ВІДГУК

офіційного опонента про дисертаційну роботу Ковиліна Єгора Романовича «Модель генерації відповідей в пошукових системах на основі неструктурованої бази знань», подану на здобуття наукового ступеня кандидата технічних наук за спеціальністю 01.05.02 – математичне моделювання та обчислювальні методи

Актуальність теми дисертації. Дисертаційна робота Ковиліна Єгора Романовича «Модель генерації відповідей в пошукових системах на основі неструктурованої бази знань» присвячена розв'язку актуальної науково-практичної задачі побудови моделі отримання знань із неструктурованих джерел у пошукових системах.

Актуальність роботи пояснюється недосконалістю процесу пошуку інформації і відсутністю цілком автоматичних моделей представлення слов'яномовних текстів. Автор зазначає, що якщо в функціоналі пошукової системи присутня необхідність працювати з користувацьким запитом, який складається з декількох неформальних критеріїв і вимагає певного семантичного аналізу, то обробка отриманих результатів повністю лягає на плечі користувача. Одним із вирішень цієї проблеми є використання процесу автоматичної генерації знань, який дозволить відразу отримати релевантні до запиту користувача знання на основі усіх знань, що містяться у системі, та вдосконалити таким чином процес пошуку даних для людини-оператора. Аналіз науково-технічної літератури показав, що існуючі моделі представлення слов'яномовних текстів, які необхідно застосувати для автоматичної генерації, спрямовані саме на опис структурованих знань і потребують залучення значної кількості лінгвістичних знань і використання попередньої семантичної розмітки, створеної лінгвістом-експертом вручну, в той же час відповідні зарубіжні моделі орієнтовані в першу чергу на обробку англомовних текстів і не можуть бути застосовані для представлення слов'яномовних текстів. Таким чином, науково-технічна задача розробки моделі генерації відповідей на основі неструктурованої бази знань в пошукових системах є актуальною.

Про актуальність роботи також свідчить те, що робота виконана відповідно до закону України «Про пріоритетні напрями розвитку науки і техніки» (Відомості Верховної Ради України (ВВР), 2001, № 48, ст.253), і стосується напряму «інформаційні та комунікаційні технології» (стаття 3).

Обраний напрямок досліджень пов'язаний із виконанням дослідних робіт кафедри комп'ютерних наук та інформаційних технологій Дніпровського національного університету імені Олеся Гончара «Методи та інформаційні технології цифрової обробки багатоканальних даних» (реєстраційний номер 0116U001297).

Структура, задачі та короткий зміст роботи. Дисертація складається із анотації двома мовами, змісту, вступу, трьох розділів, висновків, списку використаних джерел і дев'ятнадцяти додатків.

Метою дисертації є розробка моделі обробки семантично-неструктурованих документів для генерації відповідей в пошукових системах. Для досягнення мети були поставлені та вирішені наступні задачі: визначено концепції моделі вилучення інформації та архітектури системи, розроблено математичну модель представлення семантичних властивостей текстів; розроблену модель забезпечено можливістю використовувати семантично нерозмічений заздалегідь корпус текстів; розроблено математичну модель валідації текстів-кандидатів до включення у неструктуровану базу знань за ступенем їх семантичної зв'язності; розроблено математичну модель генерації текстів, її алгоритмічне та програмне забезпечення; розроблено і застосовано систему оцінок адекватності створеної моделі.

У *вступі* обґрунтовано актуальність обраної теми, її зв'язок з науковими програмами, визначено мету та задачі дослідження. Визначено та розкрито предмет, об'єкт і межі дослідження, висвітлені методи, які використані в процесі роботи. Розкрита наукова новизна та практичне значення одержаних результатів.

У *першому розділі* проаналізовано існуючі математичні моделі текстів та існуючі аналоги моделі генерації текстів на основі неструктурованої бази знань у пошуковій системі. Здійснено і обґрунтовано вибір базової концепції моделі та зазначені основні складнощі та наукові задачі, які пов'язані із її використанням у пошукових системах. Сформульована модель генерації відповідей із неструктурованої бази знань та основні кроки її роботи. Введено поняття «текстовий автомат» та визначені основні вимоги до функціонування розробленої пошукової моделі.

У *другому розділі* розкриті основні властивості семантичних мереж, критерії та алгоритм побудови семантичної моделі неструктурованого тексту, наведені результати автоматичної побудови семантичної моделі документа та проведена перевірка адекватності розробленого підходу.

У *третьому розділі* розроблено модель пошукової системи, наведено опис головних сучасних теоретичних парадигм реалізації пошукових моделей «запит-відповідь», розкриті структура і алгоритм роботи створеної моделі пошукової системи із використанням генерації текстів, показані архітектури програмного додатку і створеної неструктурованої бази знань, а також проведене тестування моделі генерації відповідей на основі неструктурованої бази знань в пошукових системах і перевірена її адекватність.

У *висновках* висвітлені основні наукові та практичні результати, які відповідно до поставленої мети дають вирішення актуальної задачі побудови моделі отримання знань із неструктурованих джерел.

Основні наукові результати досліджень і наукова новизна.

– вперше розроблено семантичну модель текстових даних, яка дозволяє отримувати кількісні показники семантичних властивостей і сенсові зв'язки між компонентами тексту без необхідності залучення лінгвістичних знань;

– вперше створено модель автоматичної класифікації знань за ступенем їх семантичної зв'язності, що використовує числові дані, отримані із розробленої семантичної моделі тексту;

– вперше побудовано модель автоматичної генерації відповідей у пошуковій системі із неструктурованої бази текстових знань яка дозволила автоматизувати роботу користувача із пошуковими системами;

– вперше розроблено систему оцінок адекватності створеної моделі генерації відповіді на основі неструктурованої бази знань;

– отримали подальший розвиток семантичні моделі текстових даних для флективно багатих мов із вільним порядком слів та методи організації пошуку інформації: створені моделі дозволяють генерувати релевантні до запиту користувача знання на основі неструктурованої бази знань.

Обґрунтованість та достовірність основних висновків і результатів.

Одержані здобувачем результати є достовірними та об'єктивними, що підтверджується зіставленням сучасних наукових і технічних досягнень в області моделювання мови і розробки алгоритмів та моделей автоматичної генерації текстів; базуються на коректному застосуванні відомих методів частотного аналізу текстів, факторного латентно-семантичного аналізу, кластерного аналізу, індукційного і дедуктивного аналізу, вимірювання, експерименту, гіпотези, припущення, комп'ютерного моделювання отриманих результатів, теорії множин і штучного інтелекту; апробацією

основних теоретичних і експериментальних результатів роботи в друкованих працях та доповідях на конференціях та наукових семінарах.

Розроблені моделі покладені в основу програмних засобів для генерації відповідей в пошукових системах на основі неструктурованої бази знань. Автором розроблено і виконано оцінки адекватності моделей та отриманих результатів. Результати проведених оцінок вказують на адекватність та обґрунтованість отриманих у роботі результатів та висунутих наукових положень.

Наукове та практичне значення роботи.

Розроблені в роботі моделі, алгоритмічні та програмні засоби, утворюють єдину теоретико-практичну основу для автоматичного отримання семантичних характеристик тексту для генерації відповідей на основі семантично неструктурованої бази знань, 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, обчислювальних методів, призначених для використання при всебічному дослідженні і створенні об'єктів та систем технічного призначення або створення нових апаратних чи апаратно-програмних засобів моделювання й обчислення. Робота містить такі напрями дослідження, визначені паспортом спеціальності: удосконалення методів і засобів математичного та комп'ютерного моделювання призначених для використання при всебічному дослідженні і створенні нових апаратно-програмних засобів моделювання й обчислення; створення і дослідження нових обчислювальних методів і алгоритмів, що забезпечують створення ефективних програмних засобів комп'ютерної реалізації.

Практичне значення роботи полягає у тому, що автором розроблено математичну модель генерації відповідей в пошукових системах на основі неструктурованої бази знань та побудовано на її основі комп’ютерну систему, яка окрім організації зручного пошукового середовища, утворює універсальний програмний фреймворк з набором інструментів для проведення автоматичної обробки текстів на семантичному рівні, доступним для будь-якого користувача у вигляді окремої бібліотеки. Задачі дисертаційної роботи є важливими і нагальними питаннями галузі цифрового моделювання текстів, вирішення яких дозволяє гнучко створювати і обробляти тематичні повнотекстові бази знань без попередньої семантичної розмітки і будувати

програмну модель текстових знань формалізованої стильової спрямованості із кількісними характеристиками семантичних властивостей, на основі яких можливо вирішувати інші завдання автоматичної обробки текстів без необхідності залучати будь-які лінгвістичні знання.

Розроблена комп'ютерна система впроваджена у міський комунальний заклад культури «Централізована система бібліотек для дітей» м. Дніпро як пошуковий інструмент обробки електронних текстів, у ТОВ «Сітал Україна» як засіб автоматичної генерації текстових інструкцій та у АТ «ДніпроАзот» як інструмент покращення процесів пошуку в системах електронного документообігу.

Оформлення дисертації та автореферату.

Дисертаційна робота написана зрозуміло і грамотно, науково-технічна термінологія використовується коректно. Структура та обсяг дисертації відповідають встановленим вимогам на здобуття наукового ступеню кандидата технічних наук. Дисертацію написано з використанням сучасної бібліографії та наукової термінології. Зміст та результати досліджень викладено лаконічно та аргументовано. Стиль викладення матеріалів дисертації та автореферату логічний. Зміст автореферату повністю відповідає основним положенням та висновкам, зробленим у дисертації. Дисертаційна робота відповідає паспорту спеціальності 01.05.02 – математичне моделювання та обчислювальні методи.

Повнота викладення результатів дисертації в опублікованих працях.

Результати дисертаційної роботи опубліковані в 14 наукових працях, в тому числі 8 статей у журналах, рекомендованих МОН України для публікації результатів дисертацій, та закордонних виданнях; у тезах доповідей та трудах міжнародних та всеукраїнських конференцій – 6 (у тому числі 1 – у НМБ Scopus).

Зауваження.

Серед недоліків дисертації слід зазначити такі.

1. При описі результатів досліджень автор використовує швидше оповідальний, в хронологічному порядку, стиль викладу, ніж власне науковий. Це без сумніву, викликає інтерес, але ускладнює розуміння.

2. Основним недоліком дисертаційної роботи є слабка формалізація і неоднозначність визначень і характеристик, як використовуваних раніше, так і введених автором.

2.1 Так, наприклад, одне з ключових положень роботи, яке автор вводить як певну гіпотезу, а потім її експериментально перевіряє, не є сформульованим однозначно, а має декілька не формалізованих формулювань. Так, наприклад, на сторінці 38 сказано: «Описаний підхід базується на висунутій гіпотезі, що семантична мітка із найбільшою сумарною вагою зв'язків із множинами речень має найбільшу концентрацію термінів, що відносяться до тематики текстового знання і має найбільшу семантичну силу для подальшого аналізу при генерації відповіді, а терміни, що не входять в предметну область, опиняться в семантичних мітках із меншою сумарною вагою зв'язків із множинами речень»; на сторінці 86: « З цього витикає гіпотеза, що якщо система орієнтується не тільки на синтаксичний і частотний, а й на семантичний рівень розуміння, то у процесі роботи із згенерованим текстом повинні відбутися зміни кількісних характеристик семантичних властивостей створеної моделі документа», на сторінці 91 : «Гіпотеза, сформульована у першому розділі, має на увазі, що утворені семантичні мітки документа мають найбільшу кількість перетинів із семантичними контурами тоді, коли вони містять найбільшу кількість семантично значимих стем у своєму складі».

У той же час треба відзначити, що в роботі представлені всі необхідні математичні викладки для формалізації даної гіпотези і якби вона була коректно сформульована, це набагато б покращило виклад і розуміння суті роботи.

2.2 У роботі немає чітких визначень розроблених автором моделей, таких як семантична модель текстових даних, модель автоматичної класифікації знань, модель автоматичної генерації відповідей та інших.

2.3 Характерним прикладом неоднозначності визначень є опис поняття репрезентативності. На сторінці 111 зазначено, що «Однією з найважливіших характеристик корпусу текстів є репрезентативність. Оскільки корпус – це своєрідна модель мови, його репрезентативність визначає достовірність отриманих на його основі даних». Однак на сторінці 113 вказано: «Під репрезентативністю розуміється складання корпусу таким чином, щоб максимально задовільнити виконання усіх можливих сторін функціонування системи».

3. В зв'язку з тим, що автор представив результати в порядку хронології, ряд положень оглядового характеру і завдання досліджень

знаходяться у другому і третьому розділах, хоча доцільніше розмістити їх у першому. В той же час наведені в першому розділі отримані результати досліджень повинні бути розташовані в відповідних їм розділах.

4. Другий розділ дисертаційної роботи складається з двох частин, кожна з яких могла бути окремим розділом. Усередині цих підрозділів текст не має розбиття на частини, не має підзаголовків, наприклад, при описі рівнів семантичної моделі. Висновки за кожним розділом також є наскрізним текстом. Все перераховане робить текст дисертації погано читабельним. Це значно ускладнює розуміння.

5. У роботі широко використовуються методи машинного навчання, такі як байєсівський класифікатор, нейронна мережа, алгоритм кластеризації k-means. Однак в роботі не представлено коректного і повного опису проведених експериментів, а саме: опису навчальної та тестової вибірки, кількості об'єктів в вибірці, кількості класів, співвідношення числа об'єктів в кожному класі та інше.

6. У роботі допущений ряд некоректних описів результатів. Наприклад, на рис. 3.17 наведено фрагмент графіку значень експертних оцінок системи, де по осі ординат представлені значення оцінок від 0 до 5 і візуально можна побачити середнє значення близько до 4. У той же час нижче сказано «Середнє значення усіх проведених експертних оцінок склалось 0,839, що вказує на задовільні результати роботи системи ». На сторінці 98 зазначено, що «отримані результати розрахунку сенсової місткості для кожного тематичного набору текстів зображені на рис. 2.21 – 2.24» і самі рисунки по осі ординат підписані як «місткість». Однак саме поняття сенсової місткості не формалізоване і не має кількісних характеристик. Вочевидь автор використовував коефіцієнт середньої семантічної ємності (формула 2.15).

7. У представленому на рис. 2.24. розподілі значень сенсової місткості для тематики «інформаційні технології» значення місткості для слабких і сильних семантичних кластерів у деяких текстів дуже близькі. Було б доцільним пояснити, з чим це пов'язано.

8. В тексті дисертації і автореферату зустрічаються некоректності як математичного, так і синтаксичного характеру. Так, наприклад, в авторефераті у формулі 5 не вказано, що матриця V є транспонованою, формула 12 позначена посиланням (14) та інше.

Наведені зауваження не знижують високий науковий рівень дисертаційної роботи і не впливають на її загальну позитивну оцінку.

Висновок. Розглянувши дисертаційну роботу Ковиліна Єгора Романовича «Модель генерації відповідей в пошукових системах на основі неструктурованої бази знань», автореферат, опубліковані наукові праці та додаткові матеріали, можна зробити такі висновки:

– дисертація відповідає паспорту спеціальності 01.05.02 – Математичне моделювання та обчислювальні методи;

– тематична спрямованість роботи є актуальною, суспільно корисною й перспективною у плані продовження розпочатих досліджень;

– зміст автореферату відповідає основним положенням дисертаційної роботи;

– дисертація є цілісною, завершеною, оригінальною, самостійною кваліфікаційною науковою працею.

Вважаю, що дисертаційна робота є завершеною науково-дослідницькою роботою, в якій отримано нові науково обґрунтовані результати, що в сукупності вирішують актуальну наукову-технічну задачу обробки семантично-неструктурованих документів для генерації відповідей в пошукових системах, та відповідає вимогам п. п. 9, 11, 12 «Порядку присудження наукових ступенів», які висуваються до кандидатських дисертацій, а її автор, Ковилін Єгор Романович, заслуговує на присудження наукового ступеня кандидата технічних наук за спеціальністю 01.05.02 – математичне моделювання та обчислювальні методи.

Офіційний опонент
професор кафедри
прикладної математики
Харківського національного
університету радіоелектроніки,
доктор технічних наук, професор Кіріченко Л. О.

Підпис проф. Кіріченко Л. О. засвідчую.

Міністерство освіти і науки України * Харківський національний університет радіоелектроніки * №02071187 *

Учений Секретар
Харківського національного
університету радіоелектроніки І.В. Магдаліна

«21» 10 2020 р.