ВІДГУК

офіційного опонента 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.

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 залишаються актуальним завданням.

Оцінка обґрунтованості наукових результатів дисертації, їх достовірності та новизни.

Під час розв'язання поставлених у дисертаційному дослідженні задач були отримані наступні результати, що включають елементи наукової новизни:

1. 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.

2. Вперше запропоновано метод виділення ключових термінів із текстового корпусу, що використовує більш широку обробку природної мови – розбиття на частини мови (Part-of-speech tagging).

3. Вперше розроблено лінгвостатистичний метод для автоматичного виділення та виявлення взаємозв'язків фразеологізмів в інформаційних потоках, з метою подальшого виявлення наративів як узагальнення сукупності фразеологізмів.

4. Вперше визначено форму візуального відображення інформаційного потоку в розрізі фразеологізмів - Ph-Di діаграму.

5. Розроблено правила визначення напрямків зв'язків між вузлами ненаправленої мережі, яка сформована з ключових слів та словосполучень тематичного текстового масиву, що змістовно відноситься до певної предметної галузі.

6. Розроблено та вперше використано метод визначення напрямків зв'язків, який базується на більш широкій обробці природної мови та використовує розбиття на частини мови (Part-of-speech tagging).

7. Представлено новий підхід до визначення вагових значень зв'язків у мережі термінів.

8. Вперше визначено цілісну технологічну схему формування мережевих моделей предметних галузей на основі текстових корпусів.

9. Вперше представлено методику використання направлених зважених мереж термінів для створення бази знань системи підтримки прийняття рішень під час розпізнавання інформаційних операцій.

10. 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 результатів дисертаційного дослідження гарантується використанням різноманітних наукових методів, включаючи методи автоматичної обробки та аналізу природної мови та комп’ютерної лінгвістики. Ці методи дозволили провести передбачувану комп'ютеризовану обробку природномовних текстів, здійснити лексичний аналіз та виявити семантичні зв'язки. Використані також методи статистичного аналізу для виділення ключових термінів із текстових даних, що сприяло об’єктивному визначенню важливих елементів. Додатково, використання методів дискретної математики, зокрема, теорії графів та складних мереж, ефективно дозволило формувати мережеві моделі предметних галузей та проводити подальший аналіз отриманих моделей.

Теоретичне ґрунтування також враховує аналіз актуальної літератури та огляд існуючих методів, враховуючи внесок вчених як вітчизняного, так і

зарубіжного наукового співтовариства. Узагальнення цих підходів дало можливість розширити та удосконалити використані методи в рамках досліджень.

Ключовим елементом дисертації стало розроблення лінгвостатистичних методів формування мережевих моделей предметних галузей на основі текстових корпусів. Ці методи дозволяють автоматизовано обробляти об'ємні тексти для подальшого аналізу та отримання цінних знань. Розгляд цих методів підкреслює їхню актуальність у вирішенні проблем, пов'язаних із зростанням обсягів текстових даних та необхідністю ефективної обробки цих даних в умовах інформаційного перевантаження.

Запропоновані в цій дисертаційній роботі лінгвостатистичні методи створюють можливості для формування мережевих моделей, що відображають структуру предметних галузей на основі текстових корпусів. Це відкриває перспективи для розуміння та аналізу інформації, яка міститься в текстах, пов'язаних з конкретними проблемними галузями. Застосування цього методу може служити для автоматизованого формування онтологічних моделей, що є ключовим елементом у концептуалізації текстових даних та їхній подальшій формалізації.

Результати дисертаційної роботи Дмитренка Олега Олександровича на тему «Інформаційні технології формування та аналізу мережевих моделей предметних галузей на основі лінгвостатистичного підходу» було використано і впроваджено при реалізації програмно-технічних засобів в середовищі Інформаційно-аналітичного ситуаційного центру КПІ ім. Ігоря Сікорського для комплексного використання та інтелектуального аналізу великих масивів неструктурованих даних різної природи, в тому числі результатів обробки текстів природної мови, в ході виконання низки держбюджетних і договірних науково-дослідних робіт (НДР) та проектів ННЦ «СЦД-Україна». Практичне значення одержаних результатів дослідження підтверджується актами впроваджень.

Отже, в дисертаційній роботі поставлене наукове завдання вирішено актуальне науково-практичне завдання, що стосується концептуалізації та подальшої формалізації у вигляді мережі термінів неструктурованих текстових даних, що містяться у тематичних інформаційних потоках виконано повністю, здобувач повною мірою оволодів методологією наукової діяльності.

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 визнається як завершена наукова праця, 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.

Мова дисертації – українська. 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. Дисертація відзначається 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.

Структура роботи включає чітко сформульовані мету та завдання, а також змістовні розділи, які підкреслюють актуальність та наукову цінність дослідження. Дисертація має стандартну структуру з вступом, чотирма розділами, висновками, списком літератури та додатків. Її загальний обсяг становить 170 сторінок, і 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.

Перший розділ дослідження присвячений аналізу сучасного стану проблеми та наукових досягнень, які стосуються теми дисертації. Проведено огляд актуальних комп’ютерно-лінгвістичних підходів і методів автоматичного аналізу текстових інформаційних потоків з метою виявлення знань у предметній галузі, пов'язаній з відповідними текстами.

У ході дослідження розглянуті різні сучасні підходи, включаючи статистичний та лінгвістичний, зосереджено увагу на методах статистичного зважування термінів. Зокрема, визначено, що серед них особливо важливою і широко використовуваною є методика TF-IDF для оцінки значущості термінів. Цей статистичний показник важливості термінів вказує на їхню значущість у

конкретному документі порівняно з корпусом загалом. Терміни з вищим значенням TF-IDF вважаються важливими для конкретного документа і менш зустрічаються в інших документах корпусу. Крім того, в першому розділі висвітлені проблеми, які можуть виникнути під час використання методів статистичного зважування. Проведено детальний аналіз основних рівнів лінгвістичної обробки текстових даних та розглянуто основні концепції семантичного пошуку, що представляє собою перспективний напрямок у сфері автоматизованого повнотекстового інформаційного пошуку.

У другому розділі розглядається цілісна методика формування направлених зважених мереж із ключових термінів, як семантичних моделей предметних галузей на основі текстових корпусів. Пропонується та вивчається новий статистичний показник важливості термінів у тексті – GTF (Global Term Frequency) – глобальна частота терміна. Цей показник визначається відношенням загальної кількості появ терміна у всіх документах корпусу до загальної кількості термінів у документах корпусу та вказує на глобальну значимість слова. Досліджено, що GTF виявляється більш ефективним у знаходженні інформаційно-важливих елементів тексту, особливо при роботі з текстовим корпусом конкретної тематики. В порівнянні з показником TF-IDF, GTF дозволяє більш ефективно виділяти терміни, які є ключовими для теми дослідження, і зустрічаються майже в усіх документах корпусу. Також у другому розділі представлено новий метод виділення ключових термінів з текстового корпусу, використовуючи більш розширену обробку природної мови, засновану на розбитті на частини мови (Part-of-speech tagging). Проведено огляд алгоритмів графів видимості (Visibility Graph algorithm – VG), які можуть бути використані для формування мережевих моделей предметних галузей. Також пропонується новий метод визначення напрямків зв’язків та їх вагових значень у мережі термінів.

У третьому розділі дисертаційної роботи запропоновано та розроблено алгоритм для побудови динамічної мережі термінів та проведено аналіз

динаміки вагових значень вузлів цієї мережі. Представлений алгоритм дозволяє вивчати зміни вагових значень конкретних ключових термінів при зміні їхньої глобальної частоти в тексті. Це може виявитися корисним при виявленні та аналізі термінологічних змін. Зокрема, можна виявляти термінологічні збагачення, які можуть бути результатом штучних утручань, пропаганди, спаму чи навмисних інформаційних атак. Використання алгоритму динамічної мережі термінів може допомогти у їх виявленні шляхом аналізу зміни вагових значень ключових термінів. Далі в розділі представлена методика порівняння текстових документів, що ґрунтується на побудові та порівнянні їх семантичних мереж. Ця методика може бути основою для систем порівняння правових документів, зокрема, в рамках парламентського контролю. Також у розділі представлено новий алгоритм побудови семантичних мереж як одного з видів онтологій, який може знайти застосування в процесі обробки запитів при проведенні інформаційного пошуку, дозволяючи визначити ступінь подібності або відмінності структури та семантики текстів.

У четвертому розділі розглядаються практичні результати застосування методики побудови мережевих моделей предметних галузей на основі текстових корпусів. Представлена технологічна схема обробки природомовного тексту за допомогою NLP функцій у мові програмування Python, включаючи виокремлення ключових термінів, біграм та триграм за визначеними шаблонами та їх статистичне зважування. Додатково, введений, реалізований та випробуваний лінгвостатистичний метод автоматичного екстрагування, спрямований на дослідження динаміки та виявлення взаємозв’язків фразеологізмів в інформаційних потоках. Запропоновано Ph-Di діаграму (Phraseme Diagram) для візуального відображення інформаційного потоку в розрізі фразеологізмів та дат. Також представлені модель середовища семантичного інформаційного пошуку та модель ранжування документів та джерел інформації в контексті проблемної галузі. Подано методику використання направлених зважених мереж термінів для формування бази знань

у системі підтримки прийняття рішень при розпізнаванні інформаційних операцій.

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 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, 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.

Всі публікації здобувача відзначаються високим науковим рівнем, у них ретельно висвітлені основні результати досліджень, з важливим особистим внеском у розкритті експериментальних аспектів. Науковий етикет дотримано у всіх публікаціях. Таким чином, наукові результати дисертації повністю відображені у наукових публікаціях здобувача.

Недоліки та зауваження до дисертаційної роботи.

1. В роботі недостатньо вдало сформульовано мету роботи: «розробити нові методи побудови мережевих моделей предметних галузей” та “розробити нові методи аналізу сформованих мереж” є задачами, що вирішуються в ході досліджень, а метою, як випливає з тексту самої роботи, є структурування природномовних даних та здобуття з них корисних знань предметних областей, і саме для цього потрібні мережеві моделі.
2. Занадто багато уваги приділяється традиційним підходам до інформаційного пошуку (за ключовими словами, булеві запити), тоді як зараз вже існує велика кількість систем лінгвістичного та семантичного розширення запитів, пошуку на основі онтологій, ймовірнісного пошуку, категоризації, оцінки метаданих тощо, які теж аналізуються у роботі.
3. У розділі 2.4, де визначається один з ключових елементів дослідження – показник важливості термінів у тексті GTF, не описується різниця між поняттями поняттями “слово”, “терм” та “термін”, а самі вони використовуються у визначеннях у подібному контексті, що значно ускладнює розуміння змісту роботи. Не пояснюється, як саме застосовано показник для знаходження ключових слів, не вказується, яка кількість слів є ключовими або яка частота є достатньою.
4. В таблицях 2.1, 2.2 наводяться результати підрахунку за різними показниками. Не визначається формально, які саме ситуації має виявляти запропонований автором показник (лише загальні твердження) краще, ніж традиційний, та нема чіткого алгоритму обробки його значень. Доцільно вводити певні метрики чи оцінки для ідентифікації таких ситуацій при порівнянні значень обох показників.
5. До визначення ступеня подібності текстових документів (розд.3.5) на сьогодні розроблено дуже багато підходів. Потрібно чіткіше

позиціонувати запропонований автором підхід та визначити його переваги порівняно з іншими

6. В тексті роботи зустрічаються синтаксично незрозумілі мовні конструкції (наприклад, “Побудувавши направлену мережу за першим правилом було отримано “) та незакінчені фрази (с.98).
7. Список використаних джерел побудовано досить незручно для сприйняття. Перші 35 посилань – це публікації здобувача, які не відокремлені від тих посилань, які характеризують поточний стан проблеми (цей перелік робіт наведено окремо у кінці роботи). Ресурси 39-46 – статистичні дані, а не матеріали наукових досліджень, які підтверджують актуальність проблематики, але не характеризують підходи інших вчених до проблеми.

Вважаю, що висловлені зауваження більше стосуються форми викладення матеріалу, а не власно результатів дослідження, і тому 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 та 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 опонент:

Старший науковий співробітник Інституту програмних систем НАН України, кандидат фізико-математичних наук, доцент _____ Ю. В. Рогушина

Підпис Рогушиної Юлії Віталіївни засвідчую:

Начальник відділу кадрів Інституту програмних систем НАН України _____ О.О.Яцюк

НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
Інститут програмних систем
Ідентифікаційний код №05540149
м. Київ

М.П. «4» квітня 2024 року