МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ЧОРНОМОРСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ПЕТРА МОГИЛИ

ЯЦУНЕНКО АНДРІЙ АНДРІЙОВИЧ

УДК 004.925.04

ІНФОРМАЦІЙНО-ПОШУКОВА СИСТЕМА АНАЛІЗУ НАУКОВО-ТЕХНІЧНИХ ПУБЛІКАЦІЙ ТА ЇХ ІНТЕЛЕКТУАЛЬНЕ ПРЕДСТАВЛЕННЯ

Спеціальність 122 – Комп’ютерні науки

Автореферат

магістерської роботи на здобуття кваліфікації
«Магістр комп’ютерних наук»

Миколаїв – 2019

Магістерська наукова робота є рукопис.

Робота виконана в Чорноморському національному університеті імені Петра Могили Міністерства освіти і науки України на кафедрі інтелектуальних інформаційних систем

Науковий керівник: професор, доктор технічних наук
Микола Тихонович Фісун,
ЧНУ ім. Петра Могили,
завідувач кафедри інженерії програмного забезпечення

Рецензент: професор, доктор технічних наук
Ігор Іванович Коваленко,
ЧНУ ім. Петра Могили,
завідувач кафедри інтелектуальних інформаційних систем

Захист відбудеться «25» лютого 2019 р. о 9^{30} на засіданні екзаменаційної комісії (ауд. 2-403) у Чорноморському національному університеті імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

З магістерською науковою роботою можна ознайомитися в бібліотеці Чорноморського національного університету імені Петра Могили за адресою: 54003, м. Миколаїв, вул. 68-ми Десантників, 10.

Автореферат представлений «23» лютого 2019 р.

Секретар
екзаменаційної комісії,
к.пед.н., доцент Н. М. Болюбаш

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Електронна інформація відіграє все більшу роль у всіх сферах життя сучасного суспільства. В інформаційних сховищах, розподілених по всьому світу, зібрані террабайти текстових даних. Розвиток інформаційних ресурсів Інтернет багаторазово посилює проблему інформаційного перевантаження.

У зв'язку із цим, виникає необхідність в швидкій розробці прикладних програмних систем (застосунків) для автоматичної або автоматизованої обробки текстів на природній мові. Прикладами такої обробки є збір і фільтрація даних з різних джерел, витяг знань, реферування, анотування і т.п.

Технологія ефективного аналізу тексту Text Mining здатна виступити в ролі репетитора, який, простудіювавши весь курс, викладає лише найбільш ключову і значущих інформацію. Таким чином, користувачеві не треба самому "просівати" величезну кількість неструктурованої інформації. Розроблені на основі статистичного та лінгвістичного аналізу, а також штучного інтелекту, технології Text Mining якраз і призначені для проведення смислового аналізу, забезпечення навігації і пошуку в неструктурованих текстах. Застосовуючи побудовані на їх основі системи, користувачі зможуть отримати нову цінну інформацію - знання.

Мета та завдання дослідження. Метою даної дипломної роботи є виявлення найбільш зручного та ефективного програмного додатку у сфері інтелектуального аналізу ПМ текстів.

Для досягнення даної мети в магістерській роботі поставлені та вирішені наступні завдання:

Завдання:

– виконати аналіз досліджень в галузі «text mining» та розглянути проблеми, які зустрічаються у цій сфері;

– виконати аналіз математичних моделей, методів та алгоритмів використовуваних під час розробки програмних застосунків, та зв’ясувати, які переваги надають ті, чи інші методи та алгоритми;

– розглянути існуючі на сьогодні системи інтелектуального аналізу текстів та порівняти їх між собою;

– провести апробацію обраної системи з розглянутих на прикладі декількох науково-технічних публікацій для демонстрації роботи такої системи.

Об’єктом дослідження є процеси інтелектуального аналізу даних текстів написаних природною мовою.

Предметом дослідження є зручний та ефективний програмний застосунок для демонстрації широких та корисних можливостей технології «text mining» у сучасному інформаційному суспільстві.

Практичне значення одержаних результатів: результати роботи використані у поточній діяльності підприємства «Добробут» при підготовці теплотрас до опалювального сезону 2018/2019 рр., що підтверджено відповідним Актом впровадження.

Публікації. Основні положення та результати магістерської роботи опубліковані у збірнику матеріалів міжнародної науково-практичної конференції.

Структура та обсяг роботи. Магістерська наукова робота складається із вступу, ____ розділів, висновків, додатків. Загальний обсяг роботи складає ____ сторінки, ___ рисунків, ____ таблиць та ____ посилань на літературні джерела.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** подано обґрунтування актуальності теми магістерської роботи, зазначено її зв’язок із науковою програмою, планами і темами, сформульовано мету та завдання дослідження, а також визначено об’єкт та предмет дослідження МНР. Зазначено базові характеристики систем для інтелектуального аналізу текстів та визначено проблематику пов’язану із їх розробкою. Також зазначено кілька відомих систем для «text mining» з існуючих на сьогодні.

У **першому розділі** магістерської роботи «**Моделі і методи в «Text mining**» проведено огляд основних понять з даної теми та визначено роль технології інтелектуального аналізу текстів у сучасному інформаційному суспільстві. Проведено просте порівняння технологій «text mining» та «data mining» і зазначено вплив, який вони надали одна одній.

Розглянуто існуючі на сьогодні методи, які активно використовуються даною технологією, такі як: індексування документів, кластеризації та факторного аналізу. Наведено приклад простого алгоритму для аналізу тексту за даною технологією для більшого розуміння її можливостей. Проаналізовано процес виявлення критеріїв для оцінки і обчислення релевантності документа запиту, на прикладі гілки інтелектуального аналізу тексту, яка використовується в сучасних пошукових системах.

Визначено проблематику створення подібних систем та проблеми які вони мають вирішувати.

У **другому розділі** магістерської роботи **«Математичні моделі, методи і алгоритми для розпізнавання текстів природною мовою»** проведено аналіз загальних методів, які використовуються в інтелектуальному аналізі текстів, як:

- класифікація;
- кластеризація;
- витяг понять;
- питання-відповідь;
- тематичне індексування;
- пошук за ключовими словами.

Розглянуто різні підходи до представлення лінгвістичних даних у системах «text mining» та наведено їх переваги та недоліки. На ряду із цим наведено характеристики різних заснованих на таких представленнях моделей даних.

Також проведено аналіз різноманітних архітектур інструментальних систем для роботи з текстами написаних природною мовою, розглянуто особливості

компонентної організації таких систем та існуючі на сьогодні види процесів обробки текстів.

У **третьому розділі** магістерської роботи **«Аналіз інформаційних технологій програмного забезпечення для аналізу текстової інформації та вибір базового ПЗ»** розглянуто велику кількість існуючих і активно використаних систем та програмних додатків для інтелектуального аналізу текстових даних. Представлені системи відносяться до різних гілок у «text mining» як: системи на базі розмітки, системи на базі анотацій, системи інтеграції поверхневої і глибинної обробки та системи розвиваючі окремі аспекти обробки тексту. Для кожної розглянутої системи наведено перелік переваг та недоліків та особливостей роботи цієї системи. Як висновок до розділу проведено аналіз характеристик, які повинна мати система з допомогою якої буде продемонстровано можливості розглянутої технології та обрано базове програмне забезпечення для роботи у 4 розділі.

У **четвертому розділі «Апробація обраного ПЗ та результатів інтелектуального аналізу текстів»** на прикладі обраної раніше прикладної системи для аналізу текстів GATE продемонстровано процес та особливості роботи під час інтелектуального аналізу науково-технічної публікації. Також зазначені результати такої обробки з вказанням виконаної під час обробки роботу.

У **п’ятому розділі «Охорона праці та безпека у надзвичайних ситуаціях»** проведений аналіз факторів виробничого середовища у приміщенні на підприємстві «Lexico», а також визначений вплив цих факторів на здоров’я та працездатність працівників. Слід зазначити, що була встановлена відповідність всіх розглянутих показників чинним санітарним нормам та виявлено, що умови праці в «Lexico» є оптимальними.

У **шостому розділі «Методичні вказівки до практичних робіт»** було підготовлено методичні вказівки до самостійної роботи для дисципліни «Лінгвістичне забезпечення інтелектуальних систем».

ЗАГАЛЬНІ ВИСНОВКИ

В результаті виконання дипломної роботи:

1. Виконано ознайомлення з основними поняттями та термінами у сфері інтелектуального аналізу тексту та виявлено місце і задачі цієї сфери у сучасному інформаційному суспільстві.

2. Досліджено методи які використовуються для аналізу текстових даних і їх взаємодія у системах «text mining».

3. Виявлено проблематику пов’язану із розробкою систем для роботи з текстами написаних природною мовою.

4. Виконати аналіз математичних моделей, методів та алгоритмів використовуваних під час розробки програмних застосунків, та зв’ясувати, які переваги надають ті, чи інші методи та алгоритми;

5. Розглянуто різні підходи до представлення лінгвістичних даних у системах «text mining» та наведено їх переваги та недоліки. Наведено характеристики різних заснованих на таких представленнях моделей даних.

6. Розглянуто існуючі на сьогодні системи інтелектуального аналізу текстів та проведено порівняльний аналіз таких систем;

7. Провести тестування обраної системи на прикладі декількох науково-технічних публікацій для демонстрації результатів роботи такої системи.

8. У спеціальному розділі з охорони праці та безпеки у надзвичайних ситуаціях проаналізовано систему заходів і засобів по запобіганню впливу на людину несприятливих факторів, які супроводжують роботу працівника ІТ-сфери. Виконано аналіз освітлення та мікрокліматичних умов на робочому місці, управління цивільним захистом на підприємстві у разі виникнення пожежі.

АНОТАЦІЯ

до магістерської наукової роботи

на тему: «**Інформаційно-пошукова система аналізу науково-технічних публікацій та їх інтелектуальне представлення**»

Студент: Яцуненко Андрій Андрійович

Керівник: д.т.н., професор Фісун Микола Тихонович

Дана магістерська наукова робота присвячена дослідженню сучасного стану технології автоматичної інтелектуальної обробки текстових даних написаних природною мовою та на основі такого дослідження пошуку зручного та ефективного програмного застосунку для демонстрації користувацьких можливостей технології «text mining».

Метою даної дипломної роботи є виявлення найбільш зручного та ефективного програмного додатку у сфері інтелектуального аналізу ПМ текстів.

Об’єктом дослідження дипломної роботи є процеси інтелектуального аналізу даних текстів написаних природною мовою.

Предметом дослідження є зручний та ефективний програмний застосунок для демонстрації широких та корисних можливостей технології «text mining» у сучасному інформаційному суспільстві.

Фахова частина магістерської наукової роботи складається з наступних розділів: Моделі і методи в «Text mining», Математичні моделі, методи і алгоритми для розпізнавання текстів природною мовою, Аналіз інформаційних технологій програмного забезпечення для аналізу текстової інформації та вибір базового ПЗ, Апробація обраного ПЗ та результатів інтелектуального аналізу текстів.

Задачі, які були виконані в процесі роботи:

– виконано аналіз досліджень в галузі «text mining» та розглянуто проблеми, які зустрічаються у цій сфері;

– виконано аналіз математичних моделей, методів та алгоритмів використовуваних під час розробки програмних застосунків, та зв’ясувано, які переваги надають ті, чи інші методи та алгоритми;

– розглянуто існуючі на сьогодні системи інтелектуального аналізу текстів та проведено їх порівняльний аналіз;

– проведено тестування обраної системи з розглянутих на прикладі декількох науково-технічних публікацій для продемонстровано результати роботи даної системи.

В спеціальній частині магістерської наукової роботи з «Охорони праці та безпеки у надзвичайних ситуаціях» проведено аналіз умов праці на робочих місцях в офісному приміщенні компанії «Lexico» за факторів виробничого приміщення, проаналізовано питання, що пов’язані з якістю освітлення в офісі та циркуляції повітря, розроблено інструкції дій працівників компанії при виникненні пожежі.

У методичній частині підготовлено методичні вказівки до самостійної роботи для дисципліни «Лінгвістичне забезпечення інтелектуальних систем».

Робота складається з ___ сторінок, ___ рисунків, ___ таблиць та ___ посилань на літературні джерела.

Ключові слова: *text, mining, аналіз, класифікація, індексація, GATE.*

ABSTRACT

to the master's scientific work
on the topic: «**Information retrieval system for the analysis of scientific and technical publications and it's intellectual presentation**»

Student: Yatsunenko Andrii
Head: doctor of technical sciences, Professor Fisun Nikolai

The given master's scientific work is devoted to research of the current state of the technology of automatic intellectual processing of text data written in natural language and based on such research to find for convenient and effective software application for demonstration of user's capabilities of technology "text mining".

The purpose of this thesis is to identify the most convenient and effective software application in the field of intellectual analysis of PM texts.

The object of study of the thesis is the processes of intellectual analysis of data texts written in natural language.

The subject of the study is a convenient and effective software application for demonstrating the broad and useful potential of technology "text mining" in the modern information society.

The specialty of the master's scientific work consists of the following sections: Models and methods in "Text mining", Mathematical models, methods and algorithms for text recognition in natural language, Analysis of information technology of software for analysis of text information and choice of basic software, Testing of selected software and results of intellectual analysis of texts.

Tasks that were completed during the process:

- an analysis of researches in the field of "text mining" was carried out and the problems encountered in this field were considered;
- an analysis of the mathematical models, methods and algorithms used in the development of software applications is carried out, and it is determined what benefits are provided by those or other methods and algorithms;
- reviewed existing systems of intellectual analysis of texts and conducted their comparative analysis;
- tested the chosen system from the examples of several scientific and technical publications considered for the results of this system.

In the special part of the master's degree work on "Occupational Safety and Security in Emergencies", an analysis of the working conditions at the work places in the office space of the company "Lexico" was conducted on the factors of the production premises, analyzed the issues related to the quality of lighting in the office and circulation of air, instructions for action of the company's employees in the event of a fire have been developed.

In the methodical part, the methodical instructions for independent work for the discipline "Linguistic provision of intellectual systems" were prepared.

The work consists of ___ pages, ___ drawings, ___ tables and ___ references to literary sources.

Keywords: *text, mining, analysis, classification, indexing, GATE.*