

Голові спеціалізованої вченої ради Д 08.084.01
49600, м. Дніпро, пр. Гагаріна, 4, НМетАУ

В І Д Г У К

офіційного опонента на дисертаційну роботу
Куроп'ятник Олени Сергіївни
«Конструктивно-продукційні моделі природомовних текстів
для виявлення запозичень у структурованих документах»,
подану на здобуття наукового ступеня кандидата технічних наук
за спеціальністю 01.05.02 – математичне моделювання
та обчислювальні методи

Дисертаційна робота, на яку дається відгук, має загальний обсяг 207 сторінок, в тому числі 161 сторінку основної текстової частини, 47 рисунків, 14 таблиць, 3 додатки на 12 сторінках та список використаних джерел із 128 найменувань на 16 сторінках.

Актуальність теми дисертації

Обробка інформації, представленої у текстовому вигляді, є затребуваною у різних сферах послуг та галузях виробництва. Серед задач такої обробки є автоматизоване анотування та реферування, перевірка орфографії та виправлення помилок, аудіо відтворення, інформаційний пошук, виявлення дублікатів тощо. Однієї з важливих задач сьогодення, зокрема в академічному середовищі, є виявлення плагіату та запозичень.

Розв'язанні задачі виявлення запозичень потребує розробки математичних моделей та методів представлення і зіставлення текстів. Її розв'язанню присвячено чимало робіт вітчизняної та закордонної наукової спільноти, проте і досі залишаються актуальними питання автоматизації процесів виявлення запозичень, особливо в умовах перманентного росту і розвитку інформаційного середовища та наявності маскування – спотворень текстів з метою приховування запозичень. В даному контексті також важливими є врахування структури документів за змістом та попередня обробка текстів. Процес виявлення запозичень також може бути ускладнений

перефразуванням та синонімією, а тому просте зіставлення лексики текстів, їх речень може бути не ефективним – необхідним є семантичний аналіз.

Саме розв'язанню такої актуальної технічної науково-практичної задачі виявлення запозичень в умовах маскування з урахуванням структури документів і присвячена дисертаційна робота Куроп'ятник О. С., подана на здобуття наукового ступеня кандидата технічних наук.

Структура, задачі та короткий зміст роботи.

Робота має «класичну» структуру, складається зі вступу, чотирьох розділів, висновків, списку літератури та додатків.

У вступі обґрунтовано актуальність роботи, сформульовано її мету та задачі, визначено об'єкт і предмет досліджень, наукову новизну й практичну цінність отриманих результатів, методи досліджень. А також зазначені внесок здобувача та основні публікації.

У першому розділі виконано огляд та аналіз відомих мовних моделей, методів і засобів обробки текстів для виявлення запозичень. Особливу увагу приділено програмній реалізації цих засобів: розглянуто майже три десятки різних програмних додатків, визначено їх основні функціональні характеристики, що безумовно є корисним з огляду на те, що автор виконує роботу не лише у науковій, а й у практичній площині. В даному розділі сформовано основні цілі та задачі дослідження.

Другий розділ є теоретичною частиною роботи. Він присвячений розробці моделей та методів представлення та обробки мови та мовних конструкцій, зокрема текстів. Оскільки автор розглядає запозичення не лише на рівні лексики, а й семантики, то моделювання починається від основ – принципів формування змісту. Для цього використовується поняття образу та розроблена автором конструктивно-продукційна образна модель мови. Для автоматизованої роботи з текстом запропонована модель його графового представлення, перевагою якого є групування ланцюжків за спільними частинами. Розроблений метод стиснення графового представлення, оснований на відомих операціях гомеоморфізму, покликаний на зменшення витрат пам'яті для збереження графів. Також у розділі запропоновано метод порівняння структурованих документів, тексти яких після обробки мають графове представлення. Особливістю формалізації даного методу є його конструювання за допомогою моделі – конструктору – з елементів носія, якими є алгоритми.

Основним математичний апаратом, використаним у даному розділі, є конструктори та операції їх перетворення, що базується на такому виді математичних числень як формальні граматики.

У третьому розділі представлено ряд експериментів для перевірки можливості застосування розроблених моделей та методів для вирішення поставлених задач. Запропоновано оригінальний підхід для формування тестових даних для перевірки здатності демаскувати запозичена. Він полягає у розробці конструктивної системи, що продукує сценарії маскувань та застосовує їх до документів. Також у даному розділі цінним є те, що за результатами експерименту встановлено перешкоди у використанні аналогів для експериментальних досліджень функціональної ефективності розроблених програмних засобів. Дана оцінка отриманим результатам, проаналізовано причини отримання саме таких значень.

Четвертий розділ напрацьовано з метою впровадження розроблених моделей та методів. Він присвячений розробці комп'ютерних реалізацій запропонованих моделей та методів. В даному розділі встановлено зв'язок між компонентами конструктивно-продукційних моделей та об'єктно-орієнтованими. Таким чином, розроблені теоретичні засади виявлення запозичень доведено до практичної реалізації.

Наведені у висновках результати повною мірою відображають суть дисертаційної роботи та мають практичне втілення.

Основні наукові результати досліджень і новизна дисертації

Вперше отримано такі результати:

- виконано формалізацію процесів формування образів людини засобами об'єктно-орієнтованого моделювання та розроблено конструктивно-продукційну модель природної мови на основі образного представлення дійсності.
- розроблено конструктивно-продукційну модель графового представлення текстів та метод їх порівняння для виявлення запозичення з урахування зміни порядку текстових складових та структурних особливостей документів, яким вони належать;
- побудовано конструктивно-продукційну модель процесів маскувань запозичень для автоматизації перевірки здатності демаскування запозичень у текстах програмами-антиплагіатами.

Отримали подальший розвиток:

- методи та засоби конструктивно–продукційного моделювання: визначено зв'язок конструктивних моделей з об'єктно-орієнтованими, представленими засобами UML, який покладено в основу комп'ютерних реалізацій запропонованих моделей представлення текстів та процесів маскування;
- методи виявлення семантичних запозичень у текстах;
- методи обробки графів для їх стиснення.

Практичне значення та рекомендації щодо використання результатів дисертації

Практичне значення полягає у доведенні теоретичних розробок до втілення у реальних програмних засобах для виявлення запозичень у структурованих документах та маскування запозичень з метою підготовки тестової бази для програм анти плагіату.

Практичне значення підтверджується рядом актів та охоронних документів, отриманих автором:

- акти впровадження результатів дисертаційної роботи та прийняття у дослідну експлуатацію програмного засобу «Система виявлення запозичень у цифровому представленні структурованих документів» в Дніпровському національному університеті ім. академіка В. Лазаряна;
- акт впровадження результатів дисертаційної роботи в філії «Проектно-конструкторське бюро інформаційних технологій» АТ «Українська залізниця» при проектуванні та розробці алгоритмів та програмних засобів порівняння та видалення тотожних частин документів;
- акт впровадження результатів дисертаційної роботи в ТОВ «СОВЛАНУТ» при виконанні проекту «Розробка системи виявлення запозичень»;
- свідоцтвом про реєстрацію авторського права на твір № 68137 від 05.10.2016;
- свідоцтвом про реєстрацію авторського права на твір № 87131 від 22.03.2019.

Результати дисертаційної роботи доцільно застосовувати при розробці та модернізації систем автоматизованого анотування, перекладу, антиплагіату

тощо, а також у також при вирішенні споріднених задач обробки природомовних текстів.

Обґрунтованість та достовірність одержаних результатів

Дисертаційна робота має достатнє обґрунтування основних наукових положень, які захищаються. Слід відзначити коректне застосування автором методів теорії графів, теорії множин, формальних граматики, статистичних методів. Достовірність одержаних результатів підтверджується зіставленням сучасних наукових і технічних досягнень в області моделювання мов і розробки алгоритмів та систем виявлення запозичень; проведення експериментів у програмно-апаратному середовищі; апробацією основних теоретичних і експериментальних результатів роботи в друкованих працях та доповідях на конференціях.

Оформлення дисертації та автореферату

Дисертаційна робота викладена сучасною науковою технічною українською мовою. Науково-технічна термінологія використовується коректно. Структура та обсяг дисертації відповідають встановленим вимогам на здобуття наукового ступеню кандидата технічних наук.

У роботі містяться посилання на наукові розробки та здобутки вітчизняних та закордонних вчених і є актуальними в розрізі сучасної наукової картини у відповідній предметній області. Робота має достатню кількість ілюстративних матеріалів.

Матеріал викладено послідовно та логічно.

Зміст автореферату повністю відповідає основним положенням та висновкам, зробленим у дисертації. Дисертаційна робота відповідає паспорту спеціальності 01.05.02 – математичне моделювання та обчислювальні методи.

Повнота викладення наукових положень, висновків і рекомендацій в опублікованих працях

За результатами дисертаційного дослідження опубліковано 24 наукових праці, в тому числі 7 статей у журналах, рекомендованих МОН України для публікації результатів дисертацій, та закордонних виданнях; у тезах доповідей та трудах міжнародних та всеукраїнських конференцій – 15. Отримано 2

свідоцтва про реєстрацію авторського права на твір. У вступі до дисертації та авторефераті чітко визначено особистий внесок здобувача.

Зауваження

Представлена робота присвячена вирішенню актуальної задачі, має чимало цікавих та оригінальних рішень, що є достатньою мірою обґрунтованими. Проте до неї є зауваження:

1. У п. 1.2.4. розглянуто моделювання не писемної мови. Доречним було б привести обґрунтування розгляду моделей такого виду мови.
2. При розгляді алгоритмів та методів порівняння тексту (п. 1.3.2) доречним було б вказати для виявлення плагіату яких видів вони застосовуються. Аналогічним є зауваження до п. 1.4, де не вказано, які методи використовуються у розглянутих програмних засобах.
3. Доцільним було б проведення експериментів з визначення здатності демаскування запозичень аналогічними програмами-антиплагіатами. При цьому при підготовці тестової бази використати розроблену сценарну модель маскування запозичень.
4. У описі схеми комплексу, який наведено у висновках, доречними були б посилання на відповідні пункти рукопису дисертації для кожного компоненту.

Наведені зауваження мають окремий характер, не знижують високий науковий рівень дисертаційної роботи і не впливають на її загальну позитивну оцінку.

Висновок

Вивчення дисертаційної роботи Куроп'ятник Олени Сергіївни «Конструктивно-продукційні моделі природомовних текстів для виявлення запозичень у структурованих документах», автореферату, ознайомлення з опублікованими науковими працями та враховуючи всі зауваження та недоліки можна зробити такий висновок:

- дисертаційні робота є завершеною оригінальною науковою працею та відповідає основним положенням паспорту спеціальності 01.05.02 – математичне моделювання та обчислювальні методи (технічні науки);
- зміст автореферату відповідає основним положенням дисертаційної роботи;

- робота є актуальною у сучасному науковому світі, присвячена вирішенню технічної задачі обробки інформації, представленої у текстовому вигляді;
- отримані результати мають як наукову, так і практичну цінність;
- робота відзначається використанням єдиного математичного апарату для моделювання усіх об'єктів та процесів, що дозволило створити єдиний комплекс моделей та засобів для вирішення поставлених задач. Також варто звернути увагу на оригінальність підходу до підготовки експерименту щодо маскування запозичень.

В ході особистої бесіди зі здобувачем встановлене повне і вільне володіння суттю роботи, гарна орієнтація у предметній області, розуміння використаних теорій та методів.

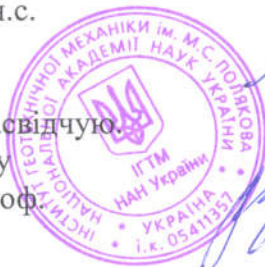
Вважаю, що дисертаційна робота є завершеною науково-дослідницькою роботою та відповідає вимогам п. п. 9, 11, 12 «Порядку присудження наукових ступенів», які висуваються до кандидатських дисертацій, а її автор, Куроп'ятник Олена Сергіївна, заслуговує на присудження наукового ступеня кандидата технічних наук за спеціальністю 01.05.02 – математичне моделювання та обчислювальні методи.

Офіційний опонент
старший науковий співробітник
відділу геомеханічних основ
технологій відкритої розробки родовищ
Інституту геотехнічної механіки
ім. М. С. Полякова НАН України,
доктор технічних наук, с.н.с.


Р. В. Кірія

Підпис проф. Кірії Р. В. засвідчую.
Учений секретар інституту
доктор технічних наук, проф.


В. Г. Шевченко



Відгук надійшов
у Раду: 19.06.2020
Вчений секретар: 