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 12 – «Інформаційні технології» за спеціальністю 122 – «Комп'ютерні науки»

Актуальність теми дисертації.

В сучасному інформаційному суспільстві виникає критична проблема – невідповідність між розвитком інформаційних систем та їх здатністю обробляти величезні об'єми неструктурованих даних з необхідною швидкістю та ефективністю. Тож актуальність полягає в потребі розробки нових підходів та методів для структуризації та аналізу цих неструктурованих текстових даних. Ключовим аспектом є процес концептуалізації текстових даних та їх подальшої формалізації у вигляді онтологічної моделі для ефективної обробки та аналізу.

З огляду на величезний обсяг текстових інформаційних потоків та динамічних масивів даних, розглянуті в дисертації лінгвостатистичні методи формування мережевих моделей предметних галузей відкривають можливості для автоматичної обробки та аналізу великих обсягів текстової інформації. Актуальність роботи також підкреслюється потребою в удосконаленні методів та технологій для вирішення цих завдань, з метою забезпечення ефективної обробки та аналізу неструктурованих текстових даних у сучасному інформаційному середовищі.

Оцінка обґрунтованості наукових результатів дисертації, їх достовірності та новизни.

Наукова новизна найважливіших результатів даного дисертаційного дослідження полягає в тому, що запропоновано та досліджено:

1. 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. 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);
3. лінгвостатистичний метод автоматичного екстрагування та виявлення взаємозв'язків фразеологізмів в інформаційних потоках з метою подальшого виявлення наративів, як узагальнення сукупності фразеологізмів;

4. метод визначення напрямків зв’язків з використанням більш широкої обробки природної мови, базуючись на розбитті на частини мови (Part-of-speech tagging);
5. новий підхід до визначення вагових значень зв’язків у мережі термінів;
6. методику використання направлених зважених мереж термінів для формування бази знань системи підтримки прийняття рішень під час розпізнавання інформаційних операцій;

Достовірність результатів дисертаційного дослідження забезпечено використанням ряду наукових методів автоматичної обробки та аналізу природної мови та методів комп’ютерної лінгвістики. Ці методи дозволяли провести попередню комп'ютеризовану обробку природномовних текстів, виконати лексичний аналіз та виявити семантичні зв'язки. Методи статистичного аналізу були використані для виокремлення ключових термінів (слів та словосполучень) із текстових даних, що також сприяло об’єктивному визначенню важливих елементів. Додатково, використання методів дискретної математики, зокрема, методів теорії графів та складних мереж, дозволило ефективно формувати мережеві моделі предметних галузей та подальше дослідження і аналіз отриманих моделей.

Одним із ключових елементів дисертації є розробка лінгвостатистичних методів формування мережевих моделей предметних галузей на основі текстових корпусів. Ці методи дозволяють автоматично обробляти великі обсяги текстової інформації з метою подальшого аналізу та отримання цінних знань. Розгляд 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.

Практичне значення одержаних результатів дослідження здобувача Дмитренка Олега Олександровича підтверджується актами впроваджень при реалізації програмно-технічних засобів в середовищі Інформаційно-аналітичного ситуаційного центру КПІ ім. Ігоря Сікорського в ході виконання низки держбюджетних і договірних науково-дослідних робіт (НДР) та проектів ННЦ «СЦД-Україна».

Отже, в дисертаційній роботі поставлене актуальне науково-практичне завдання, яке стосується концептуалізації та подальшої формалізації у вигляді мережі термінів, що містяться у тематичних інформаційних потоках неструктурованих текстових даних вирішено і виконано повністю, 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 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 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 засвоєнню 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, це дозволяє дисертації бути актуальною та зрозумілою для наукової спільноти, а також забезпечує легке порівняння з існуючими підходами та методиками.

Стиль мовлення характеризується виразністю та чіткістю, автор використовує влучні приклади та ілюстрації, що допомагають усвідомити основні ідеї та результати дослідження. Структура роботи включає чітко сформульовані мету та завдання, а також змістовні, добре збалансовані розділи.

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

У вступі викладаються мета та основні завдання дослідження, а також обґрунтовується його актуальність. Описуються проблематичні аспекти, що виникають у зв'язку з існуючими підходами, та наголошується на науковій і практичній новизні та практичному значенні досягнутих результатів. Також надана інформація про зв’язок дисертаційної роботи з науковими програмами, планами, темами. Здобувач вказує на власний внесок у дослідження, апробацію матеріалів дисертації та перелік публікацій, пов'язаних із темою дослідження.

У першому розділі здійснено огляд сучасного стану проблеми та наукових розробок, яким присвячена тема дисертації. Було розглянуто сучасні

комп’ютерно-лінгвістичні підходи та методи автоматичного аналізу текстових інформаційних потоків з метою розпізнавання знань з предметної галузі з якою змістовно пов’язані текстові дані. Було встановлено, що існує декілька підходів, зокрема такі як статистичний та лінгвістичний. Здійснено огляд існуючих методів статистичного зважування термінів, серед них найбільш відомою й використовуваною в наш час оцінкою важливості термінів є TF-IDF. Встановлено, що статистичний показник важливості терміна TF-IDF показує наскільки добре даний термін визначає документ по відношенню до корпусу. Терміни з більш високим числовим значенням TF-IDF є важливими в межах певного документа й нечасто зустрічається в інших документах корпусу. Крім цього було акцентовано увагу й на проблемах, які можуть виникати під час використання методів статистичного зважування. Детально розглянуто основні рівні лінгвістичної обробки текстових даних. Розглянуто основні ідеї семантичного пошуку, як одного із найперспективніших видів автоматизованого повнотекстового інформаційного пошуку.

У другому розділі запропоновано та представлено цілісну методику формування направлених зважених мереж із ключових термінів, як семантичних моделей предметних галузей на основі текстових корпусів. Зокрема запропоновано та досліджено новий статистичний показник важливості термінів у тексті – GTF (Global Term Frequency) – глобальна частота терміна, що визначається відношенням загальної кількості появи терміна у всіх документах корпусу до загальної кількості термінів у документах корпусу й показує, наскільки значимим є слово в глобальному контексті. Було показано, що запропонована оцінка важливості термінів на відміну від звичайного статистичного показника TF-IDF дозволяє більш ефективно знаходити інформаційно-важливі елементи тексту під час роботи з текстовим корпусом заздалегідь визначеної теми, коли інформаційно-важливий термін зустрічається майже у кожному документі корпусу. Також запропоновано новий метод виокремлення ключових термінів із текстового корпусу зі застосуванням обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging). Також у цьому розділі досліджено алгоритми графів видимості (Visibility Graph algorithm – VG), що можуть використовуватись для формування мережевих моделей предметних галузей, і запропоновано новий метод визначення напрямків зв’язків та їх вагових значень у мережі термінів.

В третьому розділі дисертаційної роботи було запропоновано алгоритм побудови динамічної мережі термінів та за його допомоги досліджено динаміку вагових значень вузлів у мережі термінів. Використовуючи алгоритм побудови динамічної мережі термінів можна досліджувати динаміку окремих ключових термінів в результаті підвищення чи зниження їх глобальної частоти вживання у тексті шляхом додавання текстових документів, які насичені чи збагачені

окремим визначеним терміном, у інформаційний потік. Показано, що такі термінологічні збагачення можуть бути штучними й викликані «інформаційними вкидами», пропагандою чи спамом. Також вони можуть бути результатом навмисних, цілеспрямованих інформаційних атак – інформаційних операцій. Тож їх виявлення може бути здійснене шляхом аналізу динаміки ключових термінів, отриманої в результаті застосування алгоритму побудови динамічної мережі термінів. У цьому ж розділі також викладено методику порівняння текстових документів, що базується на побудові та порівнянні відповідних їм семантичних мереж. Ця методика може стати основою побудови систем порівняння правових документів у рамках парламентського контролю. Також розглянуто алгоритм побудови семантичних мереж як одного із видів онтологій. Цей алгоритм також може застосовуватися в системах автоматичного реферування правової інформації з метою формування лаконічних інформаційно-насичених звітів, коротких анотацій або дайджестів. Пропонована методика може бути використана в процесі обробки запитів при проведенні інформаційного пошуку, надаючи можливість визначення ступеня подібності або відмінності структури та семантики текстів.

У четвертому розділі були висвітлені результати практичного застосування запропонованої методики побудови мережевих моделей предметних галузей на основі текстових корпусів. Спершу представлена цілісна технологічна схема виокремлення та формування ключових термінів із текстів інформаційних повідомлень, яка передбачає послідовну обробку отриманого на вхід природномовного тексту за допомогою функцій NLP бібліотек мови програмування Python, формування та виокремлення ключових слів, біграм та триграм за визначеними шаблонами, та їх подальше статистичне зважування за частотою появи у тексті. У цьому розділі дисертаційної роботи запропонований, реалізований та апробований лінгвостатистичний метод автоматичного екстрагування, дослідження динаміки і виявлення взаємозв’язків фразеологізмів в інформаційних потоках з метою подальшого виявлення наративів, як узагальнення сукупності фразеологізмів. Також запропонована форма візуального відображення інформаційного потоку в розрізі фразеологізмів і дат – Ph-Di діаграма (Phraseme Diagram) та представлені моделі середовища семантичного інформаційного пошуку та ранжування як окремих документів, так і джерел інформації, що стосуються визначеної у інформаційному запиті проблемної галузі. Наприкінці розділу розглянуто методику використання направлених зважених мереж термінів для формування бази знань системи підтримки прийняття рішень під час розпізнавання 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 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 демонструють високий науковий рівень, і в них ретельно висвітлюються основні наукові результати досліджень. Здобувач зробив значний особистий внесок у публікації, особливо у розкритті експериментальних аспектів роботи.

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

Недоліки та зауваження до дисертаційної роботи.

1. У дисертаційній роботі частину використаної термінології не визначено. Так, 16 разів у дисертації згадується категорія «прийняття рішень», але не наведено моделі та методи, які при цьому застосовуються. 10 разів зустрічається термін «база знань», але не надано його визначення та не представлено відповідну модель. Також, відсутні означення термінів «лематизація», «графемний аналіз», «синтаксичний аналіз» тощо, і, разом з тим, відсутні посилання на літературні джерела, де ці терміни визначені.

2. В огляді інструментарію, що застосовується в дослідженнях не розглядаються нейромережі, що ускладнює подальше сприйняття пов’язаного з ними матеріалу.

3. Онтологічні моделі використовуються без належної аксіоматики, що, у певній мірі затрудняє розуміння легітимності застосування цих моделей. Так, наприклад, в якості онтології зручною й ефективною моделлю представлення текстових даних може розглядатися лінгвомережева модель – модель, що представляє собою мережу із ключових термінів (слів та словосполучень), поєднаних між собою змістовними зв’язками.

4. В аналізі наукових джерел, окрім зрозумілого посилання на класичні базові роботи, на мою думку, все ж зустрічаються, посилання, на досить застарілі джерела, наприклад, Балог, В. (2005). Сучасний стан української комп’ютерної лінгвістики. Лексикографічний бюлетень. У той же час, невиправдано мало уваги приділено роботам вітчизняної наукової школи академіка Широкова.

5. В дисертаційному дослідженні не заслужено не використовуються можливості, що може надати концептографія та лексикографія.

6. Не достатньо уваги в дисертації приділено опису та використанню алгоритмів центральності з відповідними посиланнями на першоджерела.

7. У вступі у пункті «Наукова новизна отриманих результатів» завжди необхідно зазначити відмінності одержаних результатів від відомих раніше, також, обов’язково зазначати ступінь їх новизни. Деякі пункти новизни доцільно було б об’єднати. Так, пункт 8 результатів «вперше запропоновано цілісну технологічну схему формування мережевих моделей предметних галузей на основі текстових корпусі» варто було б включити в пункт новизни, пов'язаний зі створенням відповідного методу.

8. У викладенні пункту 10 результатів «вперше запропоновано методику порівняння текстових документів, що базується на побудові та порівнянні відповідних їм семантичних мереж, та на основі цієї методики запропонована модель середовища інформаційного пошуку та модель ранжування як окремих документів, так і джерел інформації.» помилково порушено причинно-наслідковий зв'язок, проте в самому описі отриманого результату цього не спостерігається.

9. У вступі деякі задачі дисертаційного дослідження у пункті «Мета і задачі дослідження» можна було об’єднати.

10. Подекуди в тексті дисертації зустрічаються стилістичні неточності, відступи від правил оформлення, форматування. Хоча вони і не впливають на можливість правильного загального сприйняття викладеного матеріалу, все ж спонукають до побажання більш чіткого і акуратного оформлення, вичитування тексту дисертації.

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», затверджені 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 опонент:

Головний науковий співробітник
Центрального науково-дослідного
інституту озброєння та військової техніки
Збройних Сил України,
доктор технічних наук, професор Олександр СТРИЖАК

5.04.24

Підпис Стрижака Олександра Євгенійовича
засвідчую:

Начальник відділу персоналу та стройового
Центрального науково-дослідного
інституту озброєння та військової техніки
Збройних Сил України Євген НОВОЖЕНІН