

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ
ФАКУЛЬТЕТ КОМП'ЮТЕРНО-ІНФОРМАЦІЙНИХ СИСТЕМ І ПРОГРАМНОЇ
ІНЖЕНЕРІЇ

МИХАЛЮК ВАСИЛЬ АНДРІЙОВИЧ

УДК 004.42

**АЛГОРИТМІЧНЕ ТА ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ
ПОБУДОВИ КОРПУСІВ МОВ САТ-ЗАСОБІВ**

123 «Комп'ютерна інженерія»

Автореферат

дипломної роботи на здобуття освітнього ступеня «магістр»

Тернопіль

Роботу виконано на кафедрі комп'ютерних систем та мереж Тернопільського національного технічного університету імені Івана Пулюя Міністерства освіти і науки України

Керівник роботи: кандидат технічних наук, доцент кафедри комп'ютерних систем та мереж
Луцків Андрій Мирославович,
Тернопільський національний технічний університет
імені Івана Пулюя,

Рецензент: кандидат фізико-математичних наук, доцент кафедри фізики
Скоренький Юрій Любомирович,
Тернопільський національний технічний університет
імені Івана Пулюя,

Захист відбудеться 27 грудня 2018 р. о 9.00 годині на засіданні екзаменаційної комісії №34 у Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56, навчальний корпус №1, ауд.1-603

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність дослідження. Оскільки з'являються нові галузі, виникають нові терміни, мова певних предметних областей постійно поповнюється новими термінами та зворотами, доцільно постійно поновлювати корпуси мов систем автоматизованого перекладу. Фактично це є задача наповнення словникової бази парами відповідників на основі семантичних зв'язків. Схожі задачі протягом останніх років успішно розв'язуються у сфері індексації пошукових запитів пошуковими машинами. Тому, важливою задачею, яку необхідно розв'язати є адаптація існуючих механізмів індексації веб-вмісту для побудови корпусів мов, які охоплюють конкретні предметні області знань. Особливої ваги має побудова корпусів української мови.

На даний момент розробки в цій сфері ведуть низка англомовних країн. Це пов'язано, по-перше, з простотою роботи з англійською, оскільки в ній відсутні складні правила відмінювання іменників, прикметників, числівників, займенників дієвідмінювання, утворення різних форм дієслів тощо. По-друге, англійська є найбільш поширеною мовою серед розвинутих країн, тобто тієї частини населення, що володіє доступом до комп'ютерних технологій та фінансових можливостей їхнього розвитку. Відомими проектами є Американський Національний Корпус, Британський Національний Корпус, Корпус Брауна, Австралійський Корпус Англійської, Колхапурський Корпус Індійської Англійської, Веллінгтонський Корпус писемної новозеландської англійської. Серед неангломовних країн найбільшими корпусами є Угорський Національний Корпус та Корпус Чеської Мови, кожен з яких містить більш ніж 180 тис. слів. Для порівняння, Корпус української мови бібліотеки Чтиво містить близько 40 тис. слів.

Важливою науково-практичною задачею є реалізація програмного забезпечення створення та опрацювання корпусів мов. Реалізація цієї задачі передбачає вибір та обґрунтування математичного та алгоритмічного забезпечення за критеріями ефективності пам'яті та часу. Відповідний математичний та алгоритмічний апарат належить до предметної області прикладної лінгвістики, яка протягом останніх років набула широкого використання в різних сферах інформаційних технологій та сферах життєдіяльності загалом. Водночас створення якісного програмного забезпечення передбачає його верифікацію шляхом використання загальнодоступних даних.

Доцільною є задача створення обширного, збалансованого, релевантного, а також вільнодоступного корпусу української мови, оскільки серед наявних корпусів існують переважно аматорські проекти цифрових бібліотек чи закриті розробки філологічних інститутів.

Створення корпусу української мови та інструментарію для роботи з ним дасть змогу підвищити швидкодію САТ-систем, а отже покращити ситуацію з перекладами іншомовних книг, видань та передач; локалізацією програм, веб-сторінок та відеоігор. Також це дозволить поглибити лінгвістичні знання про українську мову, що матиме наслідком покращення викладання, розвитку та розуміння української мови в Україні та світі.

Мета і завдання дослідження. Метою є дослідження математичного, алгоритмічного і програмного забезпечення побудови корпусів мов САТ-засобів.

Досягнення цієї мети вимагає розв'язання таких завдань:

- Аналіз сучасного стану корпусів мов. Виокремлення ключових тенденцій сучасної корпусної лінгвістики
- Формулювання основних алгоритмів (схем) побудови корпусів мов
- Вибір та обґрунтування алгоритмічного та математичного забезпечення систем формування корпусів мов.
- Відбір текстових матеріалів для дослідження.
- Реалізація прототипу системи побудови корпусу мови.

Об'єкт дослідження — процес формування корпусів мов.

Предметом дослідження є корпусна лінгвістика, латентно-семантичний аналіз, алгоритми опрацювання текстових даних, засоби функційного програмування.

Наукова новизна отриманих результатів:

- Обґрунтований вибір методу латентно-семантичного аналізу для дослідження релевантних зв'язків термів і документів на прикладі корпусу, що складається із повного набору статей англomовної Вікіпедії.
- Проаналізовано й показано доцільність використання статистичного показника TF-IDF, парадигми функційного програмування, мови програмування Scala та середовища розробки Spark для формування і опрацювання корпусів мов.

Практичне значення отриманих результатів. Досліджені методи аналізу корпусів мов спрощують процес пошуку релевантної інформації у корпусах великих розмірів та можуть бути застосовані у САТ-системах.

Апробація результатів дипломної роботи. Результати дипломної роботи магістра апробовано на міжнародних конференціях:

- VI науково-технічній конференції «Інформаційні моделі, системи та технології».
- Науково-практичній конференції «АКТУАЛЬНІ ПИТАННЯ В КОНТЕКСТІ РОЗВИТКУ СУЧАСНИХ НАУК».

Структура роботи. Робота складається з розрахунково-пояснювальної записки та графічної частини. Розрахунково-пояснювальна записка складається з вступу, 6 частин, висновків, переліку посилань та додатків. Обсяг роботи: розрахунково-пояснювальна записка — 105 арк. формату А4, графічна частина — 10 аркушів формату А1.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність й важливість цього дослідження, проведено огляд проблематики у галузях перекладів та корпусного мовознавства. Охарактеризовані основні завдання, які необхідно вирішити у дипломній роботі магістра.

У першому розділі роботи «Аналіз предметної області застосування корпусів мов у САТ» було проведено аналіз предметної області, розглянуто що таке САТ-системи, які спеціалізовані та загальні технології до них входять. Було дано означення корпусам мов, досліджено історію їх виникнення та розвитку, розібрано види корпусів та загальні вимоги до них.

У другому розділі «Методологія побудови та обробки корпусів мов» проведено аналіз предметної області прикладної лінгвістики, розглянуто метод латентно-семантичного аналізу за допомогою розкладу матриці «терм-на-документ» алгоритмом SVD-розкладання. Розглянуто показник TF-IDF, що визначає ваги термів у корпусі. Обґрунтовано використання функційного програмування для досягнення поставленої мети, застосування мови Scala та середовища Spark.

У третьому розділі «Особливості побудови і роботи системи та алгоритмів для обробки корпусів мов» було описано створення матриці «терм-на-документ», застосування методу латентно-семантичного аналізу (LSA) для аналізу корпусу, що складається із повного набору статей англomовної Вікіпедії. Було розглянуто використання Spark для попередньої обробки даних: читання, очищення та приведення в числову форму. Проведено вирахування вагових коефіцієнтів за допомогою показника TF-IDF. Було показано, як вираховується сингулярний розклад матриці (SVD), пояснено, як його інтерпретувати і використовувати отримувані матриці. Показано варіанти запитів простого пошуку по вихідній матриці, та щодо релевантності терма до терма, документа до документа, терма до документа, та багатьох термів одночасно.

У четвертому розділі «Обґрунтування економічної ефективності» зроблено обчислення показників економічної ефективності від застосування методу вибору оптимального рішення при реалізації програмних проєктів, що дало можливість зробити висновок про доцільність проведення НДР.

У п'ятому розділі роботи «Охорона праці та безпека в надзвичайних ситуаціях» розглянуто вимоги до охорони праці користувачів ЕОМ при виконанні даної науково-дослідної роботи.

У даному розділі також розкриті питання безпеки в надзвичайних ситуаціях, а саме:

— Розглянуто питання шкідливих та небезпечних факторів при використанні комп'ютерних систем та захист від них користувачів. Класифіковано шкідливі та небезпечні чинники, які впливають на користувачів. Розглянуті професійні хвороби користувачів та методи їх передбачення. Виділено основні фактори, що впливають на функціональний стан користувачів комп'ютерів та наведено шляхи мінімізації їх негативного впливу.

— Досліджено механізм проведення планування заходів цивільного захисту на об'єкті у воєнний стан з метою підготовки об'єкту до реагування на їх можливі наслідки.

У шостому розділ роботи «Екологія» Досліджено питання антропогенного забруднення навколишнього середовища, розглянуті основні джерела антропогенного забруднення в Україні та світі, а саме: промисловість, сільське господарство, транспорт, побутова сфера. Також проаналізовано основні інструментів для оцінки стану навколишнього середовища, а саме екологічні показники. До уваги бралися стандарти що мають чинність в Україні та Східній Європі.

ВИСНОВКИ

У даній дипломній роботі магістра проведено дослідження алгоритмічного та програмного забезпечення побудови корпусів мов САТ-засобів. Основні висновки і результати проведених емпіричних та теоретичних досліджень такі:

— Було проведено аналіз предметної області, розглянуто що таке САТ-системи, які спеціалізовані та загальні технології до них входять.

— Проаналізовано означення корпусів мов, досліджено історію їх виникнення та розвитку, розібрано види корпусів та загальні вимоги до них. Це дало змогу побачити перспективні напрямки розвитку галузі.

— Проведено аналіз предметної області прикладної лінгвістики, розглянуто метод латентно-семантичного аналізу за допомогою розкладу матриці «терм-на-документ» алгоритмом SVD-розкладання. Розглянуто показник TF-IDF, що визначає ваги термів у корпусі. Обґрунтовано використання функційного програмування для досягнення поставленої мети, застосування мови Scala та середовища Spark.

— Було описано створення матриці «терм-на-документ», застосування методу латентно-семантичного аналізу (LSA) для аналізу корпусу, що складається із повного набору статей англomовної Вікіпедії.

— Розглянуто використання Spark для попередньої обробки даних: читання, очищення та приведення в числову форму.

— Показано, як вираховується сингулярний розклад матриці (SVD), пояснено, як його інтерпретувати і використовувати.

— Проведено вирахування вагових коефіцієнтів за допомогою показника TF-IDF.

— Показано варіанти запитів простого пошуку по вихідній матриці, та щодо релевантності терма до терма, документа до документа, терма до документа, та багатьох термів одночасно.

— Зроблено висновок про економічну доцільність проведення НДР у відповідному розділі.

— Розглянуто системи правових, соціально-економічних, організаційно-технічних, санітарно-гігієнічних і лікувально-профілактичних заходів та засобів, спрямованих на збереження життя, здоров'я і працездатності людини під час трудової діяльності у відповідному розділі.

— Досліджено механізм проведення планування заходів цивільного захисту на об'єкті у воєнний стан з метою підготовки об'єкту до реагування на їх можливі наслідки у відповідному розділі.

— Досліджено статистику екологічних показників та питання антропогенних чинників глобального забруднення довкілля у розділі «Екологія».

СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

1. Михалюк В. Алгоритмічне та програмне забезпечення побудови корпусів мов САТ-засобів. [Електронний ресурс] / А. Луцків, В. Михалюк // Матеріали VI науково-технічної конференції „Інформаційні моделі, системи та технології“ – 2018. – Режим доступу до ресурсу: <http://elartu.tntu.edu.ua/handle/lib/25390>.

АНОТАЦІЯ

Михалюк В.А. Алгоритмічне та програмне забезпечення побудови корпусів мов САТ-засобів.

Дипломна робота магістра, 123 — Комп'ютерні системи та мережі. — Тернопільський національний технічний університет імені Івана Пулюя, Тернопіль, 2018.

В дипломній роботі магістра виконано дослідження алгоритмічного та програмного забезпечення побудови корпусів мов САТ-засобів.

В роботі розглянуто метод латентно-семантичного аналізу для дослідження релевантних зв'язків термів і документів. Розглянуто сфери застосування, компанії світового рівня що вже розглядають цю технологію.

Було описано створення матриці «терм-на-документ», застосування методу латентно-семантичного аналізу (LSA) для аналізу корпусу, що складається із повного набору статей англomовної Вікіпедії. Розглянуто використання Spark для попередньої обробки даних: читання, очищення та приведення в числову форму. Показано, як вираховується сингулярний розклад матриці (SVD), пояснено, як його інтерпретувати і використовувати. Проведено вирахування вагових коефіцієнтів за допомогою показника TF-IDF.

Показано варіанти запитів простого пошуку по вихідній матриці, та щодо релевантності терма до терма, документа до документа, терма до документа, та багатьох термів одночасно.

Ключові слова: САТ-засоби, корпус мови, латентно-семантичний аналіз, сингулярний розклад матриці, TF-IDF, матриця «терм-на-документ».

ANNOTATION

Mykhalik Vasyl. Algorithms and software of languages construction of SAT-tools.

Master's Thesis, 123 — Computer Systems and Networks. — Ternopil Ivan Puluj National Technical University, Ternopil, 2018.

In the master's thesis the research of algorithmic and software of building of buildings of languages of CAT-tools is executed.

The paper considers the method of latent semantic analysis to study the relevant relationships of terms and documents. Areas of application are considered, companies of the world level already considering this technology.

It was described the creation of the "term-document" matrix, the use of the latent semantic analysis (LSA) method for corpus analysis, consisting of a complete set of English-language Wikipedia articles. The use of Spark for preliminary processing of data is considered: reading, clearing and numerical retrieval. It shows how the Singular Value Decomposition (SVD) is calculated, explains how to interpret and use it. The weighting coefficients are calculated using the TF-IDF indicator.

It shows the options for simple searches for the original matrix, and the relevance of the term to the term, the document to the document, the term to the document, and multiple terms at the same time.

Key words: CAT-tool, language corpora, latent semantic analysis, singular value decomposition, TF-IDF, "term-document" matrix.