

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

Бойченко Олексій Сергійович

УДК 004.946

**АВТОМАТИЧНЕ ГЕОКОДУВАННЯ НОВИННОЇ ІНФОРМАЦІЇ НА ОСНОВІ
КОНТЕНТ-АНАЛІЗУ**

Спеціальність 8.05010102 – Інформаційні технології проектування

Автореферат
дисертації на здобуття освітньо-кваліфікаційного рівня магістр

Київ – 2016

Робота виконана на кафедрі автоматизації проектування енергетичних процесів та систем НТУУ «КПІ» Міністерства освіти і науки України.

Науковий керівник: кандидат технічних наук
Кузьмійх Валерій Олександрович
доцент кафедри автоматизації проектування
енергетичних процесів і систем НТУУ «КПІ»
(м. Київ)

Рецензент: кандидат технічних наук, професор
Отрох Сергій Іванович
Старший аналітик технічної дирекції ПАТ «Укртелеком»
(м. Київ)

Захист відбудеться ____ червня 2016 р., на засіданні ДЕК кафедри АПЕПС НТУУ
"КПІ" аудиторія ____.

З дисертацією можна ознайомитись у методичному кабінеті кафедри АПЕПС
НТУУ "КПІ", аудиторія ____.

Реферат підготовлено та представлено до розгляду " __ " _____ 2016 р.

Робота рекомендована до захисту " __ " _____ 2016 р.

Завідуючий кафедрою АПЕПС НТУУ "КПІ",
доктор технічних наук, професор

Лук'яненко С. О.

Відповідальний за випуск магістрів
кафедри АПЕПС НТУУ «КПІ»,
кандидат технічних наук, доцент

Гагарін О. О.

АВТОРЕФЕРАТ

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. У зв'язку з постійним зростанням кількості інформаційних ресурсів, зростає і навантаження на аналітичні центри, що оброблюють їх. Сьогодні новинна інформація в мережі Інтернет складається, аналізується та оброблюється практично без використання засобів автоматизації. Відсутність аналітичних систем, які можуть оброблювати результати пошукових запитів користувачів значно ускладнює і уповільнює пошук необхідних об'єктів.

Зв'язок роботи з науковими програмами, планами, темами

Дисертаційна робота магістра виконувалась у Національному технічному університеті України "КПІ" згідно з планом науково–дослідницьких робіт кафедри автоматизації проектування енергетичних процесів і систем.

Мета та завдання дослідження. Метою дослідження магістерської дисертації є пошук закономірностей, нових підходів щодо виявлення у тексті географічних назв, та подальшого їх використання для геокодування користувацьких запитів та новинної інформації.

Для реалізації поставленої мети були сформульовані наступні **завдання дослідження**, що визначили логіку дослідження та його структуру:

- проаналізувати сучасні системи обробки користувацьких запитів та їх функціонал;
- проаналізувати методи пошуку географічних даних в тексті;
- розробити систему автоматичного геокодування новинної інформації та користувацьких запитів.

Об'єктом дослідження є інформаційні технології проектування обробки даних.

Предметом дослідження є інформаційні технології проектування аналітичних систем пошуку географічних назв у тексті.

Методи дослідження: розв'язання поставлених задач виконувалось на базі положень аналізу текстових даних, зокрема:

- метод визначення множини досліджуваних джерел для накопичення даних, що необхідно проаналізувати;
- метод виявлення одиниць аналізу для виявлення ключових слів.

Наукова новизна одержаних результатів:

- удосконалено метод контекстного пошуку географічних назв в тексті шляхом групування етапів контент-аналізу, що дозволило пришвидшити обробку даних;
- набуло подальшого розвитку застосування методів аналізу тексту для автоматичного геокодування інформації.

Практичне значення одержаних результатів

Розроблене програмне забезпечення використовується для автоматичного геокодування користувацьких запитів.

Апробація результатів дисертації

Основні положення роботи доповідались і обговорювались на:

1. VII міжнародний науково–практичний семінар «Економічна безпека держави і науково-технічні аспекти її забезпечення» (21–22 жовтня 2015, М. Київ).
2. V Міжнародної міжвузівської школи–семінару «Методи і засоби діагностики в техніці та соціумі» (16–19 листопада 2015, м. Івано–Франківськ).
3. XIV Міжнародній науково–практичній конференції «Сучасні проблеми наукового забезпечення енергетики» (м. Київ, 18–21 квітня 2016 р.).

Публікації. Наукові положення дипломної роботи опубліковані у 3 роботах.

Ключові слова. *ГЕОКОДУВАННЯ, АНАЛІЗ ІНФОРМАЦІЇ, ІНФОРМАЦІЙНА СИСТЕМА, ОБРОБКА ДАНИХ, НОВИНИ, ГЕОГРАФІЧНІ НАЗХВИ, ПОШУК ІНФОРМАЦІЇ, КОНТЕНТ–АНАЛІЗ.*

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** розкрито поняття автоматичного геокодування новинної інформації та сучасний стан систем обробки даних. Обґрунтовано необхідність створення програмного забезпечення для реалізації процесу автоматизації процесу геокодування новинної інформації. Визначено мету наукового дослідження – пошук закономірностей, нових підходів щодо виявлення у тексті географічних назв, та подальшого їх використання для геокодування користувацьких запитів та новинної інформації.

В **першому** розділі "Сучасний стан систем автоматичного геокодування інформації" розглянуто існуючі програмні рішення для геокодування та обробки запитів. Метою даного розділу стало дослідження існуючих засобів створення складних поверхонь.

У зв'язку з постійним зростанням кількості інформаційних ресурсів, зростає і навантаження на аналітичні центри, що оброблюють їх. Сьогодні новинна інформація в мережі Інтернет складається, аналізується та оброблюється практично без використання засобів автоматизації. Тобто всі наявні дані проходять через аналітичні центри, де сортувати та оброблювати їх по деяким ознакам необхідно людині.

В сучасних умовах досить гостро постало питання економії часу. Не тільки примхи, а й щоденні потреби людини потребують пошуку оптимальних об'єктів. Сьогодні існує декілька продуктів для пошуку місцеположення об'єктів, але систем, де функції пошуку інформації та гео–прив'язки поєднувалися в зручній для мобільних пристроїв формі для одноекранного перегляду інформації не існує.

Також у першому розділі сформовано постановку задачі автоматичного геокодування інформації.

У **другому** розділі "Методичний апарат автоматичного геокодування інформації" розглядається методи аналізу текстових даних.

Особливість аналізу текстового матеріалу полягає в потребі дослідника проникати в зміст розглянутого їм твору і в акцентуванні рухомої структури тексту. Пропонується в процесі аналізу тесту, для отримання семантичного уявлення,

розглядати його як деяку ієрархічну структуру. Як правило, для цього застосовується метод шаблонів: вхідний текст розділяється за допомогою шаблонів, потім проводиться обробка отриманих даних.

Таким чином, очевидно, що використовуються різні варіанти аналізу текстових матеріалів документів, в яких виділяють два основних типи:

1. Традиційний, або класичний.
2. Формалізований, або кількісний (наприклад, контент–аналіз).

Розрізняючись між собою, вони не виключають, а взаємно доповнюють один одного, дозволяючи компенсувати недоліки кожного.

Традиційний аналіз являє собою сукупність певних логічних побудов, спрямованих на розкриття основного змісту досліджуваного матеріалу і дозволяє перетворити первісну форму містимої в документі інформації в форму, що цікавить дослідника (аналітика).

Такий аналіз включає різноманіття розумових операцій, спрямованих на інтерпретацію відомостей, що містяться в документі, з певної точки зору прийнятої дослідником в кожному конкретному випадку. Традиційний аналіз є самостійним, творчим процесом, залежним від:

- змісту та спрямованості досліджуваного документа;
- умов, цілей і завдань проведеного дослідження;
- наукової кваліфікації, багатства досвіду і творчої інтуїції дослідника (в цьому випадку він таїть в собі можливість суб'єктивної оцінки та інтерпретації досліджуваного документа).

Контент–аналіз – це алгоритм збору інформації, розроблений на основі систематичного виявлення цілей і завдань дослідження характеристик текстів (понять, дієслів, і ін.), передбачає використання певних стандартизованих процедур, які забезпечують формалізацію та вимір досліджуваних ознак, що дозволяє робити висновки про особливості досліджуваного об'єкта.

Аналітично–синтетична переробка (АСП) документів полягає в представленні кожного окремого документа або їх певної сукупності в такому вигляді, який максимально відповідає тій чи іншій завданню науково–інформаційної діяльності.

Потреба в аналізі та синтезі наукової інформації, здійснюваному спеціальними службами, виникла внаслідок прискореного розвитку науки і техніки.

У другому розділі також розглядаються технології серверної обробки даних та методи створення інтерактивного інтерфейсу користувача. Інтерактивність – це поняття, що розкриває характер і ступінь взаємодії між об'єктами. В інформаційних системах інтерактивність – це здатність системи адекватно, тобто як того очікує користувач, реагувати на його дії, виконуючи ту чи іншу операцію.

Інтерактивне відображення новинної інформації передбачає відображення всіх необхідних даних у компактному та організованому вигляді.

У **третьому** розділі "Опис програмної реалізації програмного комплексу" описується програмна реалізація системи автоматичного геокодування новинної інформації на основі контент-аналізу.

Система автоматичного геокодування новинної інформації має розподілену архітектуру та складається з бази даних для збереження обробленої інформації та налаштувань користувачів, веб сервісу для обробки текстової інформації та мобільного застосунку для взаємодії користувача з системою.

Розробка виконувалась під управлінням операційних систем Windows 10 та OS X у середовищах розробки Visual Studio Community 2015 та Xcode на мовах програмування C# та Swift з використанням фреймворків .Net Core, UIKit та MapKit с також серверу баз даних MsSql.

Можна виділити чотири основні функції системи автоматичного геокодування новинної інформації: обробка користувацьких запитів, аналіз текстової інформації, налаштування джерел інформації та формування даних для мобільного додатку. Дані функції є композитами, складеними з інших підфункцій. З метою їх виявлення було проведено функціональну декомпозицію наведених вище функцій та розбиття на підфункції.

Функція завантаження даних відповідає за завантаження даних по запиту або з заздалегідь визначених джерел. Функція аналізу текстових даних складається з наступних підфункцій:

- виділення географічних назв з текстових даних;

- геокодування знайдених географічних назв.
- Функція налаштування джерел складається з:
- налаштування rss–джерел;
- налаштування пошукових машин.

Під функцією формування відповіді на користувацький запит мається на увазі серіалізація всіх даних зібраних за користувацьким запитом.

Функція завантаження даних по запиту складається з наступних підфункцій:

- завантаження даних з пошукової машини;
- завантаження даних з rss–джерела.

Усі функції програмного забезпечення для інтерактивного відображення новинної інформації можна розділити на дві групи:

- обмін даними з веб–сервісом;
- робота з мапою.

Дані функції є композитними, тобто складаються з деякого числа підфункцій.

Функція обміну даними з веб–сервісом складається з наступних підфункцій:

- надсиання запиту до веб–сервісу;
- обробка відповіді.

В свою чергу підфункція обробки відповіді містить наступні підфункції:

- парсинг результату;
- оновлення інтерфейсу;
- оновлення мапи.

Надсилення запиту до сервісу передбачає формування JSON об’єкту з параметрами запиту, його мініфікацію, конвертацію в строку на надсилення запиту.

Взаємодія з веб–сервісом відбувається використовуючи протокол HTTP.

Парсинг результату при обробці відповіді відповідає за перетворення строкового результату у JSON об’єкт, з якого оновлюється модель даних для відображення. Оновлення інтерфейсу відповідає за оновлення візуальних компонентів відповідно з моделлю даних для відображення. Оновлення мапи аналогічно оновлює картографічний компонент.

Підсистема роботи з мапою містить:

- встановлення мітки на мапі;
- видалення мітки з мапи;
- позиціонування на заданий баундбокс.

Встановлення мітки на мапі, очевидно, встановлює “булавку” на мапу, видалення – видаляє. Позиціонування на заданий баундбокс позиціонує мапу таким чином, що б вона була максимально детальною, але в той же час що б на ній були відображені всі задані маркери.

Також третій розділ присвячений опису інтерфейсу користувача та описує типові сценарії роботи користувача з системою автоматичного геокодування новинної інформації.

Для запуску програмного забезпечення необхідно запустити додаток для обробки користувацьких запитів ("GeoNewsQuery.app") на мобільному пристрої. Після запуску програмного забезпечення користувач бачить головну форму застосунку.

Інтерфейс iOS додатку складається з трьох основних частин (вкладок): пошук, новини, налаштування. (рисунок 1).

Після введення пошукового запиту користувач натискає кнопку «Ввод», що призводить до обробки запиту та виводу результатів в область під строкою пошуку. Кожен з отриманих елементів в списку результатів обробки запиту має заголовок, коротку інформацію про результат пошуку та найкоротшу відстань до нього. При натисненні на результат застосунок переходить до відображення детальної інформації про результат.

Друга вкладка «Новини» відображає новини з джерел, обраних користувачем, в хронологічному порядку. Не прочитані новини відмічаються синьою точкою з правої сторони новини, також кількість непрочитаних новин відмічається в червоній бульбашці на вкладниці.

При перегляді детальної інформації користувачу надається можливість скористатися геокодованими результатами для навігації до об'єкту.

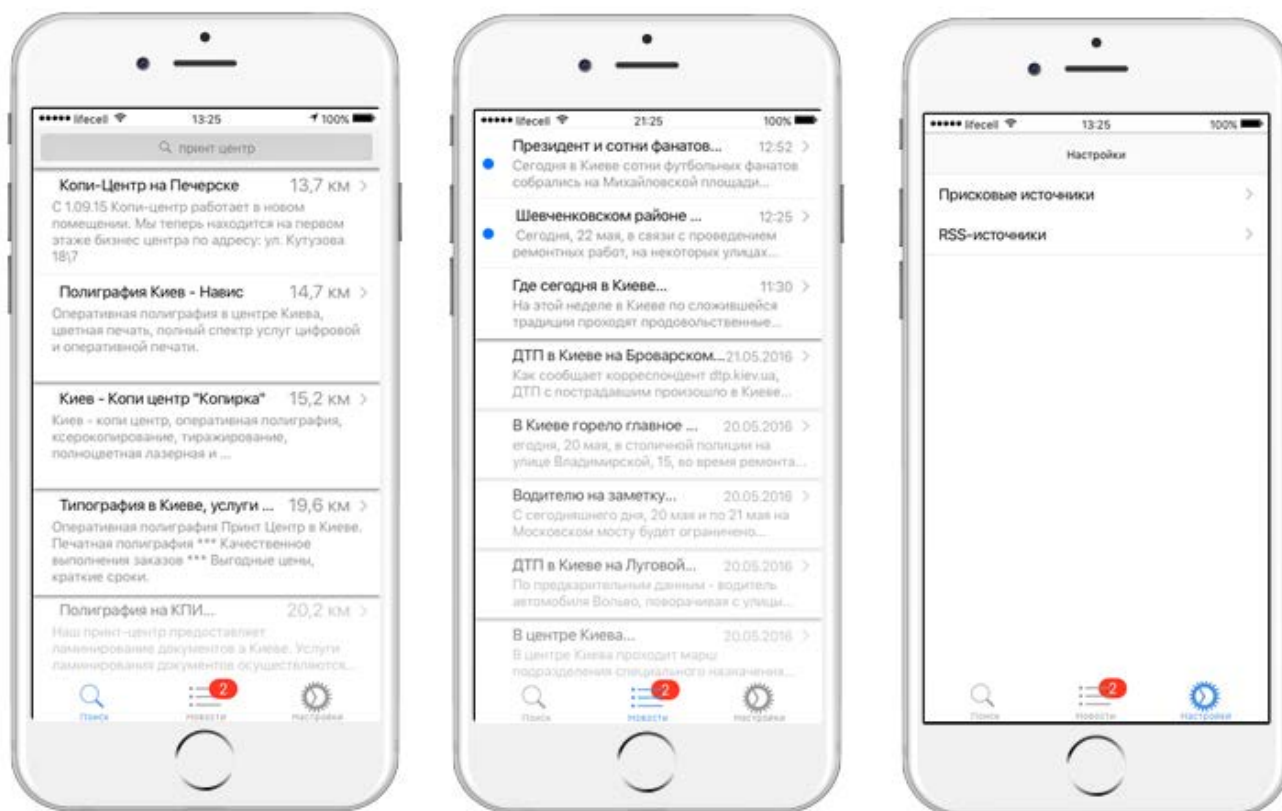


Рисунок 1 – Интерфейс iOS додатку

Екран налаштувань дозволяє обрати активні, додавати та видаляти пошукові машини та джерела новинної інформації.

Створена комп'ютерно-інформаційна система значно спрощує розуміння географічної прив'язки текстової інформації, що полегшує та прискорює визначення оптимальних об'єктів.

ВИСНОВКИ

Результати виконаного дослідження дозволяють зробити наступні **висновки**. Проаналізувавши системи аналізу та геокодування текстових даних, було виявлено, що сучасні системи не мають механізмів автоматичного аналізу новинної для виявлення географічних місць. Тому було спроектовано програмний комплекс для автоматичного геокодування новинної інформації на основі контент-аналізу, що дозволяє отримувати оброблені дані за допомогою загальнодоступних засобів (мобільних телефонів), а не оброблювати інформацію власноруч. Створене програмний комплекс може бути використаний як обгортка існуючих пошукових

сервісів, як просунутий новинний агрегатор та як службовий сервіс в організаціях та установах, пов'язаних з логістичною або геоінформаційною діяльністю.