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

ВІДГУК

офіційного опонента завідувача кафедри інтелектуальних комп'ютерних систем НТУ «ХПІ», д-ра технічних наук, професора Шаронової Наталії Валеріївни 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

Актуальність теми дисертації. Дисертаційна робота Куроп'ятник Олени Сергіївни «Конструктивно-продукційні моделі природомовних текстів для виявлення запозичень у структурованих документах» присвячена розв'язку актуальної науково-прикладної задачі автоматизованого виявлення плагіату, а в більш широкому значенні різного роду запозичень, що стає все більше затребуваним на всіх академічних рівнях в Україні.

Актуальність теми роботи визначається необхідністю істотної перебудови існуючих моделей та систем управління навчальним процесом у сучасний період, що потребує академічної доброчесності усіх учасників навчального процесу. Зазначені процеси висувають нові вимоги щодо розроблення теоретичних та методологічних основ створення та практичної реалізації відповідних інформаційних систем. Авторка вважає, що часткове вирішення задачі виявлення запозичень можливе за допомогою впровадження відповідних інформаційно-технічних та технологічних засобів.

Ретельний аналіз досліджуваної області показав, що існує велика кількість технічних засобів: web-сервісів та Windows-додатків для виявлення запозичень та плагіату. Вони різняться наявністю або відсутністю орієнтації на академічне середовище та підтримкою мов документів, що перевіряються. Автором здійснено глибинний аналіз стану питання, як з законодавчого погляду, так і з погляду існуючих теоретичних та практичних напрацювань за напрямом тематики дослідження. Слід зазначити, що, відповідно до аналізу, який здійснила авторка, багато дослідників зазначають високу ефективність використання інформаційних технологій, які базуються на дослідженнях інтелектуальних функцій людини та побудови їх математичних моделей в області семантичного опрацювання текстової інформації, що знайшло відображення в працях провідних вчених, на роботи яких вона спирається.

У роботі авторка дійшла висновку, що існуючі моделі, методи та алгоритми застосовуються розрізнено і, переважно, не мають академічної спрямованості, та не враховують структурні особливості документів при доборі матеріалів та їх зіставленні. У роботі висунуте припущення, що

доцільним є розробка моделей, методів та засобів виявлення запозичень з урахуванням структурних особливостей текстових документів.

Саме чітко формалізована науково-прикладна проблема, яка у такій постановці раніше не розглядалась, дозволила авторці визначити мету і наукові завдання дослідження. Викладене вище й обумовлює актуальність теми дисертаційного дослідження Куроп'ятник О.С.

Робота виконана відповідно до Закону України № 2623-14 від 11.07.2001 «Про пріоритетні напрями розвитку науки і техніки», Закону України «Про вищу освіту» (документ 1556-VII, редакція від 09.08.2019). Дисертація є частиною науково-дослідних робіт, виконаних на кафедрі комп'ютерних інформаційних технологій Дніпровського національного університету залізничного транспорту ім. академіка В. Лазаряна, зокрема «Прикладне конструктивне моделювання програмних сутностей» (2016 р. № держреєстрації 0116U006841) та «Підвищення конкурентоспроможності залізничного транспорту на основі уніфікованих інтелектуальних технологій процесів перевезень та *експлуатації* парків технічних систем» (2017-2018 р. № держреєстрації 0117U004392), у яких автор брала участь як виконавець.

Достовірність та обґрунтованість основних висновків і результатів роботи. Обґрунтованість і достовірність наукових результатів, висновків і рекомендацій, викладених в дисертаційній роботі, досягаються ретельним системним аналізом процесу розроблення математичних моделей для виявлення запозичень у текстах природної мови; забезпеченням використання сучасних теорій, доведенням отриманих результатів на науково-технічних конференціях і семінарах, коректним використанням математичного апарату, програмним та апаратним моделюванням. Теоретичні дослідження базуються на фундаментальних положеннях. Достовірність нових, отриманих автором результатів, підтверджується комп'ютерним моделюванням, а також результатами, які відображені у документах впровадження. Отримані теоретичні результати узгоджені з відомими фактами.

Результати дослідження обумовлені коректним використанням сучасних методів і теорій, а саме були використані: відомі методи теорії графів; математичного моделювання та теорії множин, формальних граматик, статистичні методи. Розроблені авторкою конструктивно-продукційні моделі покладені в основу програмних засобів для виявлення запозичень у структурованих документах та формування тестів для перевірки здатності демаскування запозичень. Виконано дослідження часової та функціональної ефективності алгоритмів зіставлення документів, отримані результати зіставлено з аналогом. Вищенаведене дозволяє зробити висновок про достатню обґрунтованість та достовірність результатів дисертаційної роботи.

Наукова новизна результатів дисертації. Авторка провела наукове дослідження, яке базується на комплексному використанні сучасних методів

та інформаційних технологій і отримала значні наукові результати, серед яких найважливішими є:

Уперше:

1) виконано формалізацію процесів формування образів людини засобами об'єктно-орієнтованого моделювання та розроблено так звану конструктивно-продукційну модель природної мови на основі образного представлення дійсності. Модель відрізняється від існуючих можливістю опису операцій модифікації мови і врахуванням внутрішнього виконавця, що може бути використано для продукування конструкцій різної складності та виявлення семантичних запозичень;
2) розроблено конструктивно-продукційну модель графового представлення текстів та метод їх порівняння для виявлення запозичення з урахування зміни порядку текстових складових та структурних особливостей документів, до яких вони належать;
3) побудовано конструктивно-продукційну модель процесів маскування запозичень для автоматизації перевірки здатності демаскування запозичень у текстах так званими програмами-антиплагіатами.

Отримали подальший розвиток:

4) 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 моделі графового представлення тексту для виявлення запозичень у структурованих документах та моделі процесів маскування;
5) методи виявлення семантичних запозичень у текстах: інтерпретація семантичної складової текстів виконавцем моделі природної мови дозволяє зменшити вплив зміни лексичної та синтаксичної структури тексту на виявлення запозичень;
6) методи обробки графів: на основі операцій гомоморфізму та алгоритму групування вузлів за атрибутами розроблено метод стиснення графа з текстовим навантаженням.

Наукове і практичне значення роботи. Подана в роботі сукупність моделей, розробленої методології та алгоритмічного і програмного забезпечення утворюють комплекс моделей мови, мовних конструкцій та засоби їх обробки на єдиній теоретичній основі, що дозволяє вирішити науково-технічну задачу виявлення запозичень у структурованих документах кваліфікаційного характеру, що є важливим теоретичним внеском у наукову спеціальність 01.05.02 – математичне моделювання та обчислювальні методи. Робота відповідає таким положенням формули спеціальності: удосконалення методів і засобів математичного та комп'ютерного моделювання, призначених для використання при всебічному дослідженні або створення нових апаратно-програмних засобів моделювання й обчислення. Робота містить такі напрями дослідження, визначені паспортом спеціальності:

отримання принципово нових (нетрадиційних) видів математичних моделей; створення і дослідження нових обчислювальних методів і алгоритмів, що забезпечують створення ефективних програмних засобів комп'ютерної реалізації.

Практичне значення роботи полягає у тому, що створені теоретичні основи лягли в основу розроблених конструктивно-продукційних та об'єктно-орієнтованих моделей процесів маскування запозичень, що дозволило автоматизувати процес формування тестів для перевірки здатності демаскування запозичень у програмах-антиплагіатах.

Результати, які отримано у ході дисертаційного дослідження, знайшли практичне застосування у вигляді розробленого програмного забезпечення, яке дозволяє сприяти виконанню вимог Закону України «Про вищу освіту» щодо виявлення академічного плагіату у роботах здобувачів вищої освіти та впроваджене в Дніпровському національному університеті залізничного транспорту ім. академіка В. Лазаряна. Результати дисертаційному роботи впроваджено у програмних проектах філії ПКБ ІТ АТ «Укрзалізниця» та ТОВ «СОВЛАНУТ», що підтверджено відповідними документами.

Рекомендації щодо використання результатів дисертації. Враховуючи складність предметної області оброблення текстової інформації, зокрема на рівні семантики текстових конструкцій, запропоновані в роботі елементи інформаційної технології, які базуються на розроблених конструктивно-продукційних моделях, можуть бути використані у будь-яких інформаційних системах як елемент боротьби із запозиченнями та прямим плагіатом. Можлива область застосування представлених моделей охоплює NLP-компоненти роботів і додатків, в тому числі систем перекладу і антиплагіату, навчальних та експертних систем. Розроблена модель процесів маскування дає формалізоване подання відповідних процесів для подальшої автоматизації і дає змогу застосувати інструментарій для отримання бази тестів для програм, що протидіють плагіату.

Повнота викладення результатів дисертації в опублікованих працях. Результати дисертаційної роботи опубліковані в 24 наукових працях, в тому числі 7 статей у журналах, рекомендованих МОН України для публікації результатів дисертацій, та закордонних виданнях: «Проблеми програмування» (передрук у CEUR Workshop Proceedings – проіндексовано міжнародною науково-метричною базою (НМБ) Scopus), «Штучний інтелект», «Східно-Європейський журнал передових технологій» (НМБ Index Copernicus), «Наука та прогрес транспорту» (НМЬ Index Copernicus), «Acta Cybernetica» (Угорщина, НМБ Scopus, Web of Science) , «Information Technologies & Knowledge» (Болгарія), «Системні технології» (НМБ Index Copernicus); у тезах доповідей та трудах міжнародних та всеукраїнських конференцій – 15. Отримано 2 свідоцтва про реєстрацію авторського права на твір.

Слід відзначити якісне та повне висвітлення матеріалів дисертаційного дослідження у зазначених публікаціях та апробацію результатів на міжнародних конференціях.

Аналіз змісту дисертації, її завершеності й оформлення. Побудова дисертації відповідає прийнятим для наукового дослідження рекомендаціям. Дисертація складається із анотації двома мовами, змісту, переліку умовних позначень, вступу, чотирьох розділів, висновків, списку використаних джерел і трьох додатків.

У вступі обґрунтовано актуальність теми дослідження та наукових завдань; наведено інформацію про зв'язок роботи з науковими темами; сформульовано мету й завдання дослідження; розкрито наукову новизну, практичне значення отриманих результатів та особистий внесок здобувача; наведено відомості про апробацію, публікації та впровадження результатів дослідження.

У першому розділі здійснено аналіз проблем виявлення запозичень у структурованих документах, виконано огляд та аналіз існуючих сучасних підходів до моделювання різних аспектів мови, методів, алгоритмів та програмних засобів попередньої обробки та зіставлення текстів. Проаналізовані підходи до моделювання різних аспектів мови. Проведеній аналіз показав, що розглянуті питання моделювання мови і її конструкцій є міждисциплінарними і наразі мають два напрями вирішення: філологічний (гуманітарний) та технічно-прикладний, прагматичний. Розглянуто методи та алгоритми обробки текстових конструкцій для виявлення запозичень. Методи, засновані на простому порівнянні, є чутливими до механізмів маскування запозичень. Авторка дійшла висновку, що актуальною є задача не лише зіставлення, а й попередньої обробки, необхідність якої зумовлена великими обсягами інформації та можливістю використання технік маскування, а також аналізу отриманих результатів. Крім того, розглянуті моделі, методи та алгоритми не враховують структурні особливості оцифрованих текстів, що оброблюються. Серед розглянутих програмних засобів не виявлено таких, що враховують структуру документів, текст яких перевіряється на унікальність. Розділ завершується обґрунтуванням і формулюванням мети і завдань дослідження.

У другому розділі представлені об'єктно-орієнтовані моделі смислових конструкцій та конструктивно-продукційні моделі природної мови для моделювання смислових конструкцій як результату розумової діяльності людини визначено такі поняття, як: прообраз; прообраз особистого; смисл (інтенсіонал); семантика; сигніфікат; образ; атомарний образ; думка тощо. На основі даних понять та з використанням принципів об'єктно-орієнтованого моделювання із застосуванням UML була побудована модель образів світу, що характерна для сприйняття людиною.

Використовуючи апарат конструкторів, авторкою побудована модель мовної конструкції тексту, що відображає його лексичну та синтаксичну

складову. Для вирішення задачі зіставлення текстів побудована модель графового представлення тексту.

У третьому розділі роботи представлено експериментальну частину дослідження. Проведено аналіз та виявлення подібності мовних конструкцій та структурованих документів на основі конструктивної моделі природної мови. Здійснено моделювання текстових фрагментів однієї тематики, але з різною лексично-синтаксичною структурою та різними мовами. Для наочності виконано графове представлення текстів, проведено аналіз структури та взаємозв'язків виділених образів. Для структурованих документів визначено їх образне представлення, за яким виконано їх зіставлення. В рамках експерименту структура визначалася вимогами до документів, які є пояснювальними записками до дипломних проектів. Як висновок, зміна лексики та коригування структури має незначний вплив на образне представлення текстів та документів, а тому даний підхід може бути використаний для виявлення семантичних запозичень.

На наш погляд, найбільш важливим результатом дослідження було те, що для перевірки можливості застосування моделі графового представлення текстів для виявлення запозичень в умовах маскування розроблена модель процесів маскування, що передбачає побудову та застосування сценаріїв запозичень. Модель складається з двох конструкторів: сценаріїв та модифікації текстів. Проведено комп'ютерні експерименти для дослідження часової ефективності операцій побудови графового представлення та зіставлення текстів та структурованих документів. Виконано також комп'ютерний експеримент для визначення ступеню чутливості комп'ютерної реалізації моделі графового представлення тексту до впливу сценаріїв маскування запозичень.

Четвертий розділ роботи присвячений опису результатів розробки об'єктно-орієнтованих моделей графового представлення текстів, документів та її програмної реалізації для виявлення запозичень у текстах та структурованих текстових документах. Встановлено зв'язок алгоритмічної складової моделі графового представлення тексту та логіки програми.

Результати прикладної частини дослідження імплементовано у трьох програмних додатках. Перший додаток представляє автоматизоване формування тестів на основі конструктивно-продукційної моделі сценаріїв модифікації. Другий додаток дозволяє виконувати порівняння структурованих документів з базою та визначати відсоток запозичень за розділами та документом в цілому. Порівняння ведеться згідно з розробленим шаблоном, який представлено xml-файлом. Для розділів, що не були розпізнані за шаблоном, можна задати відповідність у ручному режимі. Третій додаток дозволяє порівнювати два документи з визначенням загального відсотку запозичень та за фрагментами зі збереженням результатів; фільтрувати результати порівняння: задавати мінімально допустиму довжину запозиченого фрагменту та відстань між ними, що спрощує структурний аналіз результату перевірки; порівняти декілька

текстових документів з декількома, визначивши загальний відсоток запозичень та їх структурний склад та ін.

У висновках стисло сформульовані основні наукові і практичні результати дисертаційної роботи.

При загальному позитивному ставленні до роботи, слід відмітити, що вона не вільна від зауважень.

Зауваження до дисертаційної роботи.

1. У першому розділі проведений аналіз існуючих моделей обробки текстів природної мови з різних позицій. Доцільним було б формування критеріїв, яким повинна відповідати модель для вирішення поставлених задач, та проведення аналізу за ними. Це б дало змогу повною мірою використати позитивний досвід моделювання в даній предметній області.
2. Експерименти визначення часової ефективності обробки структурованих документів проведені на невеликій тестовій базі. Доцільним є її збільшення для виявлення «вузьких» місць з точки зору технології використання розробленого ПЗ. Крім того, ці експерименти проведені на кваліфікаційних роботах лише однієї спеціальності, що не можна вважати репрезентативною вибіркою.
3. Не визначено, за якими семантичними та синтаксичними ознаками визначаються ключові слова, які потрапляють у шаблон. Не висвітлено питання перевірки за різними шаблонами.
4. Для альтернативних програм-антиплагіатів не досліджено вплив маскувань на якість виявлення запозичень. Доцільним було б порівняння впливу маскувань на розроблену програму та альтернативні.
5. У пункті 3.3 табл. 3.8. наводиться пояснення щодо подібності образів, виділених за текстовими фрагментами, але не визначено критерій подібності. А тому даний підхід потребує доопрацювання для подальшого застосування в автоматизованих системах.
6. У пункті 2.4.1 описується запропонований конструктор-породжувач природномовного тексту. Вводяться визначення понять к-слова, к-речення та інших конструкцій. З тексту рукопису не зрозуміло, яким чином дані поняття використовуються в процесі виявлення запозичень та який вплив вони мають при формуванні результатів перевірки документів програмним засобом, описаним у пункті 4.3.
7. В дисертації розроблено дві моделі для представлення мовних конструкцій: образна модель мови (семантика), та модель графового представлення тексту (синтаксис), проте не визначено зв'язок між ними. Доречним було б наведення прикладів представлення елементів певного лінгвістичного корпусу цими двома моделями із зазначенням особливостей застосування кожної моделі.
8. У тексті автореферату нерівномірно наведені описи розділів дисертаційної роботи: розділ другий займає майже 8 сторінок, тоді як

третій – 2 стор., четвертий – 1,5 стор. Не зроблено висновків наприкінці опису кожного розділу.

Загальний висновок по роботі. Вказані у відгуку недоліки не зменшують її теоретичної та практичної цінності. Подана дисертаційна робота виконана на високому кваліфікаційному рівні. Автореферат дисертації повністю відображає основні положення самої дисертації, містить усі необхідні для його оцінки фахівцями дані і відповідає вимогам щодо оформлення.

Вважаю, що дисертаційна робота Куроп'ятник Олени Сергіївни «Конструктивно-продукційні моделі природомовних текстів для виявлення запозичень у структурованих документах», є завершеним науковим дослідженням, містить важливі наукові та практичні результати. У дисертаційній роботі вирішено актуальну науково-практичну задачу розробки моделей обробки текстової інформації з метою формалізації пошуку запозичень на основі створення конструктивно-продукційних моделей пошуку запозичень у структурованих документах. Робота відповідає напрямам досліджень паспорта спеціальності 01.05.02 – математичне моделювання та обчислювальні методи.

Дисертаційна робота за актуальністю, науковою новизною та практичною цінністю відповідає вимогам пп. 9, 11, 12 «Порядку присудження наукових ступенів», затвердженого постановою Кабінету Міністрів України від 24 липня 2013 р. № 567, які висуваються до кандидатських дисертаційних робіт, а її автор, Куроп'ятник Олена Сергіївна, заслуговує присудження їй наукового ступеня кандидата технічних наук за спеціальністю 01.05.02 – математичне моделювання та обчислювальні методи.

Офіційний опонент, завідувач кафедри
Інтелектуальних комп'ютерних систем
Національного технічного університету
«Харківський політехнічний інститут»
Міністерства освіти і науки України,
доктор технічних наук, професор Н. В. Шаронова

Учений секретар НТУ «ХПІ»
доктор технічних наук, професор О. Ю. Заковоротний

Підпис професора Шаронової Н. В. засвідчую
"28" 05 2020 р.

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ * Національний технічний університет "Харківський політехнічний інститут" * код 02071180 *

Відгук надійшов
у Раду: 04.06.2020
Вчений секретар: