УДК 004.8

В.Ю. Варнава
О.В. Сілагін

РОЗРОБКА ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПОШУКУ ІНФОРМАЦІЇ В ТЕКСТАХ

Вінницький національний технічний університет

Анотація

Дана магістерська кваліфікаційна робота присвячена розробці програмного забезпечення для вилучення інформації в тексту за запитом природньою мовою . Були розглянуті та проаналізовані існуючі програмні рішення і їх функціональні можливості, та обрано гібридний режим функціонування. Було проаналізовано різні підходи до вирішення задачі аналізу текстів. Було спроектовано програму пошуку інформації в текстах, написану мовою програмування JavaScript з використанням бібліотеки React для клієнтської частини, Node.js і Express для серверної частини та СУБД MongoDB для управління базами даних.

Ключові слова: Класифікація тексту, природна мова, веб-клієнт, клієнт-серверна архітектура.

Abstract

This master's thesis is devoted to the development of software for extracting information from text on request in natural language. Existing software solutions and their functionality were considered and analyzed, and a hybrid mode of operation was chosen. Different approaches to solving the problem of text analysis were analyzed. A program for searching for information in texts was designed, written in the JavaScript programming language, using the React library for the client part, Node.js and Express for the server part, and the MongoDB DBMS for database management.

Key words: text classification, natural language, web client, client-server architecture.

Вступ

Ми живемо в час, коли обсяги виробленої людством інформації більше, ніж будь-коли і кількість цих даних зростає з кожним днем. Однак значну користь з цієї інформації можна отримати лише при правильній обробці і аналізі цих даних. Зараз щомиті по всьому світу створюються гігабайти нових даних різного виду: робляться нові знімки, відеозапису, пишуться сотні відгуків до товарів в інтернет-магазинах, тисячі коментарів під записами на Facebook, десятки рецензій до фільмів в онлайн-кінотеатрах, ціни на акції то злітають, то падають. І велика частина цих даних в "сирому" вигляді практично марна. Щоб витягти з них якусь користь, їх потрібно відфільтрувати і обробити. [1].

Постановка задачі

Вхідними даними для програмного модуля блог-систем є: користувачі, їх тексти, кількість символів у тексті не більша за 500, кількість символів у запиті не більша за 200,мова - англійска.

Вихідними даними є зручний додаток та інформація, що була знайдена за запитом. Так як модуль матиме форму веб-додатка, він буде кросплатформеним і не залежатиме від конкретного апаратного забезпечення.

Розроблене програмне забезпечення має функціонувати у всіх браузерах та пристроях. Усі наявні елементи повинні чітко працювати без будь-яких помилок.

Початком розробки інтелектуальної технології пошуку інформації в текстах є вибір методів аналізу тексту. Далі – розробка загальної структурної схеми додатку, розробка загальної схеми алгоритму роботи додатку, створення UML-діаграм розгортування та послідовності для деталізації складових подальшої розробки. Наступним етапом є вибір технологій реалізації та безпосередня розробка інтелектуального додатку. Заключними етапами є тестування розробленого інтелектуальної технологііі розробка інструкції користувача.Можна побачити, що існуючі рішення є вузькоспеціалізованими та не покривають більшість випадків необхідності монетизації. Тому стоїть проблема розробки

компромісного рішення, що дозволить користувачу повністю отримувати кошти, зароблені з допомогою монетизації, та обрати гнучке рішення для себе, без потреби у програмуванні.

Перший вид методів це методи, в основі яких лежать техніки розпізнавання іменованих сутностей. Іменована сутність - це група слів в тексті, яка описує реальний об'єкт. Наприклад, Apple Inc., John Broun, information extraction і т.д. Пошук будь-яких іменованих сутностей ведеться в тексті за допомогою патерна. За способом знаходження паттерна методи діляться на підходи, засновані на правилах, і на статистичні підходи.

Методи, засновані на правилах, (наприклад, [1]) знаходять паттерн в тексті, редуцируючи узагальнені правила. Наприклад, з правила є числом виходять більш специфічні правила є 4-значним числом або є дробовим числом. Для обчислення правил використовуються розмічена вручну навчальна база, тому для великих текстів даний підхід не застосовується.

Метою дослідження є розширення функціональних можливостей в аналізі текстів за допомогою створення додатку для браузера, що дозволить покращити доступність системи.

Об'єктом дослідження є процеси пошуку інформації з текстах.

Предметом дослідження є алгоритми та програмне забезпечення, що організовує процес пошуку інформації з текстах

Результати дослідження

Класифікація – процес групування та організації інформацію змістовно та систематично у Перший вид методів це методи, в основі яких лежать техніки розпізнавання іменованих сутностей. Іменована сутність - це група слів в тексті, яка описує реальний об'єкт. Наприклад, Apple Inc., John Broun, information extraction і т.д. Пошук будь-яких іменованих сутностей ведеться в тексті за допомогою патерна. За способом знаходження паттерна методи діляться на підходи, засновані на правилах, і на статистичні підходи.

Методи, засновані на правилах, (наприклад, [1]) знаходять паттерн в тексті, редуцируючи узагальнені правила. Наприклад, з правила є числом виходять більш специфічні правила є 4-значним числом або є дробовим числом. Для обчислення правил використовуються розмічена вручну навчальна база, тому для великих текстів даний підхід не застосовується.

Статистичні підходи (наприклад, система Nymble]) використовують варіації EM-алгоритму для знаходження розподілів токенів по сутностей. Зокрема, в Nymble використовуються приховані Марковскі моделі. Оскільки припущення, що токени смислу розподілені по нормальному закону в усьому тексті може бути неприйнятно в разі неструктурованого джерела, даний підхід нам також нецікавий.

Методи засновані на базах данних представлені підходом, описаним Matthew Michelson і Craig A. Knoblock [4] і характеризуються використанням при аналізі змісту тексту якоїсь бази знань про об'єкти будь-якого типу. Загальна структурна схема інтелектуального модуля аналізу текстів на наявність образливих висловлювань складається з 5 основних компонентів:

- веб-клієнт;
- програмний інтерфейс додатку(API);
- панель адміністрування;
- система вилучення інформації;
- база даних.

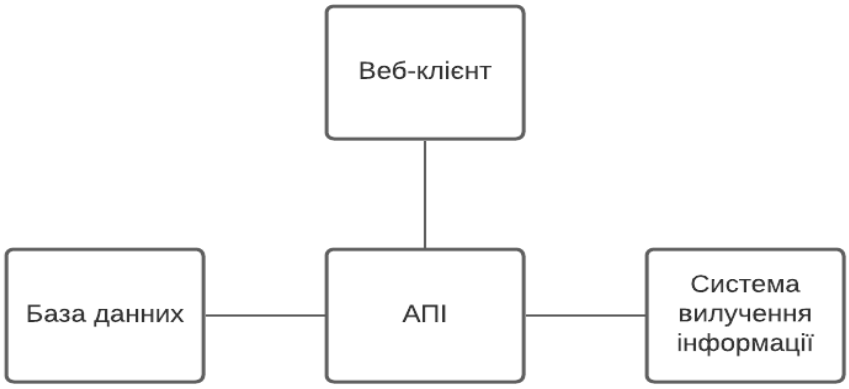

Рисунок 1 – загальна структурна схема інтелектуального модуля пошуку інформації в тексті

Для проектування фізичної та логічної організації компонентів інтелектуальної технології аналізу текстів використаємо діаграму розгортування. Діаграма розгортування – вид UML-діаграми, яка демонструє архітектуру виконання системи, включає в себе такі вузли, як апаратні або програмні середовища виконання, а також проміжне програмне забезпечення, що їх сполучує. Діаграма розгортування для інтелектуальної технології аналізу текстів включає в себе наступні вузли:

- веб-клієнт;
- сервер;
- хостинг бази даних;
- база даних;
- СУБД;
- програмний інтерфейс;
- система пошуку інформації в тестах

Усі компоненти пов’язані між собою мережею, тобто зв’язком на основі протоколів TCP/IP. Розроблена діаграма розгортування представлена на рисунку 2.5.

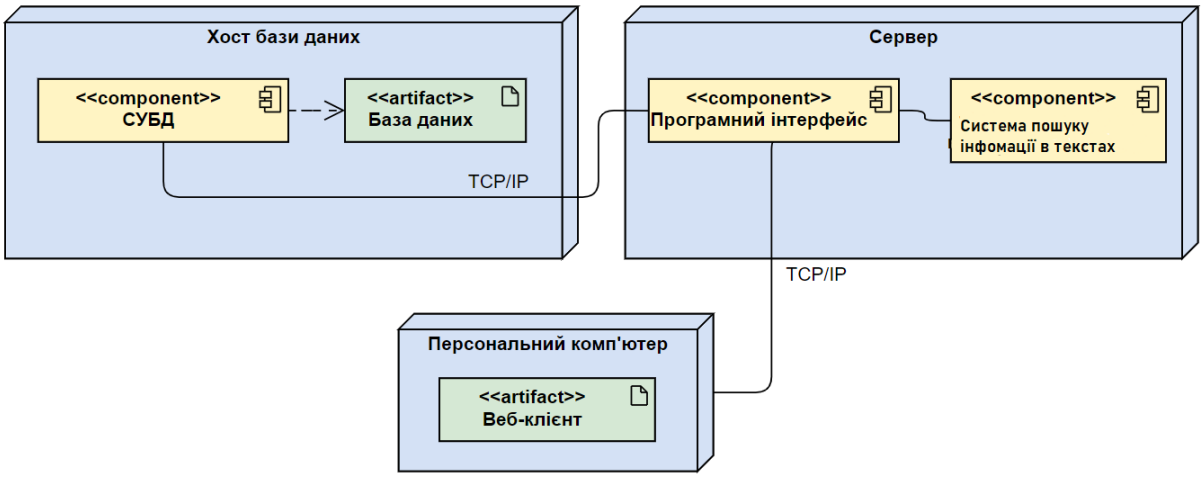

Рисунок 2 – Діаграма розгортування інтелектуальної технології пошуку інформації в текстах

Висновки

1. Проаналізовано існуючі рішення та аналоги інтелектуального додатку пошуку інформації в текстах.
2. Обґрунтовано вибір методу розв’язання задачі.
3. Розроблено загальну структурну схему інформаційної технології.
4. Розроблено основний алгоритм роботи інформаційної технології.
5. Змодельовано інформаційну технологію із використанням мови UML.
6. Обрано мову програмування, бібліотеки, фреймворки та систему управління базами даних для програмної реалізації інформаційної технології.
7. Розроблено програмну реалізацію компонентів інформаційної технології

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Matthew Michelson and Craig A. Knoblock. Creating Relational Data from Unstructured and Ungrammatical Data Sources. In Journal of Articial Intelligence Research 31 (2008), pages 543- 590, 2008.
2. MongoDB [Електронний ресурс] – Режим доступу: https://searchdata.com
3. M. Iyyer, V. Manjuhatha, J. Boyd-Graber, Hal Daume II. Deep Unordered Composition Rivals Syntactic Methods for Text Classification,
4. M. Iyyer, V. Manjuhatha, J. Boyd-Graber, Hal Daume III. Deep Unordered Composition Rivals Syntactic Methods for Text Classification.
5. Загальна характеристика UML [Електронний ресурс] – Режим доступу: http://www.informicus.ru/default.aspx?SECTION=6&id=73&subdivisionid=2.

Варнава Владислав Юрійович – студент кафедри комп’ютерних наук, Вінницький національний технічний університет, м. Вінниця. e-mail: vladyslav.varnava@gmail.com .

Сілагін Олексій Віталійович – канд. техн. наук, доцент кафедри комп’ютерних наук, Вінницький національний технічний університет, м. Вінниця. e-mail: avsilagin@vntu.edu.ua .

Varnava Vladyslav Yuriyovych – student of Computer Science Department, Vinnytsia National Technical University, Vinnytsia, e-mail: vladyslav.varnava@gmail.com.

Silagin Oleksiy Vitalyevich – Cand Sc. (Eng.), Associate Professor of Computer Science Department, Vinnytsia National Technical University, Vinnytsia, e-mail: avsilagin@vntu.edu.ua.