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. Тема дисертаційної роботи є актуальною у контексті розвитку сучасного інформаційного суспільства, адже спрямована на вирішення складних проблем, що виникають внаслідок стрімкого наростання обсягів текстових даних у Інтернеті та інших джерелах, призводячи до надмірного інформаційного завантаження. З урахуванням величезного обсягу неструктурованих текстових інформаційних потоків, що супроводжуються різноманітними інформаційними джерелами, виникає актуальна задача пошуку релевантної та важливої інформації.

Основний акцент роботи зроблено на процесі концептуалізації текстових даних та їх подальшій формалізації у вигляді онтологічної моделі з метою ефективної автоматичної обробки та аналізу комп'ютерними засобами.

Розгляд лінгвостатистичних методів для формування мережевих моделей предметних галузей, які викладено в дисертації, відкриває перспективні можливості для автоматизованої обробки та аналізу великих обсягів текстової інформації в умовах великого потоку текстових даних та динамічних масивів інформації. Актуальність роботи підкреслюється також необхідністю удосконалення методів та технологій для ефективного вирішення цих завдань, з метою забезпечення оптимальної обробки та аналізу неструктурованих текстових даних у сучасному інформаційному середовищі.

Оцінка обґрунтованості наукових результатів дисертації, їх достовірності та новизни. Найбільш важливими науковими результатами дисертаційного дослідження, які мають високий ступінь новизни, полягають в тому, що 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;
- метод виділення ключових термінів із текстового корпусу, що використовує більш широку обробку природної мови, що базується на розбитті на частини мови (Part-of-speech tagging);
- лінгвостатистичний метод автоматичного екстрагування та виявлення взаємозв'язків фразеологізмів в інформаційних потоках з метою подальшого виявлення наративів, як узагальнення сукупності фразеологізмів;
- метод визначення напрямків зв'язків з використанням більш широкої обробки природної мови, базуючись на розбитті на частини мови (Part-of-speech tagging);
- новий підхід до визначення вагових значень зв'язків у мережі термінів;
- методику використання направлених зважених мереж термінів для формування бази знань системи підтримки прийняття рішень під час розпізнавання інформаційних операцій;
- модель середовища інформаційного пошуку та модель ранжування як окремих документів, так і джерел інформації.

Достовірність результатів дисертаційного дослідження забезпечена завдяки коректному використанню ряду наукових методів, зокрема, методів автоматичної обробки та аналізу природної мови та комп'ютерної лінгвістики. Ці методи дозволили автору провести попередню комп'ютеризовану обробку природномовних текстів, виконати лексичний аналіз, виявити семантичні зв'язки

та, окрім того, забезпечити об'єктивне виділення ключових термінів із текстових даних за допомогою статистичного аналізу.

Теоретичне обґрунтування дослідження забезпечене аналізом актуальної літератури, оглядом існуючих методів та врахуванням внеску вчених як вітчизняного, так і зарубіжного наукового співтовариства у розвиток методів обробки природномовних текстів та аналізу складних мереж. Узагальнення цих підходів дозволило розширити та удосконалити використані методи для вирішення поставлених завдань.

Важливим елементом дисертації є розробка лінгвостатистичних методів формування мережевих моделей предметних галузей на основі текстових корпусів. Ці методи дозволяють автоматично обробляти великі обсяги текстової інформації для подальшого аналізу та отримання цінних знань. Використання цих методів в контексті дисертації підкреслює їхню актуальність для вирішення проблем, пов'язаних із зростанням обсягів текстових даних та необхідністю ефективної обробки цих даних в умовах інформаційного перевантаження.

Завдяки використанню у цій дисертаційній роботі лінгвостатистичних методів формуються мережеві моделі, що відображають структуру предметних галузей на основі текстових корпусів. Це відкриває широкий спектр можливостей для розуміння та аналізу інформації, яка міститься в текстах, пов'язаних з конкретною проблемною галуззю. Такий підхід може бути успішно використаний для автоматизованого формування онтологічних моделей, які є важливим елементом у концептуалізації текстових даних та їхній подальшій формалізації.

Наукові дослідження Дмитренка Олега Олександровича виконано у відділі спеціалізованих засобів моделювання Інституту проблем реєстрації інформації НАН України відповідно до плану фундаментально-прикладних наукових досліджень, що увійшли до науково-дослідних робіт: «Розробка методів і моделей підтримки прийняття рішень при розпізнаванні інформаційних операцій» (2019-2020 рр., державний реєстраційний номер: 0119U001867), «Розробити механізми підвищення живучості для забезпечення функціональної стійкості систем

організаційного управління об'єктів критичних інфраструктур» (2017-2021 рр., державний реєстраційний номер 0117U004106). Також результати роботи та практичні напрацювання були задіяні у рамках ННЦ "Світовий центр даних з геоінформатики та сталого розвитку" Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» під час науково-технічної роботи за державним замовленням на науково-технічні (експериментальні) розробки та науково-технічну продукцію «Створення інтегрованої платформи для ситуаційного аналізу соціально-економічних і безпекових явищ» (2021-2022 рр., державний реєстраційний номер: 0121U113470), у науково-дослідній роботі «Створення інформаційно-аналітичного ситуаційного центру для сценарного моделювання кризових і безпекових явищ та вивчення їх впливу на економіку і суспільство» (2021-2022 рр., нак. кер. д.т.н., проф. Ю.П.Зайченко, державний реєстраційний номер: 0121U109764), «Розробка методології та програмно-технічного комплексу для системної оцінки безпекового рівня територій України на основі супутникових даних за умов множинних військових загроз» (2023-2024 рр., державний реєстраційний номер: 0123U102015) та «Розробка програмно-технічного комплексу інтелектуального аналізу неструктурованих даних методами штучного інтелекту та OSINT для планування військових операцій» (2024-2026 рр., державний реєстраційний номер: 0124U000838). Також була здійснена реєстрація авторського права на твір № с202204275 від 19.09.2022 – Комп'ютерна програма автоматичної побудови мереж термінів на основі аналізу текстових потоків «TermsNet».

Отже, у дисертаційній роботі успішно вирішено поставлене науково-практичне завдання концептуалізації та формалізації у вигляді мережі термінів неструктурованих текстових даних в контексті тематичних інформаційних потоків. Враховуючи актуальність, важливість та достовірність отриманих результатів, здобувач повною мірою оволодів методологією наукової діяльності.

Оцінка змісту дисертації, її завершеність та дотримання принципів академічної доброчесності. За своїм змістом дисертаційна робота здобувача Дмитренка О.О. повністю відповідає Стандарту вищої освіти зі спеціальності 122 «Комп'ютерні науки» та напрямкам досліджень відповідно до освітньої програми «Комп'ютерні науки».

Дисертаційна робота є завершеною науковою працею і свідчить про наявність особистого внеску здобувача у галузь Інформаційних технологій та, зокрема, науковий напрям «Комп'ютерні науки».

Розглянувши звіт подібності за результатами перевірки дисертаційної роботи на текстові співпадіння, можна зробити висновок, що дисертаційна робота Дмитренка Олега Олександровича є результатом самостійних досліджень здобувача і не містить елементів фальсифікації, компіляції, фабрикації, плагіату та запозичень. Використані ідеї, результати і тексти інших авторів мають належні посилання на відповідне джерело.

Мова та стиль викладення результатів. Дисертаційна робота написана українською мовою. Робота характеризується високою грамотністю, послідовністю та логічною структурою викладення, що дозволяє читачеві з легкістю розуміти розвиток дослідження та логічні зв'язки між розділами. Представлення інформації чітке та зрозуміле, що сприяє сприйняттю основних понять та методів, які використовуються в дослідженні.

Дисертація демонструє високий професіоналізм та володіння загальноприйнятою термінологією у даній науковій області. Автор вміло використовує терміни та поняття, що визначені в наукових джерелах, здійснює порівняння з існуючими підходами та методиками, що дозволяє дисертації бути актуальною та зрозумілою для наукової спільноти. Також варто відзначити стиль мовлення, який характеризується виразністю та чіткістю. Використання прикладів та ілюстрації допомагають у сприйнятті основних ідей та результатів дослідження. Структура роботи включає чітко сформульовані мету та завдання, а

також змістовні розділи, які підкреслюють актуальність та наукову цінність дослідження.

Дисертація складається з вступу, чотирьох розділів, висновків, списку літератури та додатків. Загальний обсяг дисертації 170 сторінок, серед яких основну частину складають 131 сторінка.

У вступі визначаються мета та основні завдання дослідження, обґрунтовується актуальність теми. Також описуються проблематика існуючих підходів та наводиться наукова і практична новизна отриманих результатів. Також надана інформація про зв'язок дисертаційної роботи з науковими програмами, планами, темами. Здобувач також зазначає особистий внесок, апробацію матеріалів дисертації та перелічує публікації за темою роботи.

У першому розділі висвітлено сучасний стан проблеми та наукові підходи до автоматичного аналізу текстових інформаційних потоків, розглянуті комп'ютерно-лінгвістичні підходи, в тому числі статистичний та лінгвістичний, зокрема метод TF-IDF для визначення важливості термінів. Детально розглянуті рівні лінгвістичної обробки текстових даних та ідеї семантичного пошуку.

У другому розділі представлено методику формування направлених зважених мереж із ключових термінів, як семантичних моделей предметних галузей. Запропоновано новий статистичний показник важливості термінів - GTF (Global Term Frequency), який ефективно виявляє інформаційно-важливі елементи тексту. Також представлено новий метод виокремлення ключових термінів з використанням обробки природної мови та огляд алгоритмів графів видимості для формування мережевих моделей.

Третій розділ містить алгоритм побудови динамічної мережі термінів та дослідження динаміки вагових значень вузлів. Використовуючи цей алгоритм, можна виявляти та аналізувати динаміку ключових термінів та зміни їх вагових значень при зміні їхньої глобальної частоти в текстовому інформаційному потоці.

Четвертий розділ включає результати практичного використання методик. Представлено технологічну схему виокремлення та формування ключових

термінів з текстів, лінгвостатистичний метод екстрагування ключових термінів та виявлення фразеологізмів, та моделі семантичного пошуку та ранжування як текстових документів, так й інформаційних джерел. Наведено приклади використання направлених зважених мереж термінів для формування бази знань системи підтримки прийняття рішень.

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 оприлюднені й обговорювались на 19-тьох конференціях. За результатами дисертаційних досліджень опубліковано 34 наукові праці, в тому числі 5 – одноосібні. Серед них 8 наукових статей опубліковані в фахових наукових виданнях України, серед яких за спеціальністю здобувача – 6 статей, не за спеціальністю – 2, та 1 стаття опублікована у фаховому закордонному журналі, що належить до квартилю Q3 за спеціальністю здобувача 122 «Комп'ютерні науки». За матеріалами виступів на 19-ти науково-технічних конференціях опубліковано 25 робіт, серед них 9 тез доповідей наукових конференцій, 6 статей конференцій, 5 статей, що розміщені в міжнародному електронному виданні CEUR Workshop Proceedings, що індексується базою Scopus. Розширені та доопрацьовані матеріали конференцій увійшли як окремі розділи до книг за спеціальністю здобувача 122 «Комп'ютерні науки», які також індексується Scopus та WoS. Також було оформлено 1 свідоцтво про реєстрацію авторського права на твір.

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.

Усі публікації здобувача мають високий науковий рівень. У них детально розкриваються основні наукові результати дослідження. Особистий внесок здобувача у публікаціях зі співавторством вагомий, особливо у описі експериментальних частин роботи. Принципів академічної доброчесності у жодній з публікацій не порушено.

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

Недоліки та зауваження до дисертаційної роботи.

На мою думку, дисертаційна робота має наступні недоліки:

1. У вступі неточно вказано загальний обсяг та, відповідно, обсяг основної частини дисертації, оскільки «до загального обсягу дисертації не включаються таблиці та ілюстрації, які повністю займають площу сторінки».
2. Не достатньо повно представлена лексикографія. Відсутні деякі визначення та посилання на терміни, зокрема такі як: лематизація, синтаксичний аналіз та інші.
3. У дисертації не описано методи та моделі підтримки прийняття рішень, про які є згадка у роботі.
4. Достатньо мало уваги приділено перевагам запропонованої моделі середовища інформаційного пошуку. Хотілося б мати більш детальний опис переваг, а також і наявних недоліків, якщо вони неочевидні.
5. Залишились без розгляду питання щодо представленої у роботі моделі ранжування документів та інформаційних джерел, а саме: Які основні відмінності запропонованої моделі та її результатів ранжування у порівнянні з широко застосовуваними у відомих інформаційно-пошукових системах? І чи можливе застосування запропонованої моделі окремо до ранжування джерел, без попереднього ранжування документів?

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» виконана на високому науковому рівні, не порушує принципів академічної доброчесності та є закінченим науковим дослідженням, сукупність теоретичних та практичних результатів якого розв'язує наукове завдання, що має істотне значення для інформаційних технологій. Дисертаційна робота за актуальністю, практичною цінністю та науковою новизною повністю відповідає вимогам чинного 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.

Здобувач Дмитренко Олег Олександрович заслуговує на присудження ступеня доктора філософії в галузі знань Інформаційні технології за спеціальністю 122 «Комп'ютерні науки».

Офіційний рецензент:

заступник директора Інституту проблем реєстрації інформації НАН України,
член-кореспондент НАН України, д.т.н., професор _____ А. А. Крючин

М.П.

« 3 » квітня 2024 року