**ВІДГУК**

офіційного опонента на дисертаційну роботу
Яхимовича Олександра Вікторовича, подану на здобуття
наукового ступеня кандидата технічних наук
за спеціальністю 05.13.06 – інформаційні технології, на тему
**«Інформаційна технологія пошуку ключових слів на основі парсингу англомовних текстів»**

**Актуальність теми дисертації.** Надзвичайна популярність пошукових машин висуває високі вимоги до релевантності результатів пошуку, яка залежить від якості процесів знаходження ключових слів у текстовій інформації.

У відомих методах пошуку традиційно наголос робиться на аналізі статистики слів тексту, але при цьому не враховується вплив зв'язків між словами у реченнях, зокрема синтаксичних.

Це зумовлює необхідність розробки нових методів та систем, які на основі додаткової інформації про текст і його складові здійснюють пошук ключових слів. Саме така задача вирішується в дисертаційній роботі Яхимовича О. В., яка присвячена розробці методу, моделі та інформаційної технології пошуку ключових слів на основі парсингу англомовних текстів з використанням додаткової інформації універсального характеру про складні залежності між членами речення, а також дозволяє зменшити вплив вербального шуму.

Актуальність теми дисертації підтверджується також і тим, що вона виконувалась у рамках науково-дослідних робіт Вінницького національного технічного університету.

**Зв'язок роботи із науковими програмами, темами.** Дослідження, результати яких представлено в дисертації, проводились відповідно до пріоритетних тематичних напрямів науково-технічних розробок на період до 2020 року «Технології та засоби розробки програмних продуктів і систем», затверджених постановою Кабінету Міністрів України №556 від 23.08.2016 р. Зокрема, дослідження проводились згідно планів науково-дослідної роботи кафедри автоматизації та інтелектуальних інформаційних технологій Вінницького національного технічного університету. Автор брав участь у виконанні науково-дослідних робіт «Інтелектуальна інформаційна технологія образного аналізу тексту та синтезу інтегрованої бази знань природно-

мовного контенту» (№ ДР 0114U003462) та «Ідентифікація прихованих залежностей в онлайнових соціальних мережах на основі методів нечіткої логіки та комп'ютерної лінгвістики» (№ ДР 0117U000575).

**Наукова новизна отриманих результатів** полягає в тому, що:

1. 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.

2. 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.

3. 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.

4. 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.

Наукові положення, сформульовані в дисертації, досить повно обґрунтовані. Кожен пункт наукової новизни достатньою мірою підтверджений теоретичними, а також експериментальними дослідженнями.

**Ступінь обґрунтованості та достовірність наукових положень, висновків, рекомендацій, сформульованих у дисертації.** Обґрунтованість та достовірність наукових положень, висновків і рекомендацій забезпечується аргументованою постановкою мети й задач дослідження, повнотою формулювання умов, в яких вони розв’язуються та необхідними припущеннями і обмеженнями щодо застосування результатів, використанням сучасного математичного апарату та програмного

забезпечення. Теоретичні дослідження виконано з використанням методів теорії множин та теорії інформації. Розробка методів пошуку ключових слів та зменшення впливу вербального шуму базувалася на основі поєднання методів лінгвістичного аналізу англомовного тексту та статистичних методів.

Достовірність отриманих результатів підтверджується їх узгодженням із теоретичними висновками, експериментами та чисельними розрахунками, а також впровадженням розроблених моделей і методів у запропоновану інформаційну технологію.

**Значимість отриманих результатів для науки і практичного використання.** Цінність наукових результатів роботи полягає у 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.

Практична корисність роботи обумовлена тим, що створені моделі, алгоритми та програмні засоби можуть бути використані при вирішенні практичних задач комп'ютерної лінгвістики, які потребують знаходження ключових слів. Використання мово-незалежних засобів запропонованої інформаційної технології у поєднанні з необхідними, згідно з отриманою специфікацією, технологічними ресурсами лінгвістичного аналізу інших природних мов дозволить розширити область застосування інформаційної технології.

Робота впроваджена на ТОВ НВП «СПІЛЬНА СПРАВА» (м. Вінниця, 2020 р.), а також в навчальний процес кафедри АІІТ Вінницького національного технічного університету, що підтверджено виданням навчального посібника «A lexical relationships-based keywords selection in an English text». Результати експериментів показали, що запропонована інформаційна технологія одночасно збільшує у межах від 8,1% до 12,7% повноту за метрикою Жаккара та від 9,1% до 14,3% абсолютну точність пошуку ключових слів для англомовних текстів обсягом 140-1400 слів у порівнянні з аналогами.

**Повнота викладення в публікаціях та апробація роботи.** Основні наукові положення, висновки і рекомендації, які сформульовані в дисертаційній роботі, достатньо повно відображені в публікаціях здобувача і пройшли апробацію на міжнародних науково-технічних конференціях.

За темою дисертації опубліковано 21 наукову працю, в тому числі: 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.

**Структура дисертації** цілком відповідає логіці й послідовності рішення поставлених задач. Дисертація складається зі вступу, 4-х розділів, висновків, списку використаних джерел та додатків.

У вступі обґрунтовано актуальність теми дисертаційної роботи, зазначено зв’язок з науковими програмами, планами, темами, сформульовано мету та задачі, об’єкт та предмет дослідження, визначено наукову новизну та практичне значення одержаних результатів, наведено відомості про апробацію роботи, впровадження результатів та публікації.

**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 розглянуто довільний текст як множину синтаксично зв'язаних упорядкованих слів, які, в свою чергу, є підмножиною слів мови. За результатами математичного моделювання формалізовано задачу пошуку ключових слів тексту як параметричну ідентифікацію функції згортки вербальної інформації за критерієм максимуму 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, обґрунтованих 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 релевантності 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.

Список використаних джерел є достатнім, охоплює сучасні вітчизняні та зарубіжні публікації, містить 157 найменувань.

**Автореферат дисертації** ідентичний за змістом з основними положеннями дисертації і достатньо повно відображає основні наукові положення, практичну значимість і висновки. Дисертаційна робота та автореферат оформлені у відповідності з встановленими вимогами.

**Відповідність дисертації та автореферату встановленим вимогам.** Дисертація та автореферат дисертації за своєю структурою, об’ємом і оформленням відповідають вимогам, які встановлені до кандидатських дисертацій, зокрема пп. 9, 11, 12 «Порядку присудження наукових ступенів» (Постанова Кабінету міністрів України № 567, від 24 липня 2013 р.).

Автореферат дисертації за змістом відповідає основним положенням, які викладено в дисертації та не містить інформації, яка не відображена в основній роботі. Стиль викладення матеріалів досліджень та 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 п. 2.1 «Формалізація задачі пошуку ключових слів тексту», п. 2.3 «Аналіз зв'язків між лексичними одиницями тексту», розділ 3 «Розробка методів пошуку ключових слів», що включає в себе п. 3.1 «Концептуальні основи розробки методів пошуку ключових слів у тексті», п. 3.3 «Алгоритм пошуку ключових слів на основі двох послідовних методів», розділ 4 «Розробка та апробація інформаційної технології», що включає в себе п. 4.2 «Структура інформаційної технології», п. 4.5 «Визначення кількісних характеристик релевантності отриманих

результатів».

Матеріали дисертації викладені у чіткій логічній послідовності, на належному науковому рівні та повністю відповідають поставленій меті і задачам дослідження. Висновки роботи відповідають її змісту, обґрунтовані і підтверджені результатами дослідження.

**Недоліки та зауваження щодо змісту дисертації та автореферату:**

1. У пункті 1.7 «Вибір напрямку і обґрунтування задач досліджень» та інших місцях дисертації зазначено, що у відомих методах пошуку ключових слів не береться до уваги інформаційна оцінка результатів парсингу тексту. Потрібно було надати посилання на ті роботи автора, де обґрунтовується доцільність такої оцінки для побудови методу пошуку ключових слів.

2. У розділі 2 наведено вираз для функції згортки вербальної інформації в процесі пошуку ключових слів тексту (2.9), а далі описано зв'язки, які існують між словами в реченні. Не вистачає опису того, як формула (2.9) і зв'язки пов'язані між собою або прикладу обчислення для конкретного речення і його зв'язків.

3. У пункті 3.1 потрібно було описати методи поетапно, з відповідними посиланнями на модель і критерій з 2 розділу, а також навести переваги розроблених методів.

4. У тексті пункту 3.2 «Мово-незалежні особливості розроблених методів пошуку ключових слів тексту» немає згадок про наукову новизну.

5. З тексту роботи незрозуміло, до якого пункту новизни відноситься спосіб, описаний в пункті 3.4.

6. Незрозуміле відношення до розробки методів пошуку ключових слів в розділі 3 має програмно-апаратний комплекс, представлений на рисунку 3.8.

7. В авторефераті на сторінці 6 у формулі (1) символи індексації слів і знаків пунктуації збігаються. Краще було б використати різну індексацію, оскільки слова і знаки пунктуації відносяться до різних множин.

Однак зазначені зауваження не носять принциповий характер і не знижують цінності проведеного здобувачем дослідження, актуальності, новизни та практичної значущості дисертаційної роботи.

**Висновки, щодо відповідності дисертації встановленим вимогам.**

Дисертація є завершеною науковою роботою, в якій отримано нові науково-обґрунтовані теоретичні та експериментальні результати в галузі інформаційних технологій, що в сукупності вирішують актуальну науково-технічну задачу підвищення точності процесу пошуку ключових слів для

англомовних текстів шляхом розробки моделей, методів та алгоритмів пошуку ключових слів та зменшення кількості шумових слів, які покладені в основу запропонованої інформаційної технології.

Тематика та зміст дисертації відповідають паспорту спеціальності і профілю спеціалізованої вченої ради.

Вважаю, що представлена дисертаційна робота на тему «Інформаційна технологія пошуку ключових слів на основі парсингу англомовних текстів» за актуальністю обраної теми, обсягом та рівнем виконання теоретичних і експериментальних досліджень, достовірністю та обґрунтованістю висновків, новизною досліджень, значення для науки і практики відповідає вимогам пунктів 9, 11, 12 «Порядку присудження наукових ступенів», затвердженого постановою Кабінету Міністрів України від 24 липня 2013 року № 567, щодо кандидатських дисертацій, а здобувач Яхимович Олександр Вікторович заслуговує на присудження наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології.

Офіційний опонент,
професор кафедри
автоматизації та інформаційних систем,
Кременчуцького національного університету
імені Михайла Остроградського
доктор технічних наук, професор І. В. Шевченко