РЕЦЕНЗІЯ

на дисертаційну роботу
Дмитренка Олега Олександровича
на тему «Інформаційні технології формування та аналізу мережевих моделей предметних галузей на основі лінгвостатистичного підходу»,
представлену на здобуття ступеня доктора філософії
в галузі знань Інформаційні технології
за спеціальністю 122 «Комп'ютерні науки»

Актуальність теми дисертації.

Швидкий розвиток інформаційно-комунікаційних технологій та глобалізація інформаційного простору призвели до радикальних змін в обсязі та доступності інформаційних ресурсів в мережі Інтернет. З цими змінами виникає актуальна проблема інформаційного перевантаження, яке не тільки супроводжується припливом нових цінних знань, але й призводить до значного зростання обсягів неструктурованих даних, включаючи "інформаційне сміття" та дублікати.

Недостатність технологічних рішень та обмежена здатність існуючих систем обробляти величезні об'єми неструктурованих даних створюють критичну невідповідність між розвитком інформаційних систем і зростанням динамічних інформаційних потоків. Така невідповідність викликає необхідність розробки нових підходів та методів для ефективного пошуку, структуризації та аналізу неструктурованих текстових даних.

Важливим аспектом є процес концептуалізації та формалізації текстових даних у вигляді онтологічної моделі, що може значно покращити якість та точність обробки і аналізу. Розглянуті в роботі лінгвостатистичні методи формування мережевих моделей предметних галузей на основі текстових інформаційних потоків відкривають можливості для автоматизованої обробки великих обсягів текстової інформації з метою отримання цінних знань та прийняття рішень у проблемних галузях.

Однак, з огляду на швидкий розвиток інформаційного простору, дослідження та подальше удосконалення лінгвостатистичних методів залишаються невід'ємним завданням для вирішення нагальних проблем, пов'язаних із зростанням обсягів текстових даних та необхідністю ефективної обробки цих даних в умовах інформаційного перевантаження.

Робота є актуальною через стрімкий розвиток інформаційно-комунікаційних технологій та глобалізацію інформаційного простору, що призвело до значного збільшення обсягів інформаційних ресурсів в мережі Інтернет. Швидкий розвиток сучасних інформаційно-комунікаційних технологій породжує проблему інформаційного перевантаження. Це викликає не тільки приплив нових цінних знань, але й збільшення частки неструктурованих даних, включаючи "інформаційне сміття" та дублікати, що ускладнює пошук релевантної інформації. Відсутність відповідних технологічних рішень та нездатність існуючих систем обробляти величезні об'єми неструктурованих даних створюють критичну невідповідність між розвитком інформаційних систем і експоненційним збільшенням динамічних інформаційних потоків. Тож актуальність роботи полягає в необхідності розробки нових підходів та методів для ефективного пошуку, структуризації та аналізу цих неструктурованих текстових даних. Зокрема, важливим є процес концептуалізації та формалізації текстових даних у вигляді онтологічної моделі, що може покращити якість та точність обробки і аналізу. Розглянуті в роботі лінгвостатистичні методи формування мережевих моделей предметних галузей на основі текстових

інформаційних потоків відкривають можливості для автоматизованої обробки великих обсягів текстової інформації з метою отримання цінних знань та прийняття рішень у проблемних галузях. З огляду на швидкий розвиток інформаційного простору, дослідження та удосконалення лінгвостатистичних методів є актуальним завданням.

Оцінка обґрунтованості наукових результатів дисертації, їх достовірності та новизни.

У ході вивчення поставлених завдань у рамках дисертаційного дослідження були отримані нові наукові результати, що мають вагомий внесок у галузь Інформаційних технологій та мають важливе значення для подальших досліджень. Серед них здобувачем запропоновано та досліджено новий статистичний показник важливості термінів у тексті - GTF (Global Term Frequency), що відрізняється від звичайного TF-IDF та дозволяє ефективніше визначати ключові та інформаційно-важливі елементи тексту при роботі з текстовим корпусом визначеної теми. Використовуючи більш широку обробку природної мови – розбиття на частини мови (Part-of-speech tagging), вперше запропоновано метод виділення ключових термінів із текстового корпусу та метод визначення напрямків зв'язків. Також вперше розроблено лінгвостатистичний метод для автоматичного виділення та виявлення взаємозв'язків фразеологізмів в інформаційних потоках, з метою подальшого виявлення наративів і визначено форму візуального відображення інформаційного потоку в розрізі виявлених фразеологізмів – Ph-Di діаграму. Представлено новий підхід до визначення вагових значень зв'язків у мережі термінів. Вперше представлено цілісну технологічну схему формування мережевих моделей предметних галузей на основі текстових корпусів. Також вперше розроблено методику порівняння текстових документів, що ґрунтується на формуванні та порівнянні відповідних семантичних мереж, та на основі цього підходу запропоновано модель середовища інформаційного пошуку та модель ранжування як окремих документів, так і джерел інформації.

Достовірність наукових результатів дослідження експериментально доведена та гарантується використанням різноманітних наукових методів, таких як методи автоматичної обробки та аналізу природної мови та комп'ютерної лінгвістики. Методи, спрямовані на комп'ютеризовану обробку природномовних текстів, лексичний аналіз та виявлення семантичних зв'язків, підтверджують достовірність та надійність наукових результатів та висновків дисертаційної роботи. Додатково, застосування методів статистичного аналізу дозволило виділити ключові терміни в текстових даних, сприяючи об'єктивному визначенню їх важливості. У дослідженні також використовувалися методи дискретної математики, зокрема, теорії графів та складних мереж. Ці методи використовувалися для формування мережевих моделей предметних галузей на основі текстових корпусів. Теоретичне ґрунтування враховувало аналіз актуальної літератури та огляд існуючих методів, що сприяло розширенню та удосконаленню застосованих методів у рамках досліджень та забезпечило достовірність отриманих результатів.

Ключовим елементом дисертації є розроблення лінгвостатистичних методів формування мережевих моделей предметних галузей. Ці методи автоматизовано обробляють об'ємні тексти для подальшого аналізу та отримання цінних знань. Запропоновані методи створюють можливості для формування мережевих моделей, відображаючи структуру предметних галузей на основі текстових корпусів. Також у дисертації проведено експерименти на тестовому наборі даних, результати яких доводять ефективність запропонованих методів.

Отже, дисертаційне дослідження успішно вирішує актуальне науково-практичне завдання, пов'язане з концептуалізацією та формалізацією у вигляді мережі термінів неструктурованих текстових даних, розподілених у тематичних інформаційних потоках. Здобувач повною мірою оволодів методологією наукової діяльності та обґрунтував перспективи використання розроблених методів для вирішення завдань, пов'язаних зі структуризацією текстових даних та їх ефективною обробкою в умовах інформаційного перевантаження.

Оцінка змісту дисертації, її завершеність та дотримання принципів академічної доброчесності.

Зміст дисертації, її завершеність та дотримання принципів академічної доброчесності свідчить про високий рівень відповідності дисертаційної роботи здобувача Олега Олександровича Дмитренка вимогам Стандарту вищої освіти за спеціальністю 122 "Комп'ютерні науки" та відповідність освітній програмі з даного напрямку.

Дисертаційна робота є завершеною науковою працею, яка свідчить про важливий особистий внесок здобувача в науковий напрямок "Комп'ютерні науки". Результати перевірки дисертаційної роботи на текстові співпадіння підтверджують відсутність фальсифікації, компіляції, фабрикації, плагіату та запозичень.

Враховуючи, що використані ідеї, результати і тексти інших авторів мають належні посилання на відповідні джерела, можна зробити висновок про те, що дисертаційна робота є результатом самостійних досліджень здобувача та відповідає вимогам наукової доброчесності.

Мова та стиль викладення результатів.

Дисертаційна робота написана українською мовою. Стиль мовлення, використаний для викладення концепцій та результатів дослідження, є акуратним, структурованим та послідовним, що допомагає читачеві легко розуміти методологію дослідження та запропоновані методи. Розділи та параграфи передають логічну послідовність дослідження, а використання математичних методів та експериментальних результатів допомагає підкреслити обґрунтованість та наукову цінність представлених результатів. Загалом, стиль мовлення в роботі демонструє високий рівень наукової грамотності та вміння ефективно описувати складні ідеї у науковому контексті. Застосування загальноприйнятої термінології в роботі свідчить про глибоке знання та розуміння предметної області, високим професіоналізмом та володінням загальноприйнятою термінологією у галузі комп'ютерних наук. Автор ретельно використовує терміни та поняття, визначені в наукових джерелах, що робить його роботу актуальною та зрозумілою для наукової спільноти. Використані ідеї та тексти інших авторів мають відповідні посилання, що підтверджує дотримання наукової доброчесності.

Стиль викладення результатів дослідження у дисертаційній роботі відзначається високою послідовністю та логічною структурою. Це дозволяє читачеві легко розуміти хід дослідження та зв'язки між розділами. Чітке та зрозуміле представлення інформації сприяє засвоєнню основних понять та методів, використовуваних у роботі. Окрім того, слід відзначити виразність та чіткість стилю мовлення. Автор вдало використовує приклади та ілюстрації, які сприяють кращому усвідомленню основних ідей та результатів дослідження.

Структура роботи включає чітко сформульовані мету та завдання, а також розділи, які підкреслюють актуальність та наукову цінність дослідження. Дисертація має стандартну

структуру з вступом, чотирма розділами, висновками, списком літератури та додатками. Її загальний обсяг становить 170 сторінок, а основна частина складається з 131 сторінки.

У вступі автор чітко визначає мету та завдання дослідження, а також обґрунтовує його актуальність. Детально описуються проблематичні аспекти та підкреслюється наукова і практична новизна досягнутих результатів. Також подано інформацію про зв’язок роботи з науковими програмами, темами та апробацію матеріалів дисертації.

У першому розділі розглядається аналіз поточного стану проблеми та наукових досягнень, пов'язаних із темою дисертації. Виконано огляд актуальних підходів у комп'ютерно-лінгвістичних дослідженнях та методів автоматичного аналізу текстових інформаційних потоків. Спрямовано увагу на статистичні та лінгвістичні методи, зокрема на методи статистичного зважування термінів. Висвітлено особливості та важливість методики TF-IDF, яка визначає значущість термінів в документі у відношенні до всього корпусу текстів.

У другому розділі описано метод формування направлених зважених мереж з ключових термінів, які служать семантичними моделями предметних галузей на основі текстових корпусів. Запропоновано новий статистичний показник GTF (Global Term Frequency), що вказує на глобальну значимість терміна у всьому корпусі текстів. Виділено ефективність GTF у порівнянні зі стандартним TF-IDF для визначення ключових термінів у текстах. Також представлено метод виокремлення ключових термінів з використанням обробки природної мови та алгоритмів графів видимості.

Третій розділ включає розробку алгоритму для побудови динамічної мережі термінів, що дозволяє вивчати зміни вагових значень термінів при зміні їхньої глобальної частоти в тексті. Проведено аналіз динаміки вагових значень вузлів цієї мережі для виявлення термінологічних змін. Також представлено методику порівняння текстових документів за допомогою відповідних семантичних мереж.

Четвертий розділ присвячено практичним результатам застосування методик побудови мережевих моделей предметних галузей. Представлена технологічна схема обробки природомовного тексту за допомогою NLP функцій у Python, включаючи виокремлення та зважування ключових термінів. Розглянуто лінгвостатистичний метод екстрагування термінів та запропоновано Ph-Di діаграму для візуального відображення інформаційного потоку. Представлено модель середовища семантичного інформаційного пошуку та модель ранжування документів та інформаційних джерел. Запропоновано використання направлених зважених мереж термінів для формування бази знань у системі підтримки прийняття рішень в інформаційному пошуку.

Дисертаційна робота оформлена відповідно до вимог наказу МОН України від 12 січня 2017 р. № 40 «Про затвердження вимог до оформлення дисертації».

Оприлюднення результатів дисертаційної роботи.

Основні положення та результати дисертаційної роботи були представлені на 19 конференціях. Усього опубліковано 34 наукові праці, включаючи 5 одноосібних. З них 8 статей в українських фахових виданнях за спеціальністю здобувача 122 Комп'ютерні науки, 1 стаття у закордонному журналі Q3 за спеціальністю. За матеріалами конференцій опубліковано 25 робіт, серед яких 5 в міжнародних виданнях Scopus. Розширені матеріали конференцій включено до книг за спеціальністю здобувача 122 «Комп'ютерні науки», індексованих Scopus та WoS. Оформлено 1 свідоцтво про реєстрацію авторського права.

Загальна кількість публікацій у наукових виданнях, включених на дату опублікування до переліку наукових фахових видань України за спеціальністю 122

«Комп'ютерні науки» та у періодичних наукових виданнях, проіндексованих у базах даних Web of Science Core Collection та/або Scopus, з урахуванням числа співавторів та першого-третього квартилів (Q1-Q3) відповідно до класифікації SCImago Journal and Country Rank або Journal Citation Reports, становить 13 наукових публікацій.

У всіх публікаціях були дотримані принципи наукової доброчесності. Науковий рівень публікацій здобувача - високий, суттєвий особистий внесок прослідковується у кожній роботі.

Таким чином, наукові результати описані в дисертаційній роботі повністю висвітлені у наукових публікаціях здобувача.

Недоліки та зауваження до дисертаційної роботи.

1. Вважаю, що обрана УДК 004.912 не зовсім відповідає темі дослідження. Можливо, точніше буде вказати УДК 004.91.
2. Можна помітити поодинокі незначні граматичні помилки у словах та подекуди пропущені розділові знаки. Можливо, також пропущені слова у деяких словосполученнях, хоча вони й випливають з контексту, що не впливає на читабельність та розуміння висловлених ідей.
3. Шрифт на рисунку 1.4 не відповідає шрифту, що використовується у тексті дисертації. Хоч це не впливає на сприйняття представленої на рисунку схеми. Також зустрічаються поодинокі невідповідності в однозначному стилі оформлення математичних формул.
4. Присутні зайві відступи після таблиць та рисунків та зайві знаки переносу рядка між назвами рисунків і таблиць та самими рисунками та таблицями, відповідно. Присутні розриви між рисунками та назвами рисунків. Сама назва рисунків «Рисунок» написана скорочено «Рис.», проте вважаю, що для читабельності краще було б не використовувати зайвих скорочень.
5. Невідповідність нумерації сторінок у змісті. Деякі пункти змісту не відповідають головним розділам та підрозділам дисертації. Можливо, помилково був призначений заголовний стиль для вищезгаданих пунктів.
6. На мою думку, у третьому розділі «ДОСЛІДЖЕННЯ ТА АНАЛІЗ МЕРЕЖ ТЕРМІНІВ» алгоритмам HITS та PageRank приділено багато уваги.

Вважаю, що висловлені зауваження не є визначальними і не зменшують загальну наукову новизну та практичну значимість результатів та не впливають на позитивну оцінку дисертаційної роботи.

Висновок про дисертаційну роботу.

Вважаю, що дисертаційна робота здобувача ступеня доктора філософії Дмитренка Олега Олександровича на тему «Інформаційні технології формування та аналізу мережевих моделей предметних галузей на основі лінгвостатистичного підходу» свідчить про високий науковий рівень роботи. Результати дослідження є достатньо вагомими за своєю актуальністю і науковою новизною. Дисертація відповідає вимогам законодавства України, що передбачені в п. 6-9 «Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», затвердженого Постановою Кабінету Міністрів України від 12 січня 2022 р. № 44., не порушує академічні стандарти та принципи академічної доброчесності, є закінченим науковим дослідженням та має практичне значення в галузі інформаційних технологій.

Здобувач Дмитренко Олег Олександрович продемонстрував високий рівень науково-дослідницької компетентності, ретельно висвітливши теоретичні та практичні результати своєї роботи. З урахуванням вказаних факторів та відповідно до встановленого порядку, здобувач заслуговує на присудження ступеня доктора філософії в галузі знань Інформаційні технології за спеціальністю 122 «Комп'ютерні науки».

Офіційний рецензент:

старший науковий співробітник Інституту проблем реєстрації інформації НАН України, кандидат технічних наук _____ І. В. Балагура

ДЕРЖАВНА ОРГАНІЗАЦІЯ Інститут проблем реєстрації інформації Національної академії наук України код ЄДРПОУ 03363765

Підпис Балагури І.В.
ЗАСВІДЧУЮ: (Ніна КРАВЧЕНКО)
Нач. відділу кадрів ІПРІ
Національної академії наук України

М.П. "02" квітня 2024 року