ВІДГУК

офіційного опонента доктора технічних наук, професора
Хайрової Ніни Феліксівни
на дисертаційну роботу Яхимовича Олександра Вікторовича
«Інформаційна технологія пошуку ключових слів на основі парсингу англомовних текстів»,
поданої на здобуття наукового ступеня кандидата технічних наук
за спеціальністю 05.13.06 – інформаційні технології

Актуальність теми дисертаційної роботи

Дисертаційна робота Яхимовича Олександра Вікторовича, присвячена одній з важливих і актуальних задач – пошуку ключових слів в англомовному тексті. В даний час, кількість корисного електронного контенту в мережах, як локальних, так і глобальній має тенденцію збільшуватися у геометричній прогресії. Для забезпечення достовірності видачі результатів на пошукові запити необхідно вдало знаходити ключові слова з природномовного контенту.

Тому розробка інформаційної технології пошуку ключових слів на основі парсингу англомовних текстів, що базується на використанні додаткової інформації універсального характеру про складні залежності між членами речення, є без сумніву актуальною і важливою у сучасному світі.

Обрана тема та мета дисертації, що спрямовані на підвищення точності процесу пошуку ключових слів для англомовних текстів, виявляють свою актуальність та перспективність.

У дисертації наведено теоретичні узагальнення і нові вирішення наукової задачі розробки інформаційної технології пошуку ключових слів на основі парсингу англомовних текстів, а також відповідних програмних засобів її реалізації.

Таким чином, тематика дисертаційної роботи є актуальною та важливою для сучасного рівня науки й техніки.

Наукова новизна отриманих результатів

До основних наукових результатів, отриманих здобувачем особисто, належить:

Удосконалено модель пошуку ключових слів, яка, на відміну від існуючих, побудована на основі інформаційної оцінки результатів парсингу тексту та враховує результати аналізу зв'язків між лексичними одиницями тексту, що дозволило формалізувати критерій якості процесу пошуку ключових слів.

Уперше розроблено метод пошуку ключових слів, який, на відміну від існуючих, базується на знаходженні синтаксичних зв'язків між словоформами у реченнях англомовного тексту за допомогою технологічних можливостей парсингу сучасних лінгвістичних пакетів. Запропонований

метод дає змогу підвищити чисельні характеристики якості пошуку ключових слів, а саме повноту (за Жаккаром) і точність.

Удосконалено метод зменшення впливу вербального шуму на пошук ключових слів, який, на відміну від існуючих, побудовано на основі стенфордської класифікації зв'язків між лексичними одиницями речення, що дозволило підвищити якість результатів пошуку ключових слів у порівнянні з основним методом.

Набула подальшого розвитку інформаційна технологія пошуку ключових слів, яка, на відміну від існуючих, враховує додаткову інформацію процесів парсингу речень у межах послідовного застосування двох запропонованих методів, що дозволило уточнити чисельні оцінки змістовних параметрів тексту та підвищити якість пошуку його ключових слів.

Зв'язок роботи з науковими програмами, темами, планами

Актуальність теми підтверджується тим, що дисертаційне дослідження проводилось згідно з планами науково-дослідних робіт кафедри автоматизації та інтелектуальних інформаційних технологій Вінницького національного технічного університету, в тому числі в межах: «Інтелектуальна інформаційна технологія образного аналізу тексту та синтезу інтегрованої бази знань природно-мовного контенту», № ДР 0114U003462, при виконанні якого автор брала участь як виконавець окремих підрозділів; «Ідентифікація прихованих залежностей в онлайнових соціальних мережах на основі методів нечіткої логіки та комп'ютерної лінгвістики», № ДР 0117U000575, при виконанні якої автор брала участь як виконавець окремих підрозділів.

Ступінь обґрунтованості наукових положень, висновків і рекомендацій, сформульованих у дисертаційній роботі

Наукові положення, результати і висновки дисертації, отримані автором, в цілому, є достатньо доказовими та обґрунтованими.

Для розробки і оцінки інформаційної технології пошуку ключових слів на основі парсингу англомовних текстів, в роботі використані: методи теорії множин та теорії інформації, методи лінгвістичного аналізу англомовного тексту, статистичних методи, методи вимірювання у метричних просторах. Основні наукові положення дисертаційної роботи представлено моделями, методами та алгоритмами. Наведені в роботі теоретичні положення та твердження викладено у логічній послідовності та в достатній мірі аргументовано. Адекватність запропонованих методів, моделей, алгоритмів підтверджується результатами експериментальних досліджень.

Достовірність отриманих результатів забезпечено коректністю постановки завдання та формалізації математичної моделі пошуку ключових слів (розділ 2), здійсненому аналізу та класифікації зв'язків між лексичними одиницями тексту (розділ 2), розробці загального підходу до пошуку ключових слів з урахуванням зв'язків між членами речення та зменшенням

впливу вербального шуму (розділ 3), розробці відповідних методів, моделей та алгоритмів (розділ 2, 3), що стали основою розробленої інформаційної технології пошуку ключових слів з використанням інформації про синтаксичні зв'язки між словоформами у реченнях англомовного тексту (розділ 4); а також, здійсненого аналізу отриманих результатів і обчислення чисельних характеристик якості пошуку ключових слів, а саме повноту і точність (розділ 4).

Достовірність наукових положень, висновків і рекомендацій, що сформульовані в дисертаційній роботі, підтверджується апробацією та впровадженням результатів досліджень на реальному об'єкті професійної сфери діяльності, про що свідчать відповідні документи.

Практичні результати роботи, їх рівень та ступінь впровадження

Практичне значення отриманих результатів роботи полягає у наступному: формальному описі методики пошуку ключових слів англомовного тексту, створенні алгоритму її реалізації та розробці програмного забезпечення, що знаходить ключові слова на основі врахування значимих зв'язків між словоформами у реченнях англомовного тексту та подальшої фільтрації вербального шуму. Створені моделі, алгоритми та програмні засоби можуть бути використані при вирішенні практичних задач комп'ютерної лінгвістики, які потребують знаходження ключових слів, наприклад, для підвищення точності аналізу контенту сайту і підняття позиції сайту в результатах пошуку. Використання мово-незалежних засобів запропонованої інформаційної технології у поєднанні з необхідними технологічними ресурсами лінгвістичного аналізу інших природних мов дозволить розширити область застосування інформаційної технології, зокрема на українську мову.

Робота впроваджена на ТОВ НВП «СПІЛЬНА СПРАВА» (акт про результати впровадження від 10.01.2020), а також в навчальний процес кафедри автоматизації та інтелектуальних інформаційних технологій Вінницького національного технічного університету, що підтверджено виданням навчального посібника "A lexical relationships-based keywords selection in an English text".

Повнота викладення результатів досліджень в опублікованих працях

За темою дисертації з викладенням її основних результатів опубліковано 21 наукову працю, серед яких 1 стаття у закордонному фаховому періодичному виданні, що входить до SCOPUS, 1 стаття у вітчизняному періодичному виданні, що індексується у SCOPUS, 6 статей у спеціалізованих фахових виданнях України, що індексуються міжнародними бібліометричними та наукометричними базами даних, 8 публікацій в матеріалах та тезах доповідей, 2 звіти про науково-дослідну роботу. Також, отримано 1 патент України на корисну модель (Пат. 135223 UA), опубліковано закордонну монографію та електронний навчальний посібник.

Рівень і кількість публікацій та апробації матеріалів дисертації розкривають основний зміст дисертації та повністю відповідають вимогам МОН України.

Оцінка основного змісту дисертації та її структури

Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел і додатків! Основний зміст викладено на 138 сторінках друкованого тексту, містить 67 рисунки, 17 таблиць. Список використаних джерел є достатнім, охоплює сучасні вітчизняні та зарубіжні публікації, містить 157 найменувань.

Оформлення дисертаційної роботи відповідає встановленим вимогам.

У вступі до дисертації обґрунтовано актуальність теми, сформульовано мету і завдання досліджень, викладено наукову новизну та практичне значення результатів роботи, особистий внесок здобувача, наведені дані щодо апробації результатів досліджень.

У першому розділі на основі огляду та аналізу сучасної літератури в галузі пошуку ключових слів показано, що ключові слова є основним засобом упорядкування важливого масиву інформації, яким є науковий журнал або база наукових статей певного наукового напряму. Зокрема, ключові слова потрібні для систематизації множини статей, оскільки дозволяють швидше знайти статтю, групувати схожі статті, класифікувати статті у межах інших структурних групувань. З іншого боку, на основі застосування ключових слів реалізовано велику кількість методів пошуку, класифікації та оцінки інформації. Показано, що існує велика кількість доступних систем автоматичного пошуку ключових слів, розроблених і орієнтованих на обробку природних мов. Ці системи засновані на значній кількості відомих методів пошуку ключових слів, які діляться на експертно-лінгвістичні та статистичні.

У другому розділі за результатами математичного моделювання формалізовано задачу пошуку ключових слів тексту як параметричну ідентифікацію функції згортки вербальної інформації за критерієм максимума інформації у зв'язках, які поєднують n обраних ключових слів тексту T між собою та з усіма m' значущими словами цього тексту.

На основі проведеної інформаційної оцінки результатів парсингу тексту отримано формальні межі лінійного збільшення кількості інформації для значущих слів тексту, з яких обираються ключові. На відміну від існуючих моделей, така інформація враховується запропонованим підходом до пошуку ключових слів тексту.

У третьому розділі розроблено загальний підхід до пошуку ключових слів з урахуванням зв'язків між членами речення. Підхід реалізовано на базі двох методів, що забезпечують такі технологічні етапи: створення багаторівневої розмітки тексту; застосування синтаксичної розмітки, що враховує складні залежності між парами лем; зменшення вербального шуму;

вибір перших n слів з найбільшою кількістю зв'язків, де n – кількість потрібних ключових слів.

Розроблено алгоритм процесу пошуку ключових слів, алгоритм фільтрації вербального шуму та алгоритм аналізу тексту, який є базовим для двох взаємопов'язаних методів пошуку ключових слів, що пропонуються.

Представлено UML діаграму класів та діаграму зв'язків між класами, схеми алгоритмів аналізу тексту і загального підходу до пошуку ключових слів. Описано за що відповідають основні класи, розробленого програмного забезпечення, і як вони взаємодіють між собою.

У четвертому розділі розроблено інформаційну технологію пошуку ключових слів та її структуру, яка забезпечує послідовне застосування основного та додаткового методів, обгрунтованих у розділі 3. Структуру інформаційної технології розроблено за модульним принципом з метою подальшого розвитку програмного забезпечення шляхом інтеграції нових модулів в існуючу архітектуру, розширюючи таким чином функціонал системи.

Проведено серію експериментів для визначення кількісних характеристик достовірності отриманих результатів пошуку ключових слів. Аналіз результатів експерименту показав суттєві переваги запропонованої інформаційної технології за кількісними характеристиками порівняно з аналогами, якими є сайти SEO оптимізації, де є функція пошуку ключових слів. Ще однією суттєвою перевагою в порівнянні з аналогами є те, що запропонована інформаційна технологія пошуку ключових слів дозволяє повністю (до 100%) виключити шумові слова.

Проведено масштабні експериментальні дослідження, що полягають у пошуку ключових слів для дуже коротких англомовних текстів. Експериментальну базу склали близько 30 000 відібраних коротких анотацій до фільмів, що супроводжувалися заданими ключовими словами та були розбиті на категорії за жанрами. Запропонована інформаційна технологія знаходить одне і більше ключових слів, заданих автором для 42.3% опрацьованих анотацій. При цьому 51.2% усіх позитивних результатів отримано з немалою точністю від 0.2 до 0.55, а 41.1% позитивних результатів отримано з повнотою в діапазоні від 0.12 до 0.37.

У висновках наведено основні результати дисертаційної роботи та надано рекомендації щодо практичного застосування теоретичних напрацювань. Загальні висновки по роботі відрізняються чіткістю, лаконічністю, узагальнюють викладені в роботі результати досліджень.

Відповідність дисертації та автореферату встановленим вимогам

За своєю структурою, обсягом і оформленням дисертація та автореферат цілком відповідають вимогам, встановленим до кандидатських дисертацій, зокрема пп. 9, 11, 12 «Порядку присудження наукових ступенів».

Автореферат за змістом ідентичний основним положенням, що викладені в дисертації, та не містить інформації, яка не відображена в самій

роботі. Стиль викладу матеріалів досліджень, наукових положень і рекомендацій забезпечує їх адекватне і належне сприйняття.

Наукова новизна відповідає паспорту спеціальності 05.13.06 – інформаційні технології, зокрема розділ 2 «Розробка математичної моделі пошуку ключових слів», що включає в себе 2.1 «Формалізація задачі пошуку ключових слів тексту», 2.2 «Інформаційна оцінка парсингу тексту для задачі пошуку ключових слів», розділ 3 «Розробка методів пошуку ключових слів», що включає в себе 3.1 «Концептуальні основи розробки методів пошуку ключових слів у тексті», 3.3 «Алгоритм пошуку ключових слів на основі двох послідовних методів», 3.4 «Спосіб автоматичного пошуку ключових слів з використанням технології DKPro Core», розділ 4 «Розробка та апробація інформаційної технології», що включає в себе 4.2 «Структура інформаційної технології».

Недоліки та зауваження до дисертаційної роботи

До недоліків роботи, на мою думку, слід віднести:

1. Автор застосував назву для п.1.1 Загальна характеристика проблеми, хоча наукове завдання дослідження значно компактніше, а саме підвищення якості пошуку ключових слів у англомовному тексті.

2. Математична модель, що покладена автором в основу запропонованого методу пошуку ключових слів, має декларативний характер. Процес знаходження синтаксичних зв'язків між словоформами у реченнях англомовного тексту забезпечується на технологічному рівні, але було б краще формалізувати цей процес імперативною частиною математичної моделі.

3. Оскільки для побудови моделі пошуку ключових слів використано методи теорії множин, доцільно було порівняти запропоновану модель з моделлю bag-of-words, за якою текст також представляється у вигляді мультимножини.

4. В таблиці 2.1 приклади наведено для української мови, але наразі запропонована інформаційна технологія знаходить ключові слова тільки для англомовних текстів, тому краще було взяти речення на англійській мові.

5. Пункт 2.3 містить огляд універсальних залежностей, але приклади наведені тільки англійською мовою, було б непогано додати хоча б по одному прикладу на інших мовах.

6. Загалом масштабні експериментальні дослідженні в роботі були проведені для різних типів англомовних текстів - від надкоротких (з одного речення) до великих літературних творів (роман "Мобі Дік"). Результати експериментів показали підвищення якості пошуку ключових слів за допомогою розробленої інформаційної технології. Проте для підтвердження достовірності гіпотези щодо чисельних оцінок отриманого підвищення якості було варто провести експеримент для 300 текстів одного типу.

7. На жаль, робота не вільна від стилістичних недоліків та орфографічних помилок.

Відмічені зауваження не вплинули на загальну позитивну оцінку дисертаційної роботи та можуть розглядатися як рекомендації до подальших наукових досліджень та впроваджень отриманих результатів.

Загальні висновки

Дисертаційна робота Яхимовича Олександра Вікторовича на тему: «Інформаційна технологія пошуку ключових слів на основі парсингу англомовних текстів» є завершеною науковою працею, яка розв'язує актуальну наукову задачу підвищення точності процесу пошуку ключових слів в англомовних текстах.

Автореферат повністю відповідає змісту дисертації і описує суть одержаних результатів та висновків у дисертаційній роботі і оформлений згідно з вимогами.

Дисертаційна робота відповідає спеціальності 05.13.06 – інформаційні технології вимогам ДАК України, зокрема пп. 9, 11, 12 «Порядку присудження наукових ступенів», затвердженого постановою Кабінету Міністрів України від 24 липня 2013 року № 567 (зі змінами затвердженими постановою Кабінету Міністрів України від 15 липня 2020 р. № 607), які висуваються до робіт на здобуття наукового ступеня кандидата технічних наук, так як вони містять нові науково обґрунтовані результати проведених досліджень.

Автор дисертації Яхимович Олександр Вікторович заслуговує на присудження йому наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології.

Офіційний опонент,
професор кафедри
інтелектуальних комп'ютерних систем,
Національного технічного університету
«Харківський політехнічний інститут»
доктор технічних наук, професор **Ніна ХАЙРОВА**

Підпис офіційного опонента – доктора технічних наук, професора Хайрової Ніни Феліксівни засвідчую

Вчений секретар НТУ "ХПІ" проф. **Олександр ЗАКОВОРОТНИЙ**