналізу інформації, що витягується з новинної сфери освіти України;

– розглянути необхідність застосування алгоритмів аналізу тональності тексту для визначення ступеня емоційного забарвлення новин;

– спроектувати та реалізувати web-додаток з використанням технології mashup для отримання інформації з основних освітніх та новинних порталів України у сфері вищої освіти.

2 ТЕОРЕТИЧНІ ДОСЛІДЖЕННЯ

2.1 Технологія Web Mining

Глобально, цілі використання технології Web Mining зводяться до пошуку необхідної інформації та знань, незважаючи на недосконаліть пошукових систем; аналіз структур сегментів мережі, тобто структури посилань між різними сторінками та сайтами у конкретному мережевому сегменті (наприклад, використовується для аналізу цитування різних авторів); виявлення знань із веб-ресурсів – пошук ключових слів, загальних тем тощо; персоналізації інформації – створення веб-систем, які адаптуються під переваги користувача, наприклад, пропозицію схожих із вже купленими товарів; виявлення шаблонів у поведінці користувачів, щоб спрогнозувати наступні його дії та використовувати отримані знання для подальшої оптимізації сайту.

2.1.1 Цілі Web Mining

Технологія Web Mining може вирішувати в управлінні знаннями та бізнес-аналітиці такі конкретні завдання [4]:

– опис користувачів сайту;

– опис покупців до Інтернет-магазинів;

– опис типових сесій відвідувань сайту та навігаційних траєкторій користувачів;

– опис груп та сегментів відвідувачів;

– пошук залежностей під час використання сайту та його послуг.

Розглянемо зручний практичний приклад використання технології Web Mining. Компанія Google використовувала цей підхід, щоб зрозуміти, чи очікують користувачі всесвітньої мережі настання економічної кризи. У співпраці з професором економіки Університету Каліфорнії в Берклі Х.

Варіаном фахівці компанії створили інструмент Google Correlate, який призначений для відстеження статистики запитів у пошуковій системі, щоб накласти отримані результати на реальні економічні дані. З'ясувалося, що динаміка певних запитів майже повністю збігається з динамікою економічних величин, тобто її можна використовувати для прогнозування.

Ілюстрацією цієї закономірності послужив пошуковий запит «Посібник з безробіття», з'ясувалося, що його динаміка в Google збігається з динамікою кількості заяв, що подаються до служб зайнятості США.

Інструмент дозволяє користувачам завантажувати власні дані, а потім шукає пошукові запити зі схожою динамікою. Видання Forbes вирішило скористатися ним, щоби побудувати прогноз економічної активності в Україні. Було взято економічний показник «промислове виробництво». Виявилося, що його динаміка пов'язана із пошуковими запитами українських користувачів про кредити. Статистика запитів про кредитування одно чи трохи випереджає реальну статистику виробництва. У поточний період (осінь 2012 року), виходячи з цієї кореляції, можна говорити, що в Україні немає активного очікування економічної кризи.

2.1.2 Завдання Web Mining

До завдань, які вирішуються за допомогою Web Mining, відносять такі [6]:

а) пошук інформації. Для знаходження необхідної інформації користувачі користуються пошуковими ресурсами. При цьому часто використовуються прості запити за ключовими словами. Результатом виконання запиту є список сторінок, відсортований за певним індексом релевантності, що описує ступінь збігу результату із запитом. Однак існуючі пошукові механізми мають недоліки. Основним є низька точність результату, викликана недостатнім урахуванням семантичних зв'язків і контексту знайдених у тексті виразів. Індексація сегментів мережі, що

цікавлять, з використанням інтелектуального аналізу даних, що застосовує алгоритми математичної лінгвістики та обробки природних мов, є перспективним напрямком Web Mining в області пошуку інформації. Цікавий підхід описаний у статті Anupam Joshi, «Improving Web Search Engine Results Using Clustering»;

б) аналіз структури сегмента мережі. Цей метод полягає в аналізі структури посилань між різними веб-сторінками, внутрішніми та зовнішніми сайтами у виділеному мережевому сегменті. Поява цього методу була викликана необхідністю вирішення завдань, що виникають при аналізі соціальних мереж чи специфічних областей людської діяльності, або знань, наприклад, в аналізі цитування авторів [5]. Результатом такого аналізу може бути виявлений набір специфічних сторінок наступних типів:

– хаби – з такої сторінки посилання йдуть найбільш значущі ресурси у цій галузі знань чи «знайомства» з найбільш значущими користувачами соціальної мережі;

– авторитети – сторінки, на які посилаються велика кількість авторів на цю тематику або користувачі соціальної мережі, до «дружби» з якими прагне велика кількість користувачів. Топологія структури посилань представляється у вигляді спрямованого графа з позначеними вузлами відповідно до їх функціональної класифікації та дуг з вагами, що описують, наприклад, частоти переходів за посиланням. Для моделювання топології веб-посилань використовують кілька алгоритмів, наприклад, HITS;

в) виявлення знань із веб-ресурсів. Це завдання перетинається з проблемою пошуку інформації. Тільки тут дослідник вже має набір веб-сторінок, отриманих в результаті запиту. Далі потрібно зробити їхню обробку з погляду автоматичної класифікації, складання змістів, виявлення ключових слів та загальних тем. Виявлені знання можуть бути представлені у вигляді дерев, що описують структури документів або у вигляді логічних і семантичних виразів. Вирішення частини цих проблем пропонує Text

Mining – технологія автоматичного отримання знань у великих обсягах текстового матеріалу, заснована на поєднанні лінгвістичних, семантичних, статистичних та машинних методик, що навчаються;

г) персоналізація інформації [4]. Персоналізація веб-простору – завдання створення веб-систем, що адаптують свої можливості (навігація, контент, банери та інші рекламні пропозиції) під користувача на підставі зібраної та проаналізованої інформації про користувальницькі уподобання. Класичним прикладом може бути ресурс Amazon, на якому один раз замовивши дорогу книгу в твердій обкладинці, користувач починає регулярно отримувати пропозиції про купівлю подарункових видань на таку тематику. Інший приклад – на підставі аналізу кошиків замовлень користувача йому пропонуються товари, які він ніколи не замовляв, але які входять до кошика інших покупців, схожих з ним за транзакційною поведінкою. Для аналізу інформації про користувача слід найменшою мірою використовувати декларовану про себе інформацію, а скоріше ґрунтуватися на стійких шаблонах його «поведінки» в мережі – послідовності кліків усередині ресурсу, переходах на інші підресурси, періодах мережевої активності, здійснюваних покупках тощо;

д) пошук шаблонів у поведінці користувачів. Це завдання пов'язане з попередньою, але її метою є не адаптація ресурсу до переваг індивідуальних користувачів, а пошук закономірностей у шаблонах взаємодії користувача з веб-ресурсом з метою прогнозування наступних дій. Аналізовані дії користувачів можуть включати не тільки переходи за посиланнями, але й відправлення форм, прокручування сторінок, додавання до обраних сторінок і т.ін. Знайдені шаблони використовуються надалі для оптимізації структури сайту, вивчення цільової аудиторії та прямого маркетингу [6].

Розроблено безліч підходів до вирішення задачі з виявлення знань із шаблонів навігації користувачів.

З точки зору застосування алгоритмів інтелектуального аналізу даних при пошуку шаблонів поведінки користувача найчастіше використовуються

наступні методики [7]:

- кластеризація – пошук груп схожих відвідувачів, сайтів, сторінок тощо;
- асоціації – пошук спільно запитуваних сторінок, товарів, що замовляються;
- аналіз послідовностей - пошук послідовностей дій.

Найчастіше застосовується варіант алгоритму apriori, розробленого для аналізу частих наборів, але модифікованого виявлення частих фрагментів послідовностей і переходів.

Особливо цікавим є підхід кластеризації послідовностей – пошук груп користувачів зі схожими послідовностями дій. У першому етапі у цьому підході виділяються послідовності класифікованих дій користувача, наприклад, у межах однієї сесії. Потім підраховуються частоти переходів між різними діями для складання Марківського кола заданого порядку. На заключному етапі отримані Марківські ланцюги кластеризуються для виявлення груп зі схожими частотами переходів. Для прогнозування наступної дії користувача спочатку на підставі історії його дій у рамках сесії визначається група, до якої належить із найбільшою ймовірністю. Потім визначається дія, яка виконується з найбільшою ймовірністю у цій групі з урахуванням останніх дій користувача [8]. Для реалізації такого аналізу можна, наприклад, використовувати алгоритм Microsoft Sequential Clustering, який входить до Microsoft Analysis Services 2005/2008. Недоліком алгоритму Microsoft є те, що до цього часу реалізований алгоритм, який використовує Марківські ланцюги лише першого порядку.

Як приклад застосування методу аналізу послідовності дій можна навести завдання оптимізації рубрикації одного книжкового інтернет-магазину, проведену компанією spellabs. Було виявлено групу, що складається з користувачів, що переходять довгими шляхами за посиланнями на книги з різних рубрик і замовляють зрештою «ізотеричну» літературу, до цього окремо не виділену в рубрику. Так було виявлено

невраховану цільову аудиторію та оптимізовано структуру сайту.

2.1.3 Етапи та напрямки Web Mining

У Web Mining можна назвати такі етапи [8]:

– вхідний етап (input stage) – отримання «сирих» даних із джерел (логи серверів, тексти електронних документів);

– етап передобробки (preprocessing stage) – дані подаються у формі, необхідної для успішної побудови тієї чи іншої моделі;

– етап моделювання (pattern discovery stage);

– етап аналізу моделі (pattern analysis stage) – інтерпретація одержаних результатів.

Це спільні кроки, які потрібно пройти для аналізу даних мережі Інтернет. Конкретні процедури кожного етапу залежить від поставленого завдання. У цьому сенсі виділяють різні категорії Web Mining.

Web mining можна дати визначення як пошуку та аналізу корисної інформації в мережі Інтернет із застосуванням методів інтелектуального аналізу даних [9]. Цей напрямок включає три області дослідження: web content mining, web structure mining та web usage mining. Категорії Web mining представлені на рисунку 2.1.

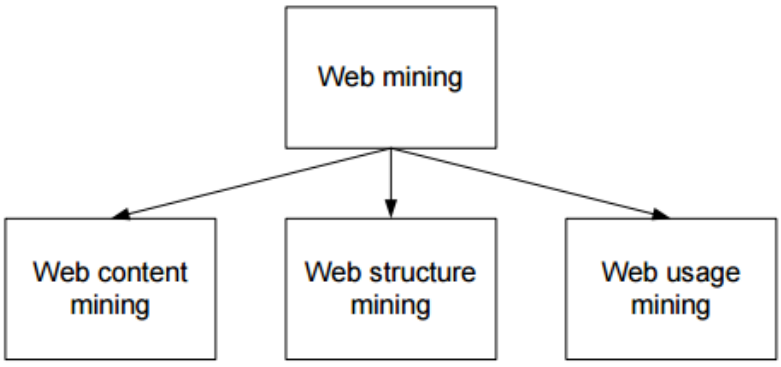

Рисунок 2.1 – Узагальнена схема категорій Web Mining

При цьому кожна їх описана категорія поділяється на більш детальні підкатегорії, кожна з яких містить свої особливості та сферу застосування. Більш детальну схему категорій Web Mining можна побачити на рисунку 2.2.

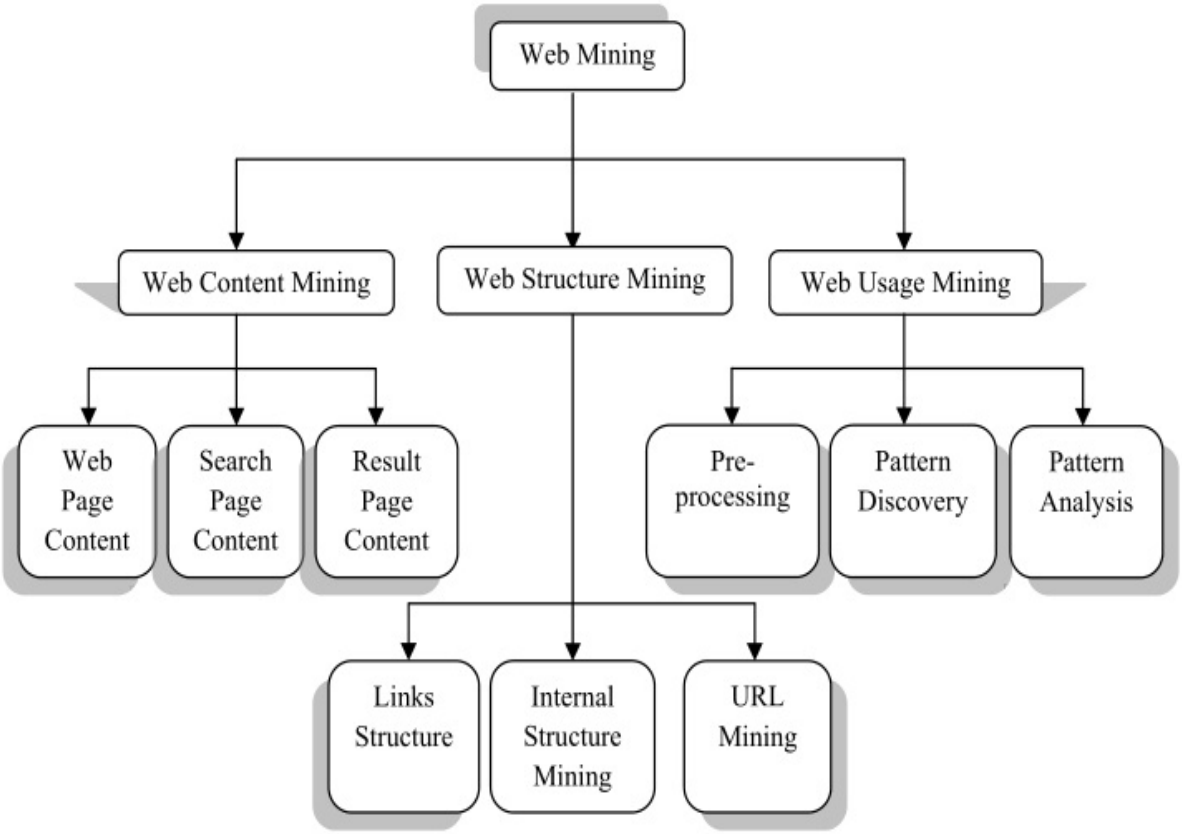

Рисунок 2.2 – Детальна схема категорій Web Mining

Web Content Mining – пошук знань у мережі Інтернет є непростим і трудомістким завданням. Саме цей напрямок Web Mining вирішує її. Воно засноване на поєднанні можливостей інформаційного пошуку, машинного навчання та Data Mining. Крім того, Web Content Mining передбачає автоматичний пошук та отримання якісної інформації з різноманітних джерел Інтернету, перевантажених «інформаційним шумом» [10].

За останні кілька років Web Content Mining став областю активних досліджень, і основні складності тут викликані гетерогенністю веб-даних та

їх низькою структуризацією, що ускладнює виділення цільової інформації. Крім того, у WCM необхідно вирішувати ряд специфічних завдань: вилучення структурованих даних із веб-сторінок з використанням методів машинного навчання та нейронних мереж; формування процедур уніфікації форматів подання даних та їх інтеграції із різних джерел; виділення оцінок продуктів та послуг у відгуках, що розміщуються на форумах, у блогах та чатах. Для відокремлення змістовної складової сторінок від службових та рекламних текстів потрібні відповідні процедури сегментації. Тут також йдеться про різні засоби кластеризації та анотування документів.

У Web Content Mining для кожного з трьох типів даних (структуровані, неструктуровані та квазіструктуровані) використовуються власні методи обробки, але незалежно від цього майже завжди виконується процедура перекладу даних із форми, призначеної для читання людиною, у форму, зручну для обробки комп'ютером. Така процедура називається data scraping, або «зрізання даних із поверхні». Перші технології screen scraping застосовувалися спочатку на мейнфреймах та пізніше на міні-комп'ютерах для надання діалогових функцій успадкованим програмам. Засобами screen scraping можна адаптувати такі програми для підтримки найпростіших «зелених» алфавітно-цифрових терміналів та таким чином налагодити режим інтерактивної взаємодії. Через багато років ця ж ідея відродилася як Web Scraping – якоюсь мірою їхня робота нагадує індексацію WWW, але її мета полягає не у складанні індексів, а у перетворенні неструктурованих даних, що існують у форматі HTML, у структуровані та збереженні їх у базах даних або електронних таблицях.

Очевидно, що найпростіше виконувати Web Content Mining для структурованих даних, тут достатньо застосувати службові процедури спочатку обходу сторінок, потім генерації та виконання пакувальника, а потім можна переходити до аналізу вмісту сторінки (Page Content Mining). Складніше з квазіструктурованими даними, що мають окремі ознаки структури, але не відповідають реляційній моделі. За цими ознаками дані

можна перетворити на структуровані, ось чому їх ще називають даними з самоописом. Проблема роботи з квазіструктурованими даними актуалізувалася з появою Інтернету, де таких даних багато. Як загальний приклад можна навести бібліографічні дані, де відома форма посилання, але в ній може бути невизначена кількість авторів або ще якісь порушення суворого формату. Те саме стосується різного роду персональних даних. Прикладом квазіструктурованих даних може бути граф перегляду сайту відвідувачем − для такого роду даних характерна змінна кількість полів, а самі поля при цьому можуть розташовуватися в довільному порядку. Особливу увагу такі дані привернули до себе з появою «людини, що читається» мови XML, що має гнучку структуру і текстовий формат обміну JSON (JavaScript Object Notation).

Web Content Mining виділяють два підходи: підхід, заснований на агентах, і підхід, заснований на базах даних.

Підхід, заснований на агентах (Agent Based Approach), включає такі системи [11]:

– інтелектуальні пошукові агенти (Intelligent Search Agents);
– фільтрація інформації/класифікація;
– персоніфіковані агенти мережі.

Приклади систем інтелектуальних агентів пошуку:

– Harvest (Brown та ін., 1994);
– FAQ-Finder (Hammond та ін., 1995);
– Information Manifold (Kirk та ін., 1995);
– OCCAM (Kwok and Weld, 1996);
– ParaSite (Spertus, 1997);
– ILA (Information Learning Agent) (Perkowitz and Etzioni, 1995);
– ShopBot (Doorenbos та ін., 1996).

Підхід, що базується на базах даних (Database Approach), включає системи:

– багаторівневі бази даних;

– системи Web-запитів (Web Query Systems).

Приклади систем web-запитів:

– W3QL (Konopnicki та Shmueli, 1995);

– WebLog (Lakshmanan та ін, 1996);

– Lorel (Quass та ін., 1995);

– UnQL (Buneman та ін., 1995 і 1996);

– TSIMMIS (Chawathe та ін., 1994).

Аналізується зміст документів: знаходяться схожі за змістом слова та їх кількість. Потім вирішується завдання кластеризації чи класифікації. Так документи групуються за смисловою близькістю. Цей напрямок може бути використаний для оптимізації пошуку індексованих документів.

Web Usage Mining – цей напрямок ґрунтується на вилученні даних із логів веб-серверів. Метою аналізу є виявлення переваг відвідувачів під час використання тих чи інших ресурсів мережі Інтернет.

Тут дуже важливо здійснити ретельну передобробку даних: видалити зайві записи лога, які цікаві для аналізу.

Web Usage Mining включає такі складові [11]:

– попередня обробка;

– операційна ідентифікація;

– інструменти виявлення шаблонів;

– інструменти аналізу шаблонів.

Кожен користувач мережі має свої індивідуальні уподобання, погляди, залежно від яких він відвідує ті чи інші ресурси. Виявивши, які сторінки та в якій послідовності відкривав користувач, можна зробити висновок про його переваги. Аналіз загальних тенденцій серед усіх відвідувачів показує, наскільки ефективно працює електронний портал, які його сторінки відвідуються найбільше, які менше.

Загальна схема роботи Web Usage Mining представлена на рисунку 2.3.

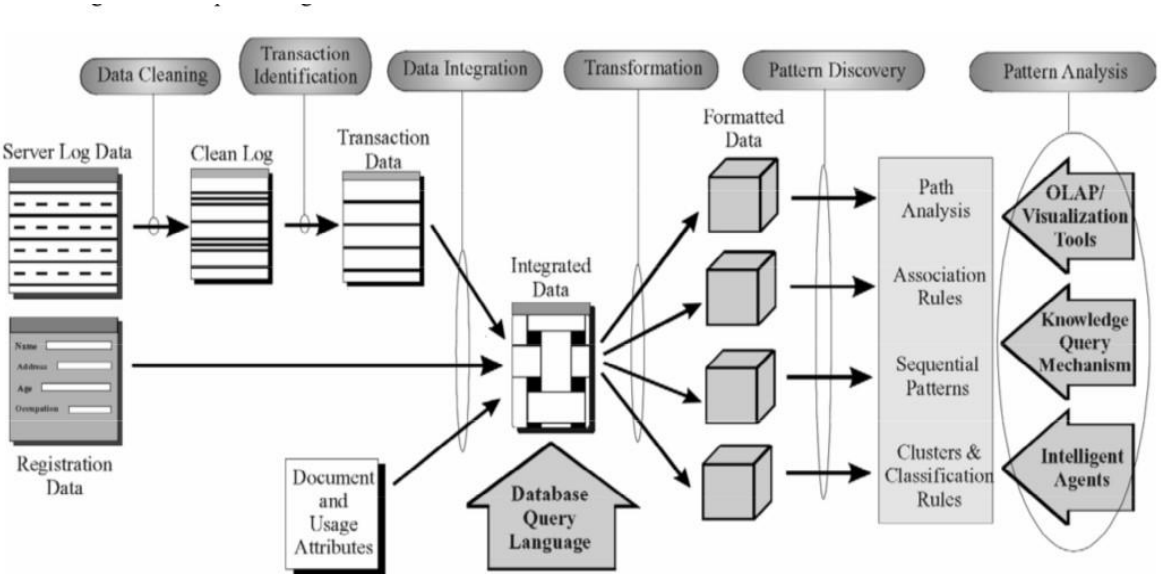

Рисунок 2.3 – Технологія Web Usage Mining

На основі цього аналізу можна оптимізувати сайт: знайти раніше не помічені проблеми у функціонуванні, дизайні та ін.

Цей напрямок Web Mining часом ще називають аналізом потоків кліків [12] – впорядковане безліч відвідувань сторінок, які переглянув користувач, потрапивши на веб-сайт.

Необхідні для аналізу дані знаходяться у логах серверів та cookie-файлах. При завантаженні веб-сторінки браузер також запитує всі об'єкти, що вставлені в неї, наприклад, графічні файли. У зв'язку з цим виникає проблема з тим, що сервер додає журнал запису про кожен такий запит. Звідси випливає необхідність попередньої обробки даних. Після того, як виділено окремі перегляди сторінок користувачем, їх об'єднують у сесію.

У таблиці 2.1 представлені найбільш популярні загальнодоступні програмні комплекси, засновані на Web Usage Mining.

Web Structure Mining. У задачі отримання Web-структур, насамперед, інтерес викликає структура гіперпосилань в межах Web-мережі. Потрібно подання документів та відносин між ними, що враховує гіперпосилання [12].

Таблиця 2.1 – Загальнодоступні програми, які використовують Web Usage Mining

Назва	Тип	Коментарі
STstat	Звіт та статистика	Є набір скриптів CGI, які генерують HTML звіти, на основі журналів доступу, які зберігає HTTP-сервер, і вони підходять практично до будь-якого програмного забезпечення HTTP-сервера, підтримують три формати журналу.
weblog_ parse	Log-файли обробки	Зчитує файл веб-журналу сервера, аналізує його, і записує лише зазначені користувачем поля, розділені знаками табуляції для полегшення обробки.
WebLog	Log-файли як аналіз	Засіб аналізу має повний доступ до журналу. Дозволяє відстежувати активність на сайті за місяць, тиждень, день та годину, відображаючи підсумкові значення показів, переданих байт та переглядів сторінок, відстежує найпопулярніші сторінки.
Analog	Аналізатор Log-файлів	Аналог програми для аналізу логів із веб-сервера. Вона повідомить, які сторінки є найбільш популярними, країни, мешканці яких відвідують сайти, а також які сайти вони переглядали, несправні посилання

Гіперпосилання моделюються з різним рівнем деталізації залежно від застосування моделі. У найпростіших моделях гіперпосилання можуть бути представлені як спрямований граф:

$$G = (D, L), \qquad (2.1)$$

де *D* – це набір вузлів, документів чи сторінок,

L – набір посилань.

Можна виділити три основні завдання, які можуть бути вирішені на підставі аналізу Web-структури:

– оцінка важливості структури Web (документа або вузла), вплив та вплив їх один на одного;

– пошук Web-документів з урахуванням гіперпосилань, що містяться в них;

– кластеризація структур їхнього можливого явного об'єднання.

Оцінка важливості веб-структур.

Л. Катц запропонував обчислення значимості Web-структур використовувати шляхи, засновані на вхідних посиланнях [13]. Відповідно до цієї ідеї кількість шляхів довжиною r від вузла i до j позначається rij. Загальна кількість шляхів різної довжини обчислюється за такою формулою:

$$Q_{ij} = \sum_{f=1}^{\infty} b^r P_{ij}^r, \qquad (2.2)$$

де величина b^r<0 повинна вибиратися таким чином, щоб забезпечити збіжність формули для кожної пари. Значимість вузла j обчислюється як сума кількостей шляхів від усіх вузлів:

$$s_j = \sum_i Q_{ij}.. \qquad (2.3)$$

У матричній формі обчислення значущості кожного вузла може бути записано у вигляді:

$$S = (I - bA)^{-1} - I, \qquad (2.4)$$

де I – одинична матриця;

A – матриця, що містить ваги зв'язків між вузлами.

Для визначення значущості впливу та впливу Web-структур широко використовуються метрики, які застосовуються для ранжування знайдених документів у пошукових системах. Так, широке застосування цієї категорії завдання Web Mining знайшла метрика, використовувана пошукової системою Google – PageRank [14].

PageRank- Статична величина, призначена для оцінки якості сторінок на підставі інформації про кількість посилань на неї.

Пошук Web-документів з урахуванням гіперпосилань.

Для пошуку Web-сторінок застосовується алгоритм Марxіорі (Marchiori's) HyperSearch. У ньому значення релевантності для сторінки p обчислюється методом, що включає релевантність сторінок, які можна досягти з p. При цьому залежність від релевантності сторінки, що досяжна, зменшується за рахунок коефіцієнта згасання, зменшеного експоненційно зі збільшенням відстані від сторінки p [15].

Іншим напрямом робіт, що використовують гіперпосилання для покращення результатів пошуку, є створення пошукового формалізму, який буде здатний обробляти запити, що включають предикати тексту та посилань. Ароцена, Менделзон та Михайла розробили структуру, яка підтримує Web-запити, які комбінують стандартні ключові слова з умовами структури навколишнього посилання.

Кластеризація Web-структур.

Для кластеризації стосовно Web-документів використовуються дві функції подібності з бібліометрики, що обчислюються для кожної пари документів p і q:

- бібліографічна пов'язаність (bibliographic coupling) – кількість документів, що цитуються обома документами p і q;
- взаємне цитування (co-citation) – кількість документів, які цитують обидва документи p і q.

2.1.4 Технологія Web Mining та інші засоби аналізу даних

Web Mining – це використання методів data mining для автоматичного знаходження та видобутку інформації з Веб-документів та сервісів. У цьому напрямі проводиться дуже багато досліджень за неймовірне зростання кількості інформації, що з'являється в Інтернеті і великим інтересом в електронній комерції.

Web Mining тісно пов'язаний із машинним навчанням та аналізом даних. Також Web mining часто асоціюється з пошуком інформації (Information Retrieval) та видобутком інформації (Information Extraction), хоча насправді це не те саме.

Web mining та пошук інформації (Information Retrieval). Пошук інформації (Information Retrieval – IR) – автоматичне знаходження всіх релевантних документів та одночасно мінімізація небажаних документів серед знайдених, а також ранжування знайдених документів за рівнем релевантності [14]. Первинна мета IR – індексування текстів та пошук важливих документів. Сучасні дослідження також відносять до IR моделювання, класифікації документів, інтерфейсів, візуалізації даних, фільтрування тощо. Якщо вважати, що Web mining – класифікація веб-документів з подальшою індексацією, тоді Web mining є частиною процесу IR. У кожному разі завдання індексування використовують методи data mining.

Web mining та вилучення інформації (Information Extraction). Видобуток інформації (Information Extraction – IE) полягає у обробці колекцій документів та поданні інформації, яку вони містять, у формі, зручній для роботи та аналізу. На відміну від IR, метою якого є знаходження релевантних документів, метою IE є знаходження релевантних даних у документах. IE полягає в аналізі структури та поданні документів, а IR – у знаходженні безлічі невпорядкованих даних [14].

Web Mining та застосування машинного навчання в Інтернет. Web

Mining – це не те ж саме, що навчання з Інтернету або методи машинного навчання, застосовані до Інтернету [15]. З одного боку, існують програми з машинним навчанням, які застосовуються в Інтернет, які не є зразками Web Mining. Прикладом цього є методологія машинного навчання, яка використовується для пошуку в Інтернет за заданою тематикою, яка наголошує на знаходженні наступних напрямів пошуку.

З іншого боку, Web Mining використовує як методи машинного навчання. Прикладом цього є специфічні алгоритми отримання інформації з авторитетних сайтів або вказівних сайтів та дослідження схеми Інтернет.

Однак, незважаючи на це, Web Mining та машинне навчання дуже близькі області досліджень та машинне навчання застосовується у Web Mining. Наприклад, недавні дослідження показують, що застосування машинного навчання може покращити процес класифікації текстів порівняно із звичайними IR технологіями. Таким чином, Web Mining перетинається із застосуванням машинного навчання до Інтернету.

2.1.5 Перспективи розвитку Web Mining

Як і для будь-якої технології, що розвивається, для Web Mining визначають шляхи його подальшого розвитку [12]. До них відносять такі:

– вибірка на певну тему. Сьогодні пошуку та стиснення інформації вже настільки розвинулися, що незабаром з'являться системи, які готували вибірку на певну тему за заданою областю даних (бази даних чи Інтернет). Технічно нескладно реалізувати це з використанням «розумної» пошукової машини та існуючих засобів семантичного стиснення інформації та знаходження смислових дублів;

– видобуток фактів (вікна фактів). Метод збирання інформації у тому, що з документів добуваються лише безперечні факти, часто дуже прості і нецікаві. Наприклад, із пропозиції «Прем'єр-міністр України «Y» підтвердив, що у 20xx році валовий прибуток зріс на 5 відсотків «можна

виділити лише один безперечний факт: «20xx році прем'єр-міністром України був «Y».

Виявилося, що зіставлення таких простих, атомарних фактів може дати нові знання. Наприклад, з газетних публікацій можна дізнатися про всю біографію «Y».

Можна припустити, що сучасні пошукові системи перейдуть від простої індексації слів у документах Інтернету до збору фактів. Технічно це не дуже важко, а атомарних фактів в Інтернеті – безліч. Факти, які збираються таким чином, мають дуже просту структуру, їх легко перетворити на знання та проводити з ними логічний висновок.

WebSQL. Сьогодні технології баз даних є надзвичайно потужним та гнучким засобом для створення запитів до добре структурованих даних. Були зроблені спроби застосувати ці технології баз в Інтернет. Але на цьому шляху є низка перешкод:

− під час виконання запиту неможливо пронумерувати всі документи в Інтернеті, оскільки вони утворюють безліч сторінок. А навіть якби це було можливо, це було б практично нереалізовано через величезну кількість інформації, яка міститься в Інтернеті;

− інформація, що міститься в Інтернеті, є слабоструктурованим. Мови запитів, розділені на добре структуровані дані, погано застосовні для створення запитів до Інтернету;

− документи HTML часто містять помилки, викликає труднощі при видобутку структури документа. Дані зберігаються у файлах різних типів: текстові, графічні, звукові, що ускладнює визначення того, чи задовольняє файл обмеження чи ні.

WebSQL – мова запитів до Інтернету, спроба застосувати технології реляційних баз даних до Інтернету. Моделюючи Веб як реляційну базу даних з двома вуртуальними відносинами Документи та Посилання, можна створювати SQL-подібні запити з обмеженнями на локальність, структуру, тип документа, дату модифікації та зміст. WebSQL – комбінування

структурних та змістовних запитів з використанням структури та топології Інтернет. Ця мова не є замінником індексних серверів, оскільки інтерфейс є занадто складним [13].

Мета використання WebSQL – полегшення розробки додатків, призначених для селективного індексування, автоматичного створення посилань тощо. WebSQL Веб розглядається як віртуальний граф, в якому документи є вершинами, з'єднані посиланнями. Тоді документ в Інтернет розглядається як кортеж віртуального відношення документа. Url (uniform resource locator) – уніфікований ідентифікатор інформаційного ресурсу – ідентифікує об'єкт, заголовок та текст містяться в HTML документі, дані про тип, довжину та останню модифікацію містяться на сервері.

Існують аналогічні підходи, які називають W3QS, W3QL та інші.

2.2 Аналіз технології mashup

Mashup – це змішування Web каналів із різних сервісів у одному додатку. Принцип створення mashup був запозичений з напряму поп-музики, де вперше з'явився напрямок музичного жанру mashup, який полягав у змішуванні вокальних та інструментальних звукових доріжок, зібраних з різних музичних творів. Mashup створюється з урахуванням Web-служб, які представляють розробникам інтерфейси прикладного програмування API. Творці API дотримуються простоти та оптимальності повторно використовуваних підтримуваних програм. Термін mashup застосовується лише до тих проектів, які використовують відкриті інтерфейси API для отримання даних послуг.

2.2.1 Поява mashup

Mashup з'явився в інтернеті завдяки появі Web 2.0. Web2.0 – методика проектування систем, які шляхом обліку мережевих взаємодій стають

краще, що більше користувачів ними користуються [17]. По суті, термін позначає проекти та сервіси, що активно розвиваються та покращуються самими користувачами: блоги, соціальні середовища тощо. Наприклад, Web сервіси: youtube, facebook – якби користувачі не додавали туди свої дані, ці Web сервіси були б порожні і нудні.

У своїй книжці один із винахідників Інтернету Tim Berners-Lee пише: «Ніхто не знає, що це означає… Якщо Web 2.0. – це ваші блоги та вікі, тоді це означає «користувачі для користувачів». Але це те саме, що сказати – Web 2.0 існує, щоб усі люди були разом». Тим самим він підкреслює, що Web 2.0 розвиватиметься і ставатиме цікавіше завдяки користувачам, які самі будуть додавати свої дані на Web сервіси.

2.2.2 Архітектура mashup

Архітектура будь-якого mashup складається з трьох основних частин, пов'язаних між собою фізично або логічно:

– провайдери API даних – це провайдери контенту, звідки береться інформація. Наприклад: у Chicagocrime використовувалася інформація з GoogleMaps та поліцейського департаменту Чикаго;

– Mashup сайт – це інтернет-додатки, на яких збирається та розміщується інформація від провайдерів контенту, інформація виходить за допомогою відкритих API;

– клієнтський браузер, який при певних налаштуваннях браузера з боку клієнта може генерувати інформацію як за мовними та регіональними налаштуваннями, так і за останніми пошуковими запитами. Завдяки чому, при запуску браузера, він сам вибирає необхідні налаштування web сервісів, позбавляючи користувача зайвої інформації, не пов'язаної не з його мовними, регіональними та пошуковими запитами [18]. Приклад реалізації налаштування браузера добре проглядається в API Google Maps. Під час запуску браузера користувача відображається інформація його мовою.

Одним словом, клієнтський браузер – це середовище, в якому додаток інтерпретується у графічному вигляді та відбувається взаємодія з користувачем.

Як було зазначено, mashup програми витягують інформацію з різних джерел даних, і навіть з різних сховищ. На рисунку 2.4 представлений приклад використання масхуп здобуванням різних сховищ даних як джерела.

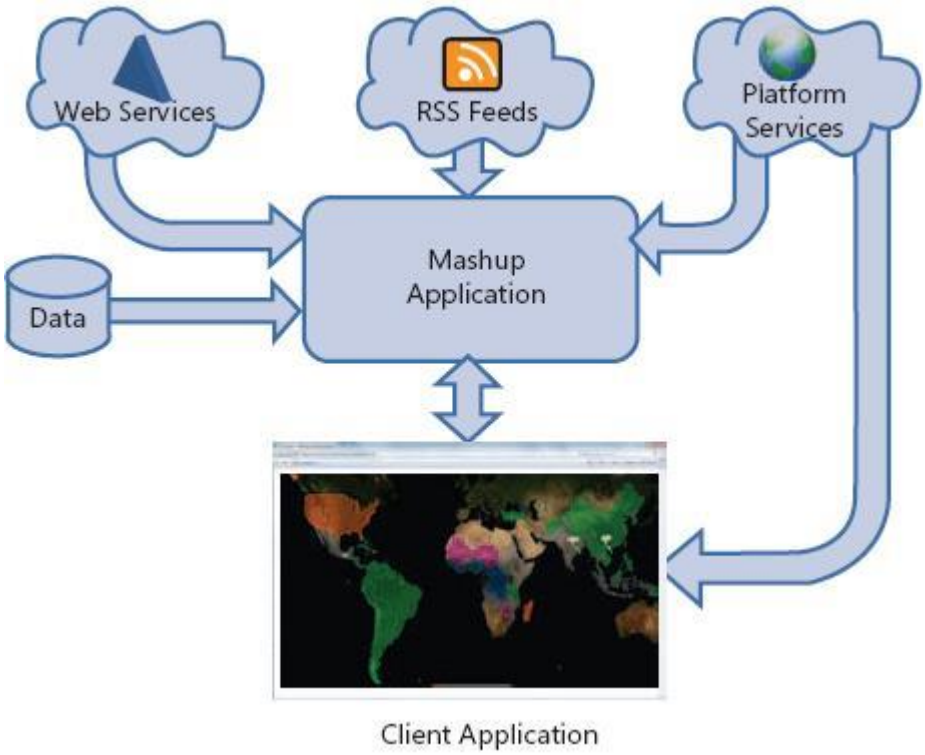

Рисунок 2.4 – Джерела даних для mashup програми

Для того, щоб mashup програми почали свою роботу, використовуються різні технології для їх реалізації.

Ajax – є модель web-додатків, що складаються з технологій:

– HTML5є співпрацею між W3C і web-додатків генеративної технології робочої групи WHATWG. Заснований на HTML, CSS, DOM, JavaScript;

– API моделі DOM – незалежна інтерфейсна модель – для

динамічного відображення та взаємодії HTML;

– XML мова розмітки, що розширюється – для асинхронного обміну даними;

– JavaScript – використання сценаріїв на стороні клієнтського браузера [17].

Враховуючи, що дані обмінюються між різними серверами та інтерпретуються на машині користувача, Ajax дозволяє це робити без перезавантаження браузера під час інтерактивної роботи користувача з web-додатком.

Soapта Rest – це платформи незалежних web-протоколів для зв'язку з віддаленими сервісами:

– Rest – архітектурний стиль програмного забезпечення для розподілу систем, таких як World Wide Web, який використовується для побудови веб-служб за допомогою HTTP та XML;

– Soap - протокол доступу до об'єктів. Використовується для обміну довільними повідомленнями у форматі XML.

Клієнти можуть використовувати SOAP та REST для взаємодії з віддаленими сервісами. Функції сервісу повністю передаються через опис повідомлень, з яких здійснюються запити і відповіді.

Atomта RSS – це сімейство форматів синдикації на основі XML. RSS дозволяє за допомогою онлайн сервісів та різних агрегаторів та каталогів імпортувати та експортувати інформацію, що дозволяє користувачеві збирати потрібну для нього інформацію у зручному для нього відео з різних сайтів. Формат Atom, який врахував недоліки RSS та реалізує той самий обмін інформації, заснований на XML.

2.2.3 Класифікація mashup додатків

Mashup програми можна розділити на два типи [17]:

– споживчі mashup програми – об'єднують дані з різних відкритих

джерел у браузері користувача. Користувач сам може брати участь у створенні нових даних, наприклад сайт arthalbum,де поєднуються картографічні дані з сервісуGoogle Map та фотографії з сервісу Flickr;

– бізнес mashup програми – створюються з використанням технології бізнес-бізнес, що дозволяє реалізувати спільні дії між підприємствами та розробниками. Добре підходять для швидкої розробки проектів, які потребують співпраці між розробниками та замовниками для визначення та реалізації бізнес-вимог. Збір та аналіз порівняльної інформації про ціни з торгових Інтернет-майданчиків дозволяє орієнтуватися у цінах, перебуваючи в одному веб-додатку.

Класифікація mashup застосувань:

– картографічні mashup програми. У наш час інформаційних технологій людство здійснює збір дуже великих обсягів даних про суб'єктів та дії; і ті, й інші забезпечуються приміткою про місце дії чи розміщення. Всі ці різні набори даних містять і дані про розміщення, які просто вимагають графічного відображення з використанням карт місцевості. Одним з найголовніших каталізаторів появи гібридних програм була поява на сайті Google програмного інтерфейсу програми (API) Google Maps. Цей сервіс став тим шлюзом, який дав можливість web-розробникам (а також любителям, початківцям та іншим) прив'язувати будь-які типи даних (буквально все що завгодно – від ядерних інцидентів до корів, що беруть участь у виставці великої рогатої худоби в Бостоні) до карт місцевості. Не можна не відзначити API від Microsoft (Virtual Earth), Yahoo (Yahoo Maps) і AOL (MapQuest), що з'явилися незабаром після цього;

– відео та фото програми. Поява сайтів, що надають місце для розміщення фотографій, та соціальних мереж, наприклад, Flickr, з API, які використовують фотографії, надані для загального використання, призвела до створення безлічі цікавих гібридних програм. Оскільки метадані цих контент-провайдерів пов'язані із зображеннями, розміщеними на їх серверах (хто зробив фотознімок, хто на ньому зображений, місце та час

зйомки тощо), розробники гібридних додатків можуть прив'язувати фотографії до іншої інформації, яку можна асоціювати з цими метаданими. Гібридна програма може, наприклад, проаналізувати текст пісні або вірша і створити мозаїку або колаж з підходящих за сюжетом фотографій або відобразити схему соціальної мережі, ґрунтуючись на метаданих фотографій (суб'єкт, тимчасова мітка та інші метадані). Такого плану mashup додаток може, наприклад, проаналізувати текст пісні або вірша і створити мозаїку або колаж з підходящих за сюжетом фотографій або відобразити схему соціальної мережі, ґрунтуючись на метаданих фотографій (суб'єкт, мітка та інші метадані);

– пошукові та торгові програми. Пошукові та торгові гібридні програми існували задовго до того, як був створений термін гібридний додаток. До настання епохи Web-API засновані на порівняльному методі інструменти – BizRate, PriceGrabber, MySimon та Google's Froogle – використовували поєднання технологій бізнес-бізнес (b2b) або захоплення дампи екрану для того, щоб збирати порівняльну інформацію про ціни. Щоб сприяти поширенню гібридних та інших цікавих web-додатків, клієнти торгових майданчиків в інтернеті, таких як eBay та Amazon, створили API для програмного доступу до свого контенту;

– новинки mashup. Джерела новин (наприклад, New York Times, BBC або Reuters) з 2002 року використовують технології синікації RSS та Atom (про які йдеться в наступному розділі) для поширення каналів новин, згрупованих з різних тем. Mashup-канали для синдикації новин можуть збирати канали новин, обрані користувачами, та представляти їх через інтернет, створюючи персональну газету, яка враховує інтереси конкретного читача.

Приклади найбільш популярних mashup додатків:

– мапи. Chicago Crime. Поліцейський департамент Чикаго mashup, який інтегрує базу даних департаменту про злочини з Google Maps для того, щоб зупинити злочинність в областях та попередити мешканців

про те, де часто скоюються злочини;

– відео та фото. Flickr – це сховище даних зображень, яке дозволяє користувачам організовувати свою колекцію зображень та обмінюватися ними. Використовуючи API Flickr, дані можна використовувати для створення мешапів;

– пошук та шопінг. Travature – Це портал про подорожі, що інтегрує двигун метапошуку авіаперельотів, гіди про подорожі та огляди готелів. Портал дозволяє користувачеві обмінюватися фотографіями та обговорювати свій досвід з іншими користувачами;

– новини. Digg. Мешап різних веб-сайтів новин, практично повністю контрольований користувачами ресурсу.

3 АНАЛІЗ ТЕХНОЛОГІЙ І ЗАСОБІВ РЕАЛІЗАЦІЇ ДОДАТКІВ ДЛЯ WEB MINING

3.1 Застосування Web Content Mining для розробки програми

Як було зазначено у цій роботі, виділяють три типи Web Mining:

– Web Usage Mining;
– Web Content Mining;
– Web Structure Mining.

Для реалізації web-додатку, метою якого є вилучення та аналіз вмісту новин у сфері вищої освіти України, найбільш підходящим типом Web Mining буде Web Content Mining, оскільки саме цей напрямок дозволяє отримувати лише необхідний вміст сторінки, такий як текст, зображення та різного роду метадані.

Сучасний обсяг даних в інтернеті пропонують великі можливості для аналізу інформації, знаходження раніше невідомих даних. Те ж саме можна сказати про більш вузьку предметну область. Наприклад, новинній сфері. Різноманітність інформації, зокрема у сфері вищої освіти України, дозволяє застосовувати різні методики аналізу даних, аналізу контенту, тексту для виявлення прихованих даних, які можуть допомогти класифікувати отриману нами інформацію, фільтрувати її відповідно до переваг, а отримати нові знання, які не видно неозброєним оком.

Так як вміст будь-якої web-сторінки містить велику кількість зайвої інформації, яка не несе в собі жодної цінності, то для отримання знань з сторінок новин, необхідно спочатку позбавиться від інформації, що забруднює. Для цього вміст кожної аналізованої сторінки спочатку парситься, що дозволяє в результаті працювати тільки з корисною інформацією, що стоїть і робить результат роботи більш чистим і правильним.

Крім цього, простого відображення новин як є, хай навіть із різних

джерел, вже може бути недостатньо. Інша справа, аналіз вмісту відображуваних новин. Серед усіх методів аналізу вмісту web-контенту широку популярність набули такі напрямки аналізу тексту, як визначення тональності тексту та виділення ключових слів.

Так, наприклад, виділення ключових слів у тексті надасть користувачеві можливість навігації безпосередньо за темами, які можу бути йому цікавими та відкидати ті, які не становлять йому ніякої користі.

Так само і визначення емоційного забарвлення тексту допоможе користувачеві бути більш підготовленим до змісту статті новин.

3.2 Аналіз тональності контенту новин

Основною метою аналізу тональності є знаходження думок у тексті та виявлення їх властивостей. Які саме властивості будуть досліджуватися, залежить вже від поставленого завдання. Наприклад, метою аналізу може бути автор, тобто особа, якій належить думка.

Аналіз тональності тексту дозволяє витягувати з тексту емоційно забарвлену лексику та емоційне ставлення авторів стосовно об'єктів, про які йдеться у тексті. Більшість сучасних систем використовують бінарну оцінку – «позитивний сентимент» або «негативний сентимент», проте деякі системи дозволяють виділяти силу тональності [19].

У сучасному світі на наш вибір у будь-яких ситуаціях найчастіше впливає думка інших людей – ми читаємо відгуки про товар, перш ніж замовити його в інтернет-магазині, дізнаємося про думку інших людей, перш ніж проголосувати на виборах за того чи іншого кандидата, довго і довго. ретельно обираємо собі ВНЗ, місце роботи та ресторан, який ми збираємось відвідати. Ця інформація становить значний інтерес для маркетологів, соціологів та багатьох інших фахівців.

Загальне визначення свідчить, що аналіз тональності текстів – це клас методів контент-аналізу, призначений для автоматичного виявлення у тексті

емоційно забарвленої лексики, і навіть думок (емоційних оцінок) автора щодо об'єктів, про які йдеться у тексті. З визначення можна зробити кілька висновків про те, де теоретично концепція аналізу тональності тексту могла знайти застосування і прояснити деякі її деталі.

По-перше, аналіз тональності текстів здатний допомогти розібратися в законах, за якими живе природна мова і навчити комп'ютер сприймати її на рівні, наближеному до людського. До недавнього часу машина розуміла тексти на абстрактному рівні - в основному, через лексеми (слова), які для неї мали форму (набір букв) і зміст (значення). Дана концепція пропонує запровадити ще одну функцію – так звану лексичну тональність тексту (у найпростішому випадку вона визначатиметься як сума лексичних тональностей кожної окремої лексеми).

По-друге, аналіз тональності здатний значно підвищити якість машинного перекладу. Відомо, що зразком машинного перекладу є результат перекладу тексту людиною – професійним перекладачем. За 50 з лишком років розробок у цій галузі дослідники переконалися у тому, що навчити машину «думати, як перекладач» можна лише зваживши на всі ті міркування, якими користується професіонал, перекладаючи той чи інший текст. Звичайно, при перекладі не обійтися без первинного аналізу тексту та окремих слів – у тому числі аналізу тональності як такої.

По-третє, метою аналізу тональності тексту може бути думка автора чи сам автор. Це найбільш цікава сфера застосування, оскільки тут бачиться не тільки спосіб делегування машині деяких повноважень вченого (наприклад, філолога, який досліджує твір того чи іншого автора), а й знову спроба наблизити спосіб мислення комп'ютера до людського. З цієї точки зору аналіз тональності, можливо, є одним із найважливіших та перспективних кроків до розвитку штучного інтелекту [20].

Думки поділяються на два типи:

– безпосередня думка;
– порівняння.

Безпосередня думка містить висловлювання автора про один об'єкт. Формальне визначення безпосередньої думки виглядає так: «безпосередньою думкою називається кортеж з п'яти елементів (entity, feature, orientation or polarity, holder, time), де entity – об'єкт, про аспект чи властивості (feature) якого автор (holder) висловив свою емоційну оцінку (orientation or polarity) у час (time)».

Тональність тексту визначається трьома факторами:

– суб'єкт тональності (автор, тобто комусь належить ця думка);

– об'єкт тональності та її властивості (сутність, щодо якої висловлюється автор);

– власне, тональна оцінка (емоційна позиція автора щодо згаданої теми).

Приклади тональних оцінок:

– позитивна;

– негативна;

– нейтральна.

Під «нейтральним» мається на увазі, що текст не містить емоційного забарвлення. Також можуть існувати інші тональні оцінки.

Якщо стоїть завдання класифікації більш ніж на два класи, то тут можливі такі варіанти для навчання класифікатора:

– плоска класифікація − навчаємо лише один класифікатор для всіх класів. Подано на рисунку 3.1;

– ієрархічна класифікація – ділимо класи на групи та навчаємо кілька класифікаторів для визначення груп. Подано на рисунку 3.2;

– регресія − навчаємо класифікатор для отримання чисельного значення тональності, наприклад, від 1 до 10, де більше значення означає більш позитивну тональність.

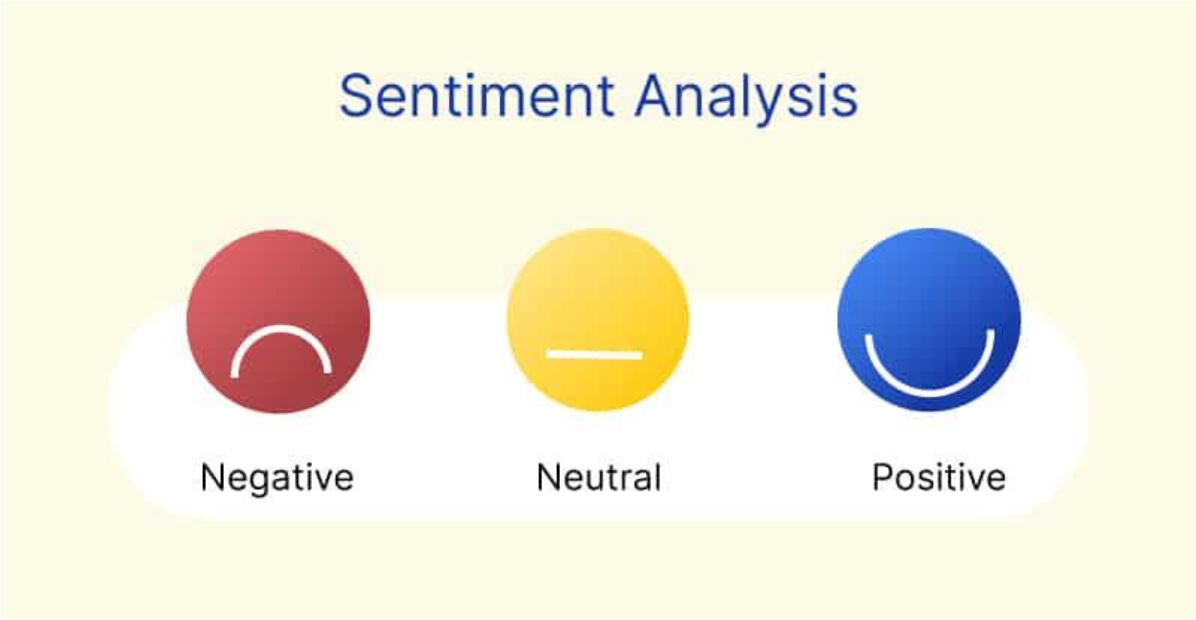

Рисунок 3.1 – Приклад плоскої класифікації

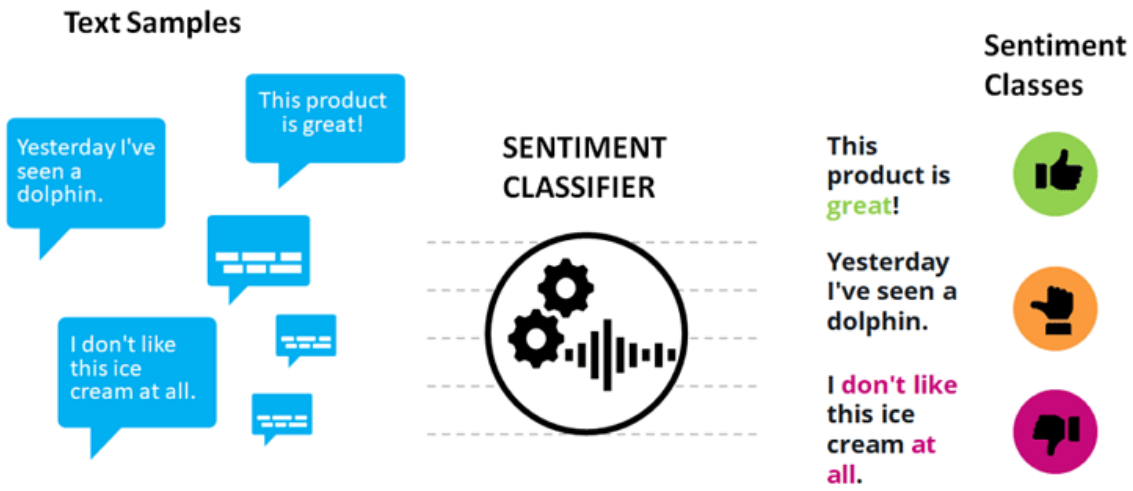

Рисунок 3.2 – Приклад ієрархічної класифікації

Комп'ютери можуть виконувати автоматичний аналіз цифрових текстів за допомогою елементів машинного навчання, такі як прихований семантичний аналіз, метод опорних векторів, «мішок слів» та семантична спрямованість у цій галузі. Більш складні методи намагаються визначити володаря настроїв (тобто людини) та мету (тобто сутність, щодо якої виражаються почуття). Щоб здобути думку у контексті, використовують

граматичні відносини між словами.

Відносини граматичної зв'язаності виходять шляхом глибокого структурного аналізу тексту. Аналіз тональності може бути поділений на дві окремі категорії [20]:

– ручний (або аналіз тональності експертами);

– автоматизований аналіз тональності.

Найбільш помітні відмінності між ними лежать у ефективності системи та точності аналізу. Програмні інструменти на основі відкритого вихідного коду використовують алгоритми машинного навчання, інструменти статистики та обробки природної мови, щоб автоматизувати аналіз тональності великих масивів текстів, включаючи веб-сторінки, онлайн-новини, тексти дискусійних груп у мережі Інтернет, онлайн-огляди, веб-блоги і соціальні медіа [22].

Аналіз тональності зазвичай визначають як із завдань комп'ютерної лінгвістики, тобто. мається на увазі, що ми можемо знайти та класифікувати тональність, використовуючи інструменти обробки природної мови (такі як тегери, парсери та ін.).

3.3 Виділення ключових слів у контенті новин

Основний зміст будь-якого документа (тексту) можна висловити з допомогою певних слів, узятих безпосередньо з цього тексту. Як правило, до кожного розгорнутого тексту можна скласти набір ключових слів різного обсягу (від 5 до 15 слів). Але взагалі кількість ключових слів може змінюватись у широких межах. В окремих випадках компресія може призвести до виділення одного основного ключового слова, яке є найчастішим знаменним словом.

Ключові слова мають ряд істотних ознак [20]:

– високий рівень повторюваності даних слів у тексті, частотність їх вживання;

– здатність знака конденсувати, згортати інформацію, виражену цілим текстом, об'єднувати його основний зміст; ця ознака особливо яскраво проявляється у ключових слів у позиції назви.

Однак вибір ключових слів є дуже непростою операцією та потребує серйозного підходу. Слід вибирати ті ключові слова, які найбільше точно відображають специфіку саме вашої теми. При цьому необхідно уникати випадкових і загальних фраз і не рекомендується повторювати кілька разів ті самі ключові слова. Тобто процес виділення ключових слів є ще й аналітичним.

Оскільки ключові фрази відображають основну ідею документа, від вилучення правильних ключових фраз залежить ефективність додатків з обробки природних мов: інформаційний пошук, питання-відповіді системи, автоматичне реферування. У пошукових системах, таких як Google та Yahoo, ключові фрази відіграють роль доповнення повнотекстового індексу та допомагають користувачам у складанні точних пошукових запитів. У питаннях-відповідях системах виділення ключових фраз з питання, дозволяє дати користувачеві більш точну відповідь. В автоматичному реферуванні фрази можуть використовуватися як семантичні метадані [21]. Таким чином, якість вилучення ключових фраз безпосередньо впливає на якість додатків з обробки природної мови.

Завдання вилучення ключових фраз відноситься до завдань контрольованого навчання або навчання з учителем. Навчання з учителем – це один із способів машинного навчання в процесі якого система навчається на тренувальному корпусі статей за допомогою прикладів. Навчання є попереднім етапом перед отриманням прогнозу. Між вхідною та вихідною інформацією (ключовими фразами) може існувати певна залежність, але вона невідома. У разі відома лише сукупність прецедентів, звана навчальної вибіркою. На основі цих даних необхідно відновити залежність, побудувати модель, здатну дати досить точну відповідь для будь-якого вхідного об'єкта.

Завдання автоматичного вилучення ключових фраз із тексту

складається з кількох етапів:

– попередня обробка тексту;
– відбір кандидатів ключових фраз;
– розрахунок ознак кожного кандидата;
– відбір ключових фраз із числа кандидатів.

У процесі попереднього відпрацювання з тексту провадиться видалення неінформативних частин (малюнки, таблиці). Кандидати ключових фраз відбираються у вигляді N-грам, не розділених розділовими знаками (крім дефісу та лапок) і стоп-словами. Де N-грама – це термін з комп'ютерної лінгвістики, що означає послідовність N елементів тексту, наприклад, слово або їх послідовність. Стоп-слова – слова, що не несуть жодного смислового навантаження, прийменники, спілки, вигуки, які часто зустрічаються в будь-якому документі. Для кожного з кандидатів ключових фраз розраховуються ознаки, які дозволяють судити про важливість кандидата у цьому документі. Набір кандидатів ключових фраз ранжується за значеннями ознак, наприклад, відповідно до їх частотності та ваги інформативності, розрахованих за однією з методик. Після ранжирування проводиться відбір перших найкращих ключових фраз із цього списку або відбираються кандидати, що перевищують встановлений мінімальний поріг значення ознаки. Найважливішим етапом завдання вилучення ключових фраз є розрахунок їх ваг інформативності, який дозволяє оцінити їх значущість по відношенню один до одного в документі.

Визначення ключових фраз у сфері новин вищої освіти України можна використовувати для спрощення навігації користувача за контентом новин. Крім цього, ключові слова та фрази можуть стати в основу тегів за контентом, що вивчається. Також додавання ключових слів до мета даних поточного контенту зможе підвищити релевантність даної новини в пошукових системах. Ще однією метою для якої можуть бути ключові слова виступає певна категоризація контенту за ключовими словами.

3.4 Опис засобів реалізації програми

Проаналізувавши завдання, які були поставлені, а також вивчивши дані про предметну область, було обрано саме технологію mashup для web-додатку, оскільки саме дана технологія дозволяє об'єднувати в одному місці інформацію з різних джерел даних.

Будь-який web-додаток складається з клієнтської та серверної частин. При цьому реалізації кожного з описаних елементів використовують свої технології.

Так, для реалізації web-сервісу на стороні сервера використовуються різноманітні технології та будь-які мови програмування, здатні здійснювати виведення у стандартну консоль, такі як ASP.NET, Java, PHP та ін. На стороні клієнта для реалізації GUI зазвичай застосовують HTML і CSS. Для формування та обробки запитів, створення інтерактивного та незалежного від браузера інтерфейсу найчастіше використовують ActiveX, Adobe Flash, Adobe Flex, JavaScript, Silverlight.

Для реалізації web сервісу в рамках цієї роботи було обрано такі технології:

– для реалізації серверної частини програми - Java EE 7;

– для реалізації клієнтського інтерфейсу – HTML5 та CSS3;

– для формування запитів – JavaScript, Ajax.

Як мову програмування на серверній стороні було обрано технологію Java EE з наступних причин:

– J2EE є єдиним середовищем розробки, що підтримує багатоплатформність;

– в основі J2EE виступає лише одна мова програмування - Java;

– наявність технологій розробки зовнішнього інтерфейсу – JSP, JSTL;

– наявність технологій для управління життєвим циклом програми – Maven;

– наявність фреймворків для керування залежностями – Spring.

Крім реалізації серверної та клієнтських частин, додаток містить бізнес логіку, яка ґрунтується на обробці даних, що зберігаються на описаних вище освітніх ресурсах. Так, основним інструментом парсингу даних є бібліотека boilerpipe. Бібліотека boilerpipe забезпечує алгоритми виявлення та видалення надлишків у даних навколо основного текстового змісту веб-сторінки. Бібліотека вже містить конкретні стратегії для загальних завдань, а також може бути легко розширена для індивідуальних налаштувань завдання. Вилучення вмісту документа дуже швидко (у мілісекундах) і, як правило, досить точно.

Так як вміст будь-якої web-сторінки містить велику кількість зайвої інформації, яка не несе в собі жодної цінності, то для отримання знань з сторінок новин, необхідно спочатку позбавиться від інформації, що загрожує. Для цього вміст кожної аналізованої сторінки спочатку псується, що дозволяє в результаті працювати тільки з корисною інформацією, що стоїть і робить результат роботи більш чистими і правильним.

Так, для отримання посилань на сторінки з новинами зі сторінки http://ua.osvita.ua/news/, яка має структуру, представлену на рисунку 4.1 задано селектор, представлений у прикладі.

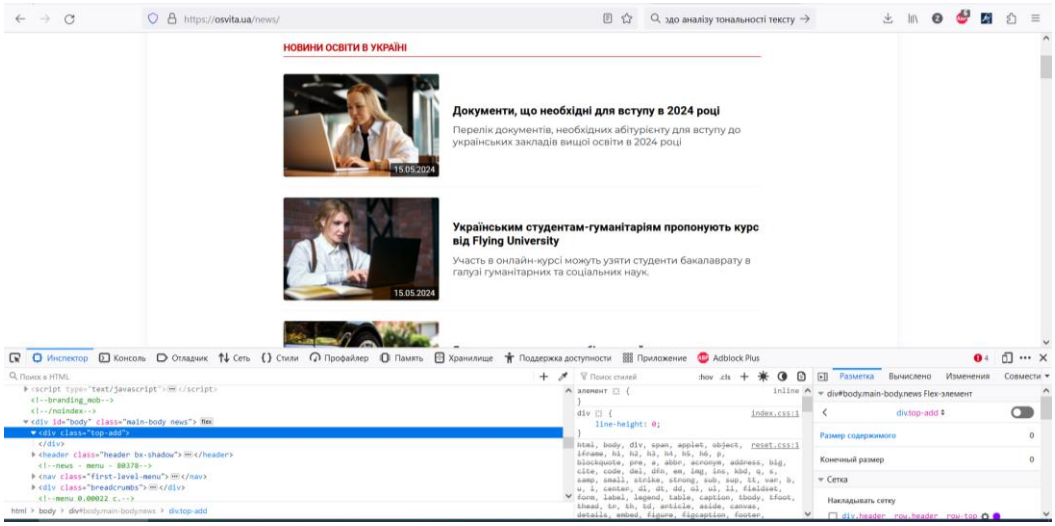

Рисунок 3.4 – Приклад DOM структури елемента

3.4.1 Інструментарій для отримання посилань на сторінки з новинами

У цьому додатку обробляється контент із конкретних сторінок. Але оскільки новини – це інформація, яка постійно оновлюється та додається, ми не можемо одразу задати конкретний список сторінок з новинами. Для цього необхідно виконувати конкретні посилання на сторінки з контентом з основної, постійної сторінки сайту.

Для цих цілей чудово підходить бібліотека JSOUP. Розглянемо приклади роботи з бібліотекою soup. Java-бібліотека súp призначена для аналізу HTML-сторінок (парсингу), дозволяючи отримати необхідні дані, використовуючи DOM, CSS та методи в стилі jQuery. Бібліотека підтримує специфікації HTML5 і дозволяє ширяти сторінки, як це роблять сучасні браузери. Бібліотеку можна підсунути для аналізу URL, файлу або рядка.

Soup – це Open-source Java бібліотека для роботи з реальним HTML. Вона забезпечує дуже зручний API для вилучення та маніпулювання даними, використовуючи найкращі DOM, CSS, та JQuery-подібні методи. Soup реалізує WHATWG HTML5 специфікацію, і розбирає HTML в ту саму модель DOM, як це роблять сучасні браузер на кшталт Chrome і Firefox:

– Soup може очистити та розібрати HTML з URL, файлу чи рядка;

– Soup може знайти та витягти дані, використовуючи обхід DOM або CSS селектори;

– Soup дозволяє маніпулювати HTML елементами, атрибутами та текстом;

– Soup забезпечує очищення наданої користувачем інформації white-list, для запобігання XSS атак;

– також Soup видає «акуратний» HTML.

Soup призначений для роботи з різними видами HTML, що існують у реальному світі, включаючи належним чином підтверджений HTML з неповним непідтвердженим набором тегів.

3.4.2 Інструментарій для отримання контенту зі сторінки

Для отримання посилань на конкретні сторінки з новинами використовується наступний сервіс: Parser API Readability.

Спочатку сервіс передбачався як простий інструмент для читання, який дозволяє перетворювати будь-яку web сторінку на зручний для читання вигляд. Він був запущений в (як експеримент) в 2010 році, в Нью-Йорку. Проект Readability змінює формат веб-сторінки на спрощений варіант HTML Arc90. Для цього використовується jsdom. Також цей продукт підтримує велику кількість кодувань таких як GB2312, і ще readability підтримує відносні URL-адреси, завдяки чому зображення продовжують відображатися.

Для використання даного інструменту не потрібно проходити авторизацію через OAuth, достатньо просто отримати application id та використовувати його для некомерційних цілей. Далі отриманий appId використовується як параметр під час передачі запитів на парсинг вмісту сторінки.

Розглянемо приклад використання Readability Parser API для отримання необхідного контенту зі сторінки http://ua.osvita.ua/vnz/46980/.

Робимо наступний запит:

```
GET
https://www.readability.com/developers/api/parser?url=http://
osvita.ua/vnz/46980/&token=<TOKEN>
```

В результаті такого запиту маємо наступну відповідь із сервера у форматі JSON (лістинг 3.1).

Лістинг 3.1 – Відповідь із сервера на запит

```
{ "domain": "ru.osvita.ua",
 "next_page_id": null,
 "url": "http://ua.osvita.ua/vnz/46980/",
 "short_url": "http://rdd.me/w5vvv5ef",
 "author": null,
 "excerpt": "Введіть e-mail адресу:",
```

Продовження лістингу 3.1


```
"direction": "ltr",
 "word_count": 165,
 "total_pages": 0,
 "content": "<div><div class="article"><table
class="topimg"><tr><td><img
src="http://ua.osvita.ua/doc/images/news /469/46980/28-
001_i.jpg" alt="ЗВО завищує вартість дипломів про освіту"
title="ЗВО завищує вартість дипломів про освіту"></td>
<td class="vbottom"></td></tr></table><p id="divtomail">Введіть
e-mail адресу: <br></p>Міністерство освіти та науки підкреслює,
що гранична вартість виготовлення документа про вищу освіту
державного зразка не має перевищувати 34-х гривень. Про це
йдеться в листі</a> освітнього відомства до керівників
навчальних закладів, викликаного численними зверненнями про
завищення вартості. виготовлення дипломів.</p><p>Зокрема, у
Міносвіти нагадують, що згідно <a
href="http://ua.osvita.ua/legislation/law/2235/"
target="_blank">закону "Про вищій освіті"</a> Заклади вищої
освіти мають самостійно виготовляти документи про вищу освіту.
На виконання зазначених норм Кабінетом Міністрів у березні було
прийнято постанову <a
href="http://ua.osvita.ua/legislation/Vishya_osvita/46747/"
target="_blank">№ 193</a> "Про документи про вищу освіті
(наукові ступеня) державного зразка", яким затверджено перелік
інформації, що має міститися у документах про вищу освіту, а
також встановлено, що гранична вартість виготовлення документа
про вищу освіту не повинна перевищувати двох неоподатковуваних
мінімумів доходів громадян.</p><p >"З метою неухильного
дотримання законодавства про вищу освіту пропонуємо в строк до
20 травня 2024 року забезпечити оприлюднення на офіційному веб-
сайті навчального закладу інформації про вартість документа про
вищу освіту державного зразка, який видається навчальним
закладом", - йдеться у листі Міністерства освіти та науки.</p>
<div id="soc2" class="social-likes"><p class="facebook"
title="Поділитись посиланням на Фейсбуці">Facebook</p><p
class="twitter" title="Поділитись посиланням у Твіттері
">Twitter</p><p class="vkontakte" title="Поділитись посиланням
у Вконтакті">Вконтакті</p><p class="plusone" title="Поділитися
посиланням у Google+">Google+</p ></div><p class="info"><a
href="http://osvita.ua"
title="osvita.ua">Освіта.ua</a><br>19.05. 2024</p></div><div
class="avatar"><img
src="http://ua.osvita.ua/doc/images/users/avatar_t.gif"
id="avatar" alt=" Аватар"></div></div>",
 "date_published": null,
 "dek": null,
 "lead_image_url":
"http://osvita.ua/doc/images/news/469/46980/28-001_i.jpg",
 "title": "ЗВО завищують вартість дипломів про освіту",
 "rendered_pages": 1
}
```

3.4.3 Інструментарій для обробки контенту новин

Для інтелектуального аналізу вмісту статей новин вибраний сервіс Alchemy API.

Модуль Alchemy реалізує API та дозволяє працювати з кількома інтерфейсами, які призначені для аналізу змісту сайту через сервіс Alchemy. Alchemy API дозволяє веб-розробникам обробляти зміст сайту і надавати йому певні мета-дані.

Сервіс аналізує зміст документа і витягує з нього семантичні дані: інформацію про людей, місця, компанії, тему статті, мову якою вона написана і так далі. Використання цього сервісу дозволяє покращити семантику сайту. Цей модуль зосереджує увагу на таких речах:

– ключові слова: у змісті виділяються ключові слова, ця техніка схожа на роботу Yahoo! Term Extraction, але дає якісніший результат, що корисніше для пошукової оптимізації;

– об'єкти: у змісті виділяються терміни, пов'язані з людьми, компаніями, організаціями, містами, географією тощо, що схоже на роботу Open Calais;

– загальна концепція: це схоже на пошук ключових слів, але тут на чільне місце ставиться розуміння зв'язку між окремими словами і загальними поняттями.

Для роботи із сервісом Alchemy потрібно зареєструватися на ньому, щоб отримати API-ключ. Крім цього, потрібно самостійно завантажити Alchemy SDK і розпаковувати в кореневу папку модуля (або в папку libraries, якщо використовується Libraries API).

AlchemyAPI забезпечує:

– REST API endpoints;

– SDK на всіх основних мовах програмування;

– трансформація мета-даних у різних форматах;

– 24/7 телефонна підтримка.

Sentiment Analysis API надає механізми виявлення позитивного чи негативного настрою у будь-якому документі. Алгоритм аналізу тональності Alchemy API виділяє слова, які мають позитивний чи негативний відтінок. Він також розуміє заперечення (наприклад, «цей автомобіль добре» проти «цей автомобіль не є хорошим») і модифікатори (тобто «цей автомобіль добре» проти «цей автомобіль дійсно хороший»). API аналізує тональність на документах великих та малих, у тому числі новинних статей, повідомлень у блозі, огляди продукції, зауваження та твіти. У лістингу 3.2 представлений запит для аналізу тональності, що передається в параметрах вмісту. Вміст взято зі сторінки http://ua.osvita.ua/abroad/higher_school/united-states/46992/.

Лістинг 3.2 – Приклад запиту для аналізу тональності

```
POST
http://access.alchemyapi.com/calls/text/TextGetTextSentiment?a
pikey=<APIKEY>&outputMod=json&text=<CONTENT>
```

В результаті отримуємо відповідь, наведену в лістингу 3.3.

Лістинг 3.3 – Результат аналізу тональності тексту

```
{
 "status": "OK",
 "TotalTransactions": "1",
 "language": "українська",
 "docSentiment": {
 "mixed": "1",
 "score": "0.397149",
 "type": "positive"
 }
}
```

Keyword Extraction API здатний знаходити ключові слова в тексті та ранжувати їх. Дане API працює як з URL чи HTML документами, і з текстовими файлами. AlchemyAPI автоматично визначає мову вмісту, а потім виконує відповідний аналіз. Алгоритм виділення ключових слів

використовує складні статистичні алгоритми та технології обробки природної мови для аналізу вмісту та виявлення відповідних ключових слів.

У лістингу 3.4 подано запит на виявлення ключових фраз у тексті, поданому у статті http://ua.osvita.ua/abroad/higher_school/ united-states/46992/.

Лістинг 3.4 – Приклад запиту на виділення ключових слів

```
POST
http://access.alchemyapi.com/calls/text/TextGetRankedKeywords?
apikey=<APIKEY>&text=<CONTENT>&outputMode=json&maxRetrieve=10
```

В результаті такого запиту маємо наступну відповідь, наведену в лістингу 3.5.

Лістинг 3.5 – Приклад виділення ключових слів

```
{
"status": "OK",
"language": "українська",
"keywords": [
{
"relevance": "0.94916",
"text": "безкоштовне навчання"
},
{
"relevance": "0.848755",
"text": "Ліги Плюща"
},
{
"relevance": "0.784212",
"text": "фінансова допомога"
},
{
"relevance": "0.775266",
"text": "Prep Schools"
}
}
```

3.5 Опис використовуваних інформаційних ресурсів

На даний момент серед українського web-простору є досить небагато

порталів, спрямованих на освіту. А ті, що існують, часто не завжди справляються зі своїм завданням – донести до користувача максимально повну підтверджену інформацію, що охоплює всі сфери освіти – починаючи зі шкільної початкової освіти та закінчуючи вищими навчальними закладами.

Таким чином, безперечним виходом з цієї ситуації буде комбінування різних джерел у рамках одного додатка, яке дозволить з легкістю орієнтуватися серед основних освітніх новин, без необхідності багаторазового відвідування різних інформаційних освітніх ресурсів, при цьому витрачаючи мінімальний час.

Після детального вивчення основних web-орієнтованих ресурсів, спрямованих на сферу вищої освіти України, було виділено такі:

– сайт міністерства освіти та науки України: http://www.mon.gov.ua – є першоджерелом основних новин та постанов щодо освіти в Україні. Містить інформацію не тільки про вищу освіту, але також про всі інші освітні галузі. Дозволяє переглядати стрічку новин за датою додавання, а також по регіонах;

– освітній портал: http://www.osvita.org.ua – містить ексклюзивні новини української та зарубіжної освіти, законодавство, інформацію про ЗНО, навчання за кордоном, бізнес-освіту, вивчення іноземних мов, а також свіжу інформацію від найкращих навчальних закладів України;

– інформаційний портал освіти: http://osvita.ua – популярний тематичний ресурс в українському Інтернеті, присвячений освіті в Україні та за її межами, є актуальним джерелом інформації у цій сфері. Тематика матеріалів, які розміщуються на сайті, відповідає інтересам широкої цільової аудиторії: педагогів, абітурієнтів, студентів, учнів, батьків;

– один з найпопулярніших порталів новин Сьогодні.ua – http://www.segodnya.ua/. Даний портал новин надає читачам свіжі та актуальні новини політики, економіки, подій в Україні та світі, а також значні події спорту, культури, шоу-бізнесу та інших сфер життя. На цьому

порталі надається велика кількість актуальних новин у сфері освіти в Україні. Сайт постійно оновлюється, надаючи повну інформацію читачеві в будь-який час доби;

– сайти деяких вищих навчальних закладів на прикладі Харківського національного університету радіоелектроніки – http://nure.ua. Сайти вищих навчальних закладів містять переважно новини в рамках конкретного університету, але це дозволить користувачеві сайту залишатися в курсі не тільки глобальних новин, але й знати більш локальні події та новини, які можуть його зацікавити.

4 ПРАКТИЧНА РЕАЛІЗАЦІЯ

4.1 Опис поведінки користувача

Система, що проектується, спрямована на кінцевого користувача, який цікавиться сферою вищої освіти в Україні. Це може бути як студент, батьки, викладачі, так і абітурієнти та люди, які вже закінчили вищі навчальні заклади. Загальна взаємодія користувача із системою представлена на рисунку 4.1.

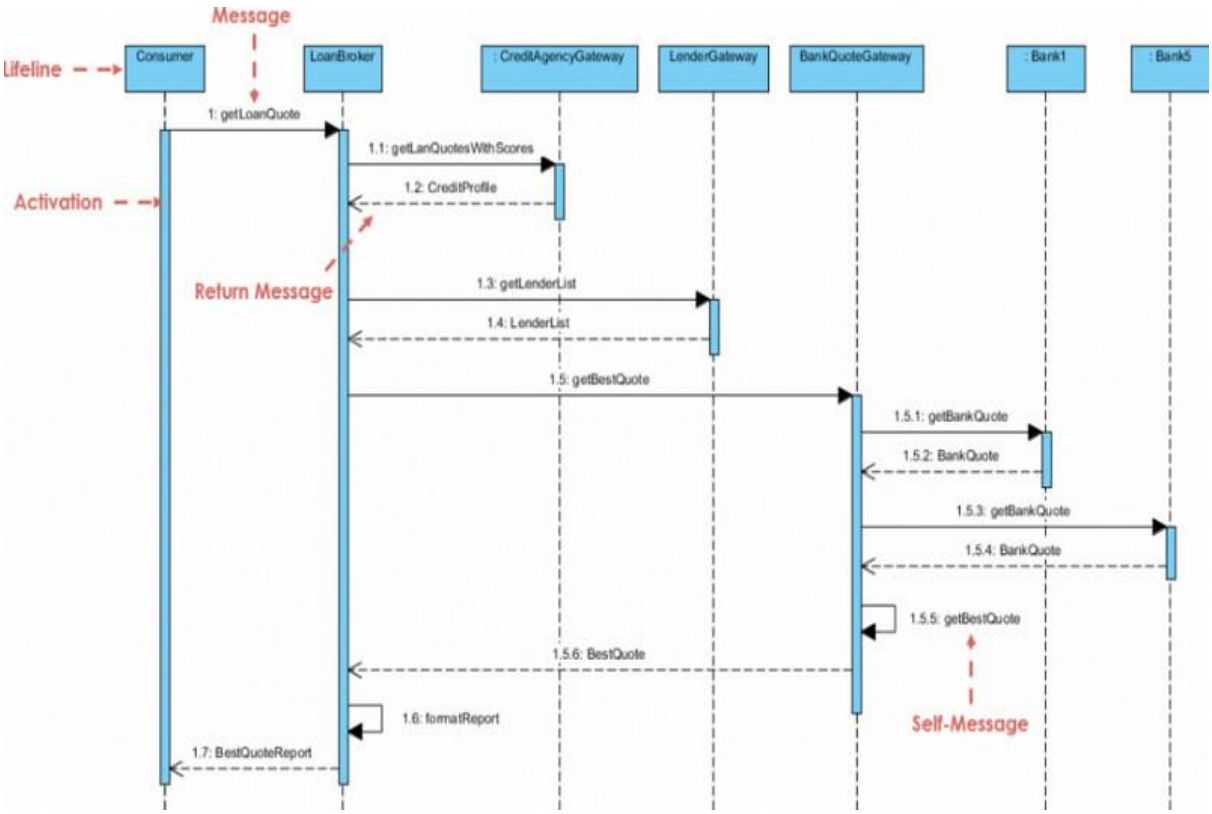

Рисунок 4.1 – Діаграма послідовності

Користувач заходить на веб-додаток через інтернет, тим самим надсилаючи запит на сервер. Далі, залежно від фільтрів, якщо такі було обрано користувачем, йде отримання даних зі сховища новин, які вже були оброблені додатком. Після цього йде запит на отримання нових даних з

вибраних джерел новин Отримавши нові дані – сервер їх аналізує поле чого зберігає нову інформацію. Групує нові дані з отриманими зі сховища – і видає їх користувачеві, оновлюючи у своїй, у разі потреби, доступні фільтри.

4.2 Функціональна структура mashup програми

На рисунку 4.2 представлено функціональну структуру web-додатка формату mashup для отримання інформації з новинної сфери вищої освіти України.

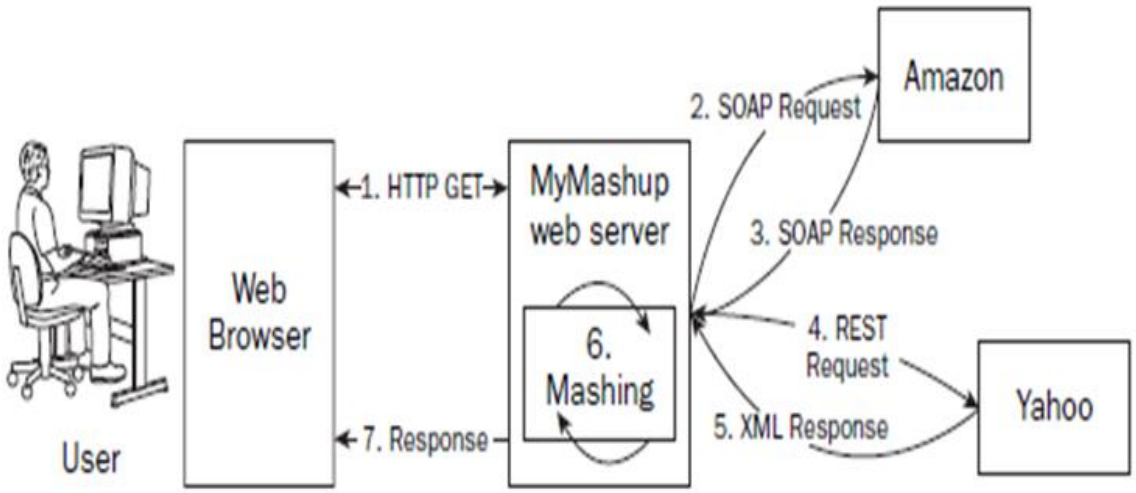

Рисунок 4.2 – Функціональна структура mashup програми

Вилучення сторінок з контентом новин. Ця функціональність передбачає парсинг головних сторінок описаних вище новинних сайтів з метою отримання посилань на конкретні сторінки з контентом новин.

Вилучення контенту новин. Ця функціональність передбачає парсинг сторінки з конкретною новиною та вилучення з неї лише корисної інформації та фільтруванням контенту, який не містить жодних корисних даних, таких як реклама, новини, посилання на інші сторінки.

Обробка контенту новин. Під обробкою контенту новин мається на увазі очищення отриманих на попередньому етапі даних. Вилучення необхідних зображень, даних про дату публікації і, звичайно, самого вмісту сторінки.

Далі отриманий вміст обробляється для:

– виділення ключових слів;

– визначення тональності вмісту статті новин.

Збереження отриманої інформації. Цей етап передбачає збереження в реляційній базі даних, пов'язаних на попередньому етапі даних, таких як дата публікації, назва, посилання на оригінальну статтю, список ключових слів, а також емоційне забарвлення тексту з цифровим показником.

Відображення новин користувачеві. Ця функціональність має на увазі рендеринг повченої інформації користувачеві з можливістю переходу на оригінальний сайт, а також фільтруванням новин за ключовими словами та тональністю статей.

4.3 Узагальнена діаграма класів програми

Діаграма класів (class diagram) служить уявлення статичної структури моделі системи у термінології класів об'єктно-орієнтованого програмування. Діаграма класів може відображати, зокрема, різні взаємозв'язки між окремими сутностями предметної області, такими як об'єкти та підсистеми, а також описує їхню внутрішню структуру та типи відносин.

На рисунку 4.3 представлена узагальнена діаграма класів web-програми, що розробляється у форматі mashup.

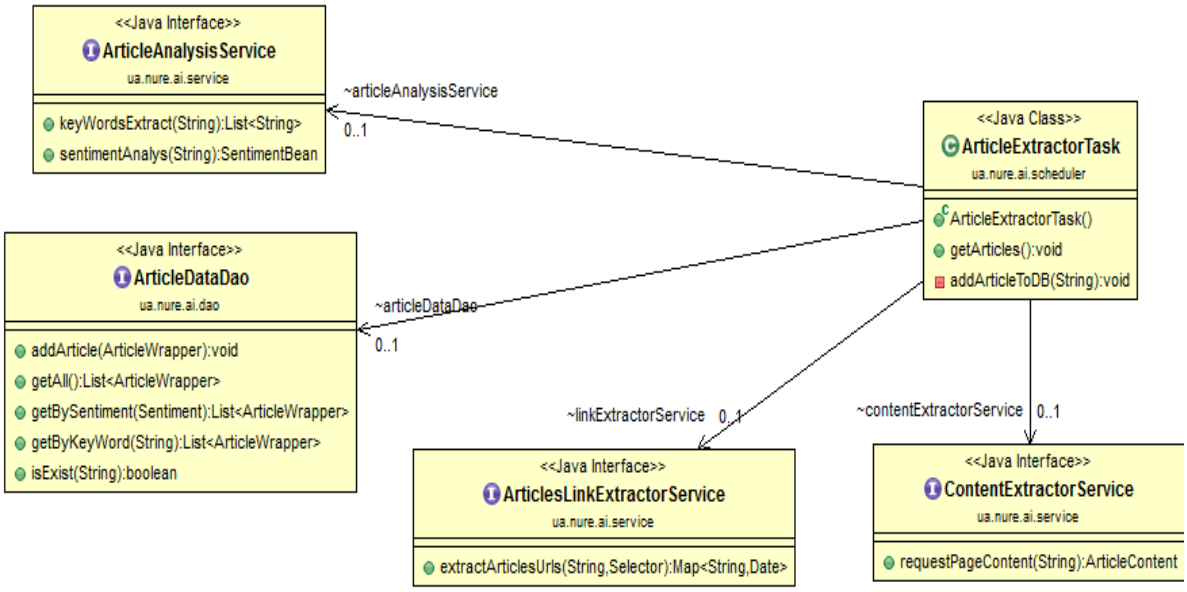

Рисунок 4.3 – Узагальнена діаграм класів програми

Дана діаграма включає в себе лише основні інтерфейси програми, що розробляється:

– ArticleAnalysisService – цей клас відповідає за основну обробку вмісту статті новин. Містить у собі два методи, перший з яких відповідає за вилучення ключових слів з контенту, що обробляється. Другий метод займається аналізом тональності запропонованого вмісту;

– ArticleDataDao – це сой DAO, який відповідає за збереження та отримання обробленої інформації зі сховища новин – MySql бази даних;

– ArticleLinkExtractorService – даний сервіс інкапсулює роботу з JSOUP бібліотекою для парсингу гавної сторінки новин на сайті та визначення посилань на сторінки, що містять основну новинну інформацію;

– ContentExtractorService – даний сервіс здійснює взаємодію з Readability Parser API для визначення на сторінці інформації, яка представляє з себе безпосередньо новину та супутні зображення.

4.4 Опис розробленого web-додатку

Грунтуючись на вибраних сайтах, що представляють новинну інформацію у сфері вищої освіти України, а також на вибраних методах аналізу web контенту, було розроблено web-додаток, який виконує такі основні функції:

– вилучення списку сторінок новин з перерахованих інформаційних сайтів для отримання набору посилань на web-сторінки, які містять інформацію у сфері вищої освіти України;

– інтеграція програми з сервісом Readability Parser API для позбавлення від непотрібної інформації, що присутня на сторінках з контентом новин;

– інтеграція програми з сервісом AlchemyAPI для отримання прихованої інформації про аналізовані новини, такої як емоційна спрямованість конкретної новини на основі аналізу тональності, так і отримання ключових слів та фраз;

– збереження нової знайденої інформації у реляційній базі даних для швидкого доступу;

– швидке вилучення новин з різних джерел даних та з бази даних для відображення користувачеві;

– фільтрація новин відповідно до емоційної спрямованості тексту, а також за ключовими словами та датою публікації.

Під час постановки завдання розглядалися сайти, контент яких доступний лише українською мовою. Але в ході реалізації, через неможливість підтримки української мови сервісом AlchemyAPI, список інформаційних сайтів було скорочено. У результаті додаток, що розробляється, в даний момент в змозі аналізувати інформацію з наступних інформаційних ресурсів: osvita.ua, nure.ua, segodnya.ua. На підставі описаної функціональності програми був розроблений інтерфейс програми, що розробляється. Вигляд головної сторінки подано на рисунку 4.4.

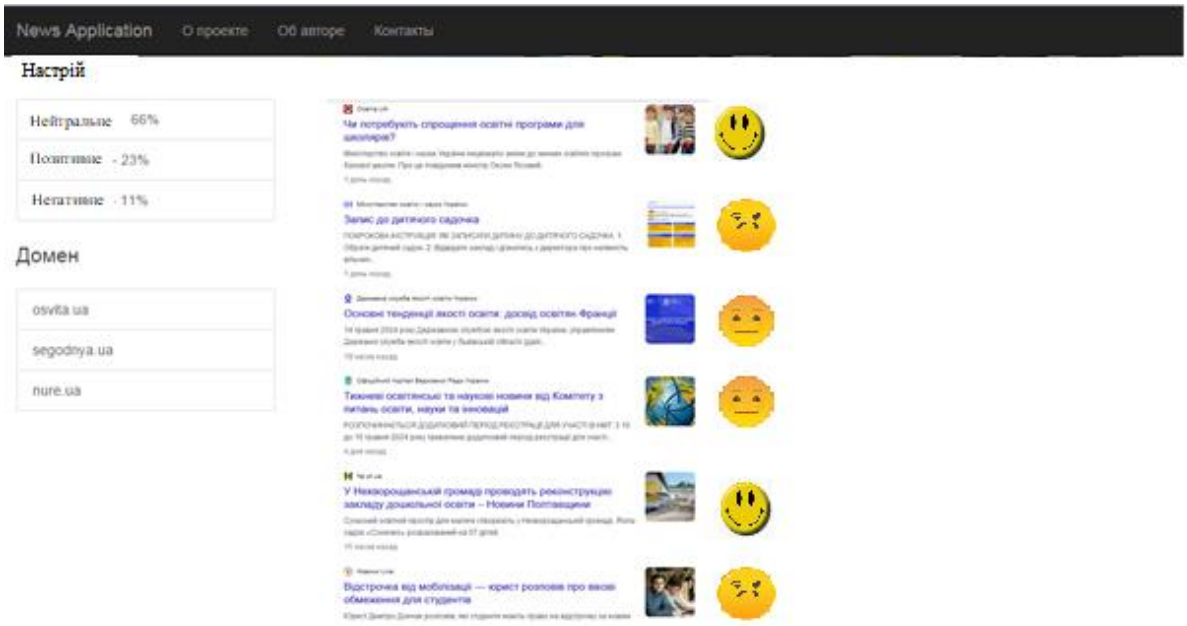

Рисунок 4.4 – Інтерфейс домашньої сторінки програми

Це основна сторінка, яка поєднує в собі посилання на всі повчені новини, а також дозволяє користувачеві фільтрувати відображувану інформацію. В основі фільтрів лежить результат обробки даних, описаний вище. Так, на рисунку 4.5 можна побачити основні критерії фільтрації новин.

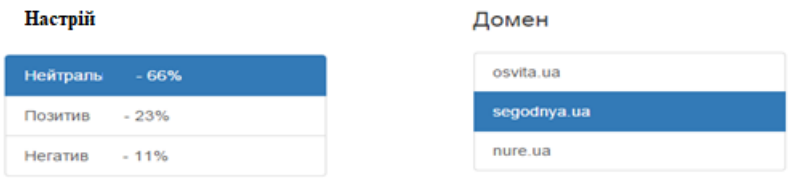

Рисунок 4.5 – Фільтрування домену та тональності

Крім фільтрів за категоріями «Домен» та «Настрій» користувач має можливість фільтрувати новини згідно з отриманими в результаті аналізу ключовими словами, які також доступні ліворуч від основної новинної інформації.

ВИСНОВКИ

У ході кваліфікаційної роботи проаналізовано поняття Web Mining, вивчено його основні види, цілі, сфери застосування, а також його зв'язок з іншими методами аналізу даних. Розкрито переваги та необхідність використання Web Mining у діяльності суб'єктів мережевої економіки: електронних магазинів, аукціонів, бірж. Його результати дозволяють оптимізувати роботу порталу, знайти похибки в дизайні, дізнатися про переваги відвідувачів, виділити серед них групи. Так з'являються можливості залучення більшої кількості клієнтів, підвищення їх лояльності, що в результаті призводить до зростання прибутку.

Крім вивчення поняття Web Mining, вивчено поняття mashup додатку, що дозволяє відображати в одному місці дані з різних джерел. Описані основні засоби реалізації mashup додатків, його архітектура та основні типи.

Описані теоретичні дослідження вплинули на розробку web-додатку, основною метою якого є збирання інформації у сфері вищої освіти України з різних інформаційних веб-орієнтованих джерел. Представлено функціональну модель такої програми, детально описано основні джерела даних, представлено діаграму поведінки користувача на сайті, а також представлено узагальнюючу діаграму класів програми. Розглянуті деякі методи аналізу даних, такі як виділення ключових слів у тексті та аналіз тональності тексту для подальшого їх застосування для аналізу інформації, отриманої з сайтів новин України.

Використання сторонніх сервісів допомогло більш автоматизувати процес вилучення контенту зі сторінки та звільненням від контенту, який не несе жодної інформації та знань. В цьому допоміг сервіс Readability Parser API.

Також використання стороннього сервісу Alchemy API надало можливість аналізувати виділений на попередньому етапі текстовий контент. Як аналізи були обрані виділення ключових слів у тексті, а також

аналіз тональності контенту новин.

Розроблений додаток за технологією mashup може бути корисним для всіх, хто зацікавлений у сфері вищої освіти України та вважає за краще отримувати новини з веб-орієнтованих джерел, не витрачаючи багато часу на це, оскільки додаток об'єднує всі основні освітні новини в рамках одного сайту.

Також, реалізована можливість фільтрації відображуваних новин значно спрощує використання програми та вносить більше гнучкості. Можливість перегляду новин не тільки відразу з усіх джерел, а лише з тих, що вибирає сам користувач – дозволяє прибирати джерела даних, яким користувач не довіряє.

Відображення новин згідно з їхньою емоційною складовою також є безперечною гідністю програми, оскільки дозволяє користувачу переглядати лише ті новини. які відповідає його настрою на даний момент.

Таким чином, можна зробити висновок, що технологія Web Mining дозволяє створювати програми зовсім іншого рівня з функціональністю та таким аналізом контенту, які були не доступні раніше.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Web Mining: інтелектуальний аналіз даних у мережі Internet. *Портал Знань*. URL: http://www.znannya.org (дата звернення: 10.05.2024).

2. Чубукова І.А. Data Mining: 2-ге вид. БІНОМ: Лабораторія знань, 2006. 382 с.

3. Eirinaki M., Vazirgiannis M. Web Mining для Web Personalization. *ACM Transactions on Internet Technology*. 2003. Vol.3. February. 260 с.

4. Khalil F., Li J., Wang H. A Framework of Combining Markov Model with Association Rules. *Australasian Data Mining Conference*. 2006. Vol.61.

5. Mashups: The New Breed of Web App. *IBM Developer Works*. URL: http://www.ibm.com/developerworks/x-mashups/index.html (дата звернення: 10.05.2024).

6. Агєєв М.С., Вершинніков І.В., Добров Б.В. Вилучення значущої інформації з web-сторінок для завдань інформаційного пошуку. *Інтернет-математика 2005. Автоматична обробка веб-даних*. 2005. 19 с.

7. Бабіна Д.В., Вороний С.М., Малащук Є.В. Підвищення ефективності отримання знань на основі інтелектуального аналізу та структурування інформації. *Штучний інтелект*. 2005. 15 с.

8. Миргород В.С., Личканенко І.С., Мазур Д.М., Родрігес Р.А. Метод інтелектуального аналізу інтернет-сторінок. Інформатика та комп'ютерні технології. ДонНТУ, 2012.

9. Markov Z., Larose D.T. Data-mining the Web: uncovering patterns in Web content, structure, and usage. Hardcord, 2007. 105 с.

10. Query languages for the WWW. URL: http://www.db.fmi.uni-passau.de/uni/WS97-98/Seminar/Ausarbeitungen/optimization.ps (дата звернення: 10.05.2024).

11. Mendelzon A., George A.M., Milo T. Querying the World Wide Web. *Int. J. on Digital Libraries*. 1997. 200 с.

12. Logical methods in computer science. URL: http://www.lmcs-

online.org (дата звернення: 11.05.2024).

13. Сфери застосування Data Mining. URL: http://bug.kpi.ua/stud/work/RGR/DATAMINING/spheresofapplication.html (дата звернення: 15.05.2024).

14. Mashup-додатки – еволюція SOA: Частина 1. Web 2.0 та основні концепції. URL: http://www.ibm.com/developerworks/ua/library/ws-soa-mashups/ (дата звернення: 16.04.2024).

15. Mashups: The new breed of Web app. URL: http://www.ibm.com/developerworks/xml/library/x-mashups/index.html (дата звернення: 24.04.2024).

16. Гібридні програми: нове покоління web-додатків. URL: http://www.ibm.com/developerworks/ua/library/x-mashups/ (дата звернення: 15.05.2024).

17. Li D., Kaichang D., Shi X. Mining asociation rules with linguistic cloud models, Research and Development in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science. 1998. 90 с.

18. Chia-Hui Ch., Mohammed K., Moheb R.G., Khaled F.S. A survey of Web Information Extraction. *IEEE Transactions on Knowledge and Data Engineering*. 2006. № 18/10. P. 50–54.

19. Інформаційний технології в лінгвістиці, обробка інформації. URL: http://ua.wikiversity.org/wiki/Інформаційний_технології_в_лінгвістиці/Обробка_інформації (дата звернення: 15.05.2024).

20. Galitsky B., Huanjin Ch., Shaobin D. Inversion of Forum Content Based on Authors' Sentiments on Product Usability. *AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0*. 2019. 56 с.