Вінницький національний технічний університет

Міністерство освіти і науки України

Кваліфікаційна наукова праця

на правах рукопису

ЯХИМОВИЧ ОЛЕКСАНДР ВІКТОРОВИЧ

УДК 004.891

ДИСЕРТАЦІЯ

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПОШУКУ КЛЮЧОВИХ СЛІВ НА ОСНОВІ ПАРСИНГУ АНГЛОМОВНИХ ТЕКСТІВ

05.13.06 – інформаційні технології

Технічні науки

Подається на здобуття наукового ступеня кандидата технічних наук

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____О. В. Яхимович

Науковий керівник: Бісікало Олег Володимирович

доктор технічних наук, професор

Вінниця – 2021

АНОТАЦІЯ

Яхимович О. В. Інформаційна технологія пошуку ключових слів на основі парсингу англомовних текстів. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 «Інформаційні технології». – Вінницький національний технічний університет, Вінниця, 2021.

Завдання пошуку ключових слів тексту виникає у бібліотечній справі, лексикографії та термінознавстві, а також в усіх задачах інформаційного пошуку. В даний час обсяги і динаміка інформації, яка підлягає обробці в цих областях, роблять особливо актуальною задачу автоматичного пошуку ключових слів, які можуть використовуватися для створення і розвитку термінологічних ресурсів, а також для ефективної обробки документів – індексування, реферування, кластеризації та класифікації. Популярність та затребуваність пошукових машин для кожного користувача Інтернету висуває високі вимоги до релевантності результатів пошуку, яка прямо пропорційна якості процесів знаходження ключових слів у текстовій інформації.

Існує значна кількість доступних систем автоматичного пошуку ключових слів, розроблених і орієнтованих на обробку природних мов. В основу роботи таких систем покладено методи пошуку ключових слів тексту, які умовно можна розділити на дві категорії – експертно-лінгвістичні та статистичні. Перша група методів ґрунтується на значеннях слів, отриманих експертним шляхом, зокрема використовують словники з зібраними семантичними даними про кожне слово або онтології предметних областей. Однак, лінгвістичні методи ресурсоємні на ранніх етапах. Тому найбільша кількість відомих напрацювань у напряму експертно-лінгвістичної обробки текстів відома для визнаної мови міжнародного спілкування – англійської. З іншого боку,

традиційні статистичні методи, які є фактично мово-незалежними, супроводжуються значними обсягами «вербального шуму», що суттєво впливає на якість пошуку ключових слів. Внаслідок цього статистичні методи зазвичай супроводжуються емпіричними процедурами, налаштованими на конкретний клас задач, що суттєво звужує їхню область застосування.

Підвищення якості процесу пошуку ключових слів вимагає залучення додаткової інформації, бажано універсального, а не специфічного характеру. Зокрема, відсутня формальна постановка та розв'язок задачі пошуку ключових слів як згортки інформації у тексті. Не враховуються у відомих методах пошуку ключових слів результати аналізу зв’язків між лексичними одиницями тексту, а інформаційна оцінка результатів парсингу тексту не береться до уваги як складова відповідного критерія якості.

Перспективними є гібридні методи пошуку ключових слів, для яких швидкість статистичної обробки тексту підсилюється можливостями сучасних лінгвістичних пакетів, що мають найбільш розвинутий функціонал, у першу чергу, для англійської мови. Тому актуальним науковим завданням є підвищення якості пошуку ключових слів у англомовному тексті шляхом розробки інформаційної технології пошуку ключових слів на основі парсингу англомовних текстів.

Мета дисертаційного дослідження полягає у підвищенні якості пошуку ключових слів у англомовному тексті.

Для досягнення вказаної мети в роботі розв’язуються такі основні задачі:

1. Аналіз й порівняльна характеристика відомих підходів, методів та засобів пошуку ключових слів.

2. Побудова математичної моделі процесу пошуку ключових слів на основі інформаційної оцінки результатів парсингу тексту.

3. Розробка методу пошуку ключових слів на основі визначення зв'язків між словоформами.

4. Розробка методу зменшення впливу вербального шуму на пошук ключових слів.

5. Експериментальне дослідження результатів запропонованих методів у порівнянні з аналогами.

6. Розробка та апробація інформаційної технології пошуку ключових слів англомовного тексту.

Науковою новизною виконаного дисертаційного дослідження визначено:

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.

Практична цінність отриманих в дисертації результатів 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 подальшою фільтрацією 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, згідно з отриманою специфікацією, технологічними ресурсами лінгвістичного аналізу інших природних мов дозволить розширити область застосування інформаційної технології, зокрема на українську мову.

Основні результати та дисертаційна робота в цілому апробовані на восьми науково-практичних конференціях:

- міжнародній Інтернет-конференції «Молодь в технічних науках: дослідження, проблеми, перспективи» (МТН-2015), м. Вінниця, 23-26 квітня 2015 р.;

- першій міжнародній конференції «Адаптивні технології управління навчанням», м. Одеса, 23-25 вересня 2015 р.;

- третій міжнародній науковій конференції «Вимірювання, контроль та діагностика в технічних системах» (ВКДТС-2015), м. Вінниця, 27-29 жовтня 2015 р.;

- XLIV, XLV, XLVI, XLVII та XLVIII науково-технічних конференціях підрозділів ВНТУ факультету комп'ютерних систем та автоматики, м. Вінниця, щорічно у 2015 – 2019 рр.

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 наукових досліджень кафедри автоматизації та інтелектуальних інформаційних технологій Вінницького національного технічного університету, а саме у науково-дослідних роботах «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).

Робота впроваджена на ТОВ НВП «СПІЛЬНА СПРАВА», 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” та актом про результати впровадження від 10.01.2020. 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.

Ключові слова: інформаційна технологія, ключові слова, вербальний шум, лінгвістичний пакет, DKPro Core, синтаксичний аналіз, словосполучення.

ABSTRACT

Yahimovich O. V. Information technology of searching keywords based on parsing English texts. – Qualification research paper, manuscript copyright.

Thesis for the degree of a candidate of technical sciences in specialty 05.13.06 «Information technology». – Vinnytsia National Technical University, Vinnytsia, 2021.

The task of searching keywords in the text arises in the library business, lexicography and terminology, as well as in all tasks of information retrieval. Currently, the amount and dynamics of information to be processed in these areas, make it especially important to the automation process of searching keywords in texts that can be used to create and develop terminological resources, as well as for efficient document processing - indexing, abstracting, clustering and classification. The extraordinary demand and popularity of search engines for every Internet user puts forward high demands on the relevance of search results, which is directly proportional to the quality of the technological process of searching keywords in textual information. Therefore, the topic of the thesis is relevant.

There are a lot of automatic searching keywords systems available, designed and focused on natural language processing. The work of such systems is based on methods of searching keywords in the text, which can be divided into two categories - expert-linguistic and statistical. The first group of methods is based on the meanings of words obtained by experts, in particular, use dictionaries with collected semantic data about each word or ontology of subject areas. However, linguistic methods are resource-intensive in the early stages. Therefore, the largest number of known developments in the direction of expert-linguistic processing of texts is known for the international communication language - English. On the other hand, traditional statistical methods, which are, in fact, language-independent, are accompanied by significant amounts of "verbal noise", which significantly affects the quality of

searching keywords in texts. As a result, statistical methods are usually accompanied by empirical procedures tailored to a specific class of problems, which significantly narrows their scope of its usage.

Increasing the quality of the searching keywords process requires additional information, preferably universal rather than specific knowledge. In particular, there is no formal description and solution of the problem of searching keywords as a convolution of information in the text. The known methods of searching keywords do not take into account the results of the analysis of relationships between lexical units of the text, and the informational evaluation of the results of parsing the text is not taken into account as part of the relevant quality criterion.

The hybrid methods of keyword search, for which the speed of statistical text processing is enhanced by the capabilities of modern linguistic packages that have the most advanced functionality, especially for the English language, are promising. Therefore, the actual scientific task is to increase the quality of searching keywords in English text by developing information technology for searching keywords based on parsing English texts.

The purpose of the qualification research paper is to improve the quality of searching keywords in English text.

In order to achieve the mentioned purpose, the following basic tasks are solved in the work:

1. Analysis and comparative characteristics of known approaches, methods and ways of searching keywords.

2. Formalization of a mathematical model of the searching keywords process based on information evaluation of the results of text parsing.

3. Development of a method for searching keywords based on determining the relationships between word forms.

4. Development of a method to reduce the impact of verbal noise for searching keywords.

5. Experimental research of the results of the proposed methods in comparison with analogues.

6. Development and approbation of information technology for searching keywords in English text.

The scientific novelty of the qualification research paper is:

1. The model of searching keywords has been improved, which, unlike the existing ones, is based on the information evaluation of parsing text results and takes into account the results of analysis of relationships between lexical units of text, which allowed to formalize the quality criterion of searching keywords process.

2. For the first time, searching keywords method has been developed, which, unlike the existing ones, is based on finding syntactic relationships between word forms in sentences of English text with the help of technological capabilities of parsing of modern linguistic packages. The proposed method allows to improve the numerical characteristics of searching keywords quality, namely completeness (according to Jacquard) and accuracy.

3. The method of reducing the impact of verbal noise for searching keywords has been improved, which, unlike the existing ones, is based on the Stanford classification of relationships between lexical units of a sentence, which has improved the quality of results of searching keywords compared to the main method.

4. The information technology of searching keywords has been further developed, which, unlike the existing ones, takes into account additional information of sentence parsing processes within the bounds of the consistent use of the two proposed methods, which allowed to refine numerical estimates of content parameters of the text and improve the quality of searching keywords.

The practical value of the results obtained in the qualification research paper is as follows: formalization of the searching keywords method of the English text, creation of an algorithm for its implementation and development of software that searching keywords based on complex relationships between word forms in sentences

of English text and subsequent filtering of verbal noise. Created models, algorithms and software can be used to solve practical problems of computational linguistics that require searching keywords, for example, to improve the accuracy of site content analysis and raise the position of the site in search results. The use of language-independent means of the proposed information technology in combination with the necessary, according to the obtained specification, technological resources of linguistic analysis of other natural languages will expand the area of information technology, in particular for Ukrainian texts.

The results of the qualification research paper were reported and discussed at 8 scientific and technical conferences:

- international internet conference «Youth in Science: Research, Problems, Prospects» (MTH-2015), Vinnytsia National Technical University, April 23-26, 2015;
- the first international conference «Adaptive technologies of learning management», Odessa, September 23-25, 2015;
- the third international scientific conference «Measurement, control and diagnostics in technical systems» (ВКДТС-2015), Vinnytsia National Technical University, October 27-29, 2015;
- XLIV, XLV, XLVI, XLVII and XLVIII scientific and technical conferences of teaching staff, staff and students of Vinnytsia National Technical University, Faculty of Computer Systems and Automation, annually in 2015 - 2019.

The results of the qualification research paper which are presented, was conducted in accordance with the priority thematic areas of scientific and technical development for the period up to 2020 "Technologies and tools for software and systems development", approved by the Cabinet of Ministers of Ukraine №556 from 23.08.2016. In particular, the research was conducted according to the research plans of the Automation and Intelligent Information Technologies Department of Vinnytsia National Technical University, namely in the research works «Intelligent information technology of figurative analysis of text and synthesis of integrated knowledge base

of natural language content» (№ 0114U003462) та «Identification of hidden dependencies in online social networks based on the methods of fuzzy logic and computational linguistics» (№ 0117U000575).

The results of the qualification research paper were implemented at LLC «Spilna Sprava» and also to the educational process of the Automation and Intelligent Information Technologies Department of Vinnytsia National Technical University. This is confirmed by the publication of a textbook “A lexical relationships-based keywords selection in an English text” and act on the results of implementation from January 10, 2020. The results of the experiments showed that the proposed information technology simultaneously increases in the range from 8.1% to 12.7% the completeness of the Jacquard metric and from 9.1% to 14.3% the absolute accuracy of searching keywords for English texts of 140-1400 words in comparison with analogues.

Keywords: information technology, keywords, verbal noise, linguistic package, DKPro Core, syntactic analysis, phrase.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

[1] O.V. Bisikalo, W. Wójcik, O.V. Yahimovich, and S. Smailova, "Method of determining of keywords in English texts based on the DKPro Core", *Proceedings of SPIE 10031, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016*, 100314T, 2016. DOI:10.1117/12.2249225.

[2] O. Bisikalo, A. Lisovenko, O. Jahumovuch, S. Trachenko, and M. Pradivliannyi, "System of computational linguistic on base of the figurative text comprehension", *Proceedings of the 2016 IEEE 1st International Conference on Data Stream Mining and Processing (DSMP 2016)*, pp. 69-74, 2016. DOI: 10.1109/DSMP.2016.7583510.

[3] О.В. Бісікало, та О.В. Яхимович, "Метод визначення ключових слів англомовного тексту на основі DKPro Core", *Технологический аудит и резервы производства: Информационные технологии*, Том 1, № 2(21), с. 26-30, 2015. ISSN 2226-3780.

[4] О.В. Бісікало, А.І. Лісовенко, О.В. Яхимович, та С.С. Траченко, "Визначення змістовних ознак тексту на основі аналізу зв'язків між лексичними одиницями", *Вісник НТУ «ХПІ». Серія "Механіко-технологічні системи та комплекси"*, № 21 (1130), с. 83-89, 2015. ISSN 2411-2798.

[5] О.В. Бісікало, та О.В. Яхимович, "Знаходження ключових слів англомовного тексту за допомогою інструментальних засобів пакету DKPro Core", *Інформаційні технології та комп'ютерна інженерія*, № 2(34), с. 10-14, 2015. ISSN 1999-9941.

[6] О.В. Бісікало, та О.В. Яхимович, "Автоматизоване визначення лексичних технологій з тезаурусу технічного спрямування", *Оптико-електронні інформаційно-енергетичні технології*, № 1 (31), с. 26-38, 2016. ISSN 1681-7893.

[7] О.В. Бісікало, О.В. Яхимович, та Я.В. Яхимович, "Розробка методу фільтрації вербального шуму в процесі пошуку ключових слів англомовного тексту", *Технологический аудит и резервы производства: Информационные технологии*, № 6(44), с. 33–41, 2018. ISSN 2226-3780.

[8] С.Д. Штовба, О.В. Штовба, О.В. Яхимович, та М.В. Петричко, "Вплив синтаксичних зв'язків у реченнях на якість ідентифікації токсичних коментарів в соціальній мережі", *Наукові праці ВНТУ*, № 4, с. 1-8, 2019. DOI: https://doi.org/10.31649/2307-5376-2019-4-35-42.

[9] О.В. Яхимович, "Визначення ключових слів англомовного тексту з використанням технології DKPRO CORE", *Молодь в технічних науках: дослідження, проблеми, перспективи (МТН-2015) : Матеріали міжнародної Інтернет-конференції*, Вінниця, 2015, с. 72-74. ISBN 978-966-924-027-9.

[10] О.В. Бісікало, О.В. Яхимович, А.І. Лісовенко, та Траченко С.С., "Підтримка діалогу з навчальним контентом", *Адаптивні технології управління навчанням: матеріали першої міжнародної конференції*, Одеса, 2015, с. 97-100.

[11] О.В. Бісікало, О.В. Яхимович, А.І. Лісовенко, та Траченко С.С., "Моделювання процесів побудови парадигматичних зв'язків між словоформами на основі вимірювання текстової інформації", *Вимірювання, контроль та діагностика в технічних системах (ВКДТС-2015)*, Вінниця, 2015, с. 119-121.

[12] О.В. Яхимович, "Застосування інструментальних засобів пакету DKPRO CORE для визначення ключових слів англомовного тексту", *XLIV науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2015, с. 7.

[13] О.В. Яхимович, "Визначення ключових слів з тексту повідомлень мікроблогів", *XLV науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2016, с. 1119-1121.

[14] О.В. Яхимович, "Колізія при знаходженні ключових слів", *XLVI науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2017, с. 1312-1314.

[15] О.В. Яхимович, "Зменшення вербального шуму при визначенні ключових слів", *XLVII науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2018, с. 1463-1465.

[16] О.В. Яхимович, "Формалізація задачі визначення ключових слів тексту", *XLVIII науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2019, с. 1136-1138.

[17] О.В. Бісікало, Р.Н. Квєтний, С.Г. Кривогубченко, Л.Є Азарова, та О.В. Яхимович, "Синтез інтегрованої бази знань природно-мовного контенту", ВНТУ, Вінниця, Д/б № 0114U003462, 2015.

[18] О.В.Бісікало, Р.Н. Квєтний, С.Г. Кривогубченко, А.І. Лісовенко, та О.В. Яхимович, "Розв’язання семантико-залежних задач обробки природно-мовних об’єктів на основі бази знань", ВНТУ, Вінниця, Д/б №0114U003462, 2016.

[19] О. В. Бісікало, А. І. Лісовенко, О. В. Яхимович, та В. В. Шолота, "Спосіб автоматичного пошуку ключових слів з використанням технології DKPro Core", *МПК G06F 17/21, G06F 17/27, G06F 17/28. № и 2019 00016*, Черв. 25, 2019.

[20] O. Bisikalo, and A. Yahimovich. *Keyword search based on lexical relationships in the text*. Beau Bassin-Rose Hill, Mauritius: Lap Lambert Academic Publishing, 2019. ISBN 978-620-0-00314-0.

[21] O. Bisikalo, and A. Yahimovich. *Lexical relationships-based keywords selection in english texts*. Vinnytsia, Ukraine: VNTU, 2020.

ЗМІСТ

ВСТУП

Обґрунтування вибору теми дослідження. Завдання пошуку ключових слів тексту виникає у бібліотечній справі, лексикографії та термінознавстві, а також в усіх задачах інформаційного пошуку. В даний час обсяги і динаміка інформації, яка підлягає обробці в цих областях, роблять особливо актуальною задачу автоматичного пошуку ключових слів, які можуть використовуватися для створення і розвитку термінологічних ресурсів, а також для ефективної обробки документів – індексування, реферування, кластеризації та класифікації. Популярність та затребуваність пошукових машин для кожного користувача Інтернету висуває високі вимоги до релевантності результатів пошуку, яка значно залежить від якості пошуку ключових слів у текстовій інформації.

Значний внесок у розвиток математичних моделей та методів семантичного аналізу природномовних текстів закладено закордонними дослідниками Н. Хомскі, Д. Маккарті, Т. Виноградом, К. Менінгом, Д. Журафски, Колмогоровим А.М., Мельчуком І.А., Апресяном Ю.Д., Поспєловим Д.А., Сосніним П.І., плідно працювали та продовжують дослідження у цьому напряму вітчизняні науковці Шабанов-Кушнаренко Ю.П., Бондаренко М.Ф., Широков В.А., Шлезінгер М.І., Анісімов А.В., Шаронова Н.В., Дарчук Н.П., Бісікало О.В., Хайрова Н.Ф, Шевченко І.В. тощо.

Натепер існує значна кількість доступних систем автоматичного пошуку ключових слів, розроблених і орієнтованих на обробку природних мов. В основу роботи таких систем покладено відомі методи пошуку ключових слів тексту, які умовно можна розділити на дві категорії – експертно-лінгвістичні та статистичні. Перша група методів ґрунтується на значеннях слів, отриманих експертним шляхом, зокрема використовують словники з зібраними семантичними даними про кожне слово або онтології предметних областей. Такі методи ресурсоємні на ранніх етапах – розробка онтологій, наприклад,

вельми трудомісткий процес [1]. Тому найбільша кількість відомих напрацювань у напряму експертно-лінгвістичної обробки текстів відома для визнаної мови міжнародного спілкування – англійської. З іншого боку, традиційні статистичні методи, які є фактично мово-незалежними, супроводжуються значними обсягами «вербального шуму», що суттєво впливає на якість пошуку ключових слів. Внаслідок цього статистичні методи зазвичай супроводжуються емпіричними процедурами, налаштованими на конкретний клас задач, що суттєво звужує їхню область застосування.

Підвищення якості процесу пошуку ключових слів вимагає залучення додаткової інформації, бажано універсального, а не специфічного характеру. Зокрема, відсутня формальна постановка та розв'язок задачі пошуку ключових слів як згортки інформації у тексті. Не враховуються у відомих методах пошуку ключових слів результати аналізу зв’язків між лексичними одиницями тексту, а інформаційна оцінка результатів парсингу тексту не береться до уваги як складова відповідного критерія якості.

Потрібно звернути увагу на гібридні методи пошуку ключових слів, для яких швидкість статистичної обробки тексту підсилюється можливостями сучасних лінгвістичних пакетів, що мають найбільш розвинутий функціонал, у першу чергу, для англійської мови. Тому *актуальним* науковим завданням є підвищення якості пошуку ключових слів у англомовному тексті шляхом розробки інформаційної технології пошуку ключових слів на основі парсингу англомовних текстів.

Зв’язок роботи з науковими програмами, планами, 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 наукових

досліджень 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).

Мета і завдання дослідження. Мета дослідження полягає у підвищенні якості пошуку ключових слів у англомовному тексті.

Для досягнення поставленої мети необхідно розв’язати наступні задачі:

1. Аналіз й порівняльна характеристика відомих підходів, методів та засобів пошуку ключових слів.
2. Побудова математичної моделі процесу пошуку ключових слів на основі інформаційної оцінки результатів парсингу тексту.
3. Розробка методу пошуку ключових слів на основі визначення зв'язків між словоформами.
4. Розробка методу зменшення впливу вербального шуму на пошук ключових слів.
5. Експериментальне дослідження результатів запропонованих методів у порівнянні з аналогами.
6. Розробка та апробація інформаційної технології пошуку ключових слів англомовного тексту.

Об’єкт дослідження – процес обробки вербальної інформації для пошуку ключових слів у англомовному тексті.

Предмет дослідження – моделі, методи та технологічні засоби пошуку ключових слів у англомовному тексті.

Методи досліджень. Для побудови моделі пошуку ключових слів використано методи 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. Під час оцінки релевантності запропонованої інформаційної технології застосовано методи вимірювання у метричних просторах.

Наукова новизна отриманих результатів.

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.

Практичне значення отриманих результатів. Прикладні результати дисертаційного дослідження полягають 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.

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, згідно з отриманою специфікацією, 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 АІІТ Вінницького національного технічного університету, що підтверджено виданням навчального посібника "A lexical relationships-based keywords selection in an English text" та актом про результати впровадження від 10.01.2020). 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.

Особистий внесок здобувача. Усі результати, які складають основний зміст дисертаційної роботи, отримані здобувачем самостійно. У роботах [146], [142], [125], [126], [130], [94] здобувачеві належать усі теоретичні та практичні результати. У роботах, опублікованих у співавторстві, здобувачу належать: [116] – розробка програмного забезпечення для методу пошуку ключових слів на

основі знаходження синтаксичних зв’язків між словоформами у реченнях англомовного тексту за допомогою технологічних можливостей парсингу лінгвістичного пакету DKPro Core; [147] – запропоновано використання заміни займенників на відповідні до них іменники для зменшення впливу вербального шуму на пошук ключових слів; експериментально підтверджено, що результати пошуку ключових слів розробленим методом містять менше стоп-слів у порівнянні з частотним словником; [101] – запропоновано підхід до інформаційної оцінки процесів парсингу тексту у межах задачі пошуку ключових слів; [10] – запропоновано методику зменшення впливу вербального шуму на пошук ключових слів за допомогою вилучення слів, які відносяться до неінформативних частин мови, а також слів, що відносяться до списку стоп-слів; [156] – експериментально підтверджено доцільність використання розробленого методу пошуку ключових слів, оскільки він забезпечує кращу релевантність результатів пошуку ключових слів у порівнянні з аналогами за критеріями точності і повноти; [45] – запропоновано методику побудови онтологій на основі знаходження відношень між термінами, яка може бути використана для поліпшення якості інформаційного пошуку; [152] – запропоновано метод зменшення впливу вербального шуму на пошук ключових слів на основі стенфордської класифікації зв'язків між лексичними одиницями речення; [144] – запропоновано використовувати інформаційної технології пошуку ключових слів для покращення ідентифікації токсичних коментарів в соціальних мережах; [134], [119] – розробка програмного забезпечення для модулю побудови семантичного графу на основі зв'язків між словами у словосполученнях, отриманих з тексту, а також огляд сучасних лінгвістичних пакетів; [135] – запропоновано використання модульного підходу до фільтрації вербального шуму, що складається з модулів: виключення словосполучень кандидатів з не інформативними типами зв’язків, заміни займенників на іменники в словосполученнях кандидатів, виключення

ключових слів, які належать до попередньо визначеного переліку не інформативних для семантичного аналізу частин мови та списку стоп-слів; [132], [133] – розробка програмного забезпечення на основі пакету DKPro Core для задачі виявлення інформативних ознак тексту; запропоновано метод класифікації тексті, що забезпечує підвищення кількісних характеристик релевантності результатів, а саме повноту і точність; [107], [112] – проведено аналіз зв'язків між лексичними одиницями тексту; реалізація програмного експерименту з використанням лінгвістичного пакету DKPro Core на англомовних текстах різної довжини; здійснено розрахунок чисельних характеристик якості отриманих результатів за критеріями повноти та точності та доведено ефективність використання розробленого методу в порівнянні з аналогами.

Апробація матеріалів дисертації. Основні результати та дисертаційна робота в цілому апробовані на восьми науково-практичних конференціях:

- міжнародній Інтернет-конференції «Молодь в технічних науках: дослідження, проблеми, перспективи» (МТН-2015), м. Вінниця, 23-26 квітня 2015 р.;

- першій міжнародній конференції «Адаптивні технології управління навчанням», м. Одеса, 23-25 вересня 2015 р.;

- третій міжнародній науковій конференції «Вимірювання, контроль та діагностика в технічних системах» (ВКДТС-2015), м. Вінниця, 27-29 жовтня 2015 р.;

- XLIV, XLV, XLVI, XLVII та XLVIII науково-технічних конференціях підрозділів ВНТУ факультету комп'ютерних систем та автоматики, м. Вінниця, щорічно у 2015 – 2019 рр.

Публікації. За темою дисертації з викладенням її основних результатів опубліковано 21 науковій праці, серед яких 1 стаття у закордонному фаховому періодичному виданні, що входить до SCOPUS, 1 стаття у виданні, що входить

до 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 (Додаток А).

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел і додатків. Основний зміст викладено на 137 сторінках друкованого тексту, містить 67 рисунки, 17 таблиць. Список використаних джерел містить 157 найменувань. Загальний обсяг 193 сторінки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Ю.С. Ершов, "Выделение ключевых слов в русскоязычных текстах", *Молодежный научно-технический вестник. - М.: ФГБОУ ВПО "МГТУ им. Н.Э. Баумана"* № ФС77-51038, с. 70-79, 2014.

[2] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang, "Automatic keyword extraction from documents using conditional random fields", *Journal of Computational Information Systems №4,* pp. 1169-1180, 2008.

[3] R. Feldman, and J. Sanger, "Task-Oriented approaches", *The text mining handbook: advanced approaches in analyzing unstructured data, Cambridge: Cambridge Univ. Pr.*, pp. 58-62, 2013.

[4] J. Mijic, B. Bašic, and J. Šnajder, "Robust keyphrase extraction for a large-scale Croatian news production system", *FASSBL7*, pp. 59-66, 2010.

[5] G. Palshikar, "Keyword extraction from a single document using centrality measures", *Lecture Notes in Computer Science Pattern Recognition and Machine Intelligence,* pp. 503-510, 2007.

[6] R. Mihalcea, and P. Tarau, "Textrank: Bringing order into text", *Proceedings of the 2004 conference on empirical methods in natural language processing,* pp. 404-411, 2004.

[7] S. Lahiri, S. Choudhury, and C. Caragea, "Keyword and keyphrase extraction using centrality measures on collocation networks", arXiv preprint arXiv:1401.6571, 2014.

[8] F. Boudin "Comparison of Centrality Measures for Graph-Based Keyphrase Extraction", *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP),* pp. 834-838, 2013.

[9] К. Н. Даркулова, и Г. Ергешова, "Необходимость выделения ключевых слов для свёртывания текста", [9] *Лингвистический анализ научного текста. VI Международная студенческая электронная научная конференция,*

Южно-Казахстанский государственный университет им. Мухтара Ауэзова Шымкент, с. 30-35, 2014.

[10] O. Bisikalo, A. Lisovenko, O. Jahumovuch, S. Trachenko, and M. Pradivliannyi, "System of computational linguistic on base of the figurative text comprehension", *Proceedings of the 2016 IEEE 1st International Conference on Data Stream Mining and Processing (DSMP 2016)*, pp. 69-74, 2016. DOI: 10.1109/DSMP.2016.7583510.

[11] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge", *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 216-223, 2003. doi:10.3115/1119355.1119383

[12] Y. HaCohen-Kerner, "Automatic extraction of keywords from abstracts", *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 843-849, 2003.

[13] N. Pudota, A. Dattolo, A. Baruzzo, and C. Tasso, "A New Domain Independent Keyphrase Extraction System", *Communications in Computer and Information Science Digital Libraries*, pp. 67-78, 2010. doi: 10.1007/978-3-642-15850-6_8

[14] В. Шорошев и Е. Маевский, "Методические основы эффективного поиска информации в сети интернет", *Правове, нормативне та метрологічне забезпечення системи захисту інформації в Україні*, № 4, с. 81-88, 2002.

[15] А.М. Андреев, Д.В.Березкин, В.В. Сюзев, и В.И. Шабанов, "Модели и методы автоматической классификации текстовых документов", *Вестн. МГТУ. Сер. Приборостроение,* МГТУ, №3, с. 64-94, 2003.

[16] T. Joachims, "Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization", *Carnegie-mellon Univ Pittsburgh PA Dept of computer science*, No. CMU-CS-96-118, pp. 143-151, 1996.

[17] N. Fuhr, N. Govert, M. Lalmas, and F. Sebastiani, "Categorisation tool: Final prototype. Deliverable 4.3, Project LE4-8303 «EUROSEARCH»", *Commission of the European Communities*, 30 p. 1999.

[18] L. Larkey, and W. Croft, "Combining classifiers in text categorization. In Proceedings of SIGIR 96", *19th ACM International Conference on Research and Developmentin Information Retrieval*, pp. 289-297, 1996.

[19] T. Joachims, "A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization", *In Proc. of the ICML'97*, pp. 289-297, 1997.

[20] M. Litvak, M. Last, H. Aizenman, I. Gobits, and A. Kandel, "DegExt - A language-independent graph-based keyphrase extractor", *Advances in intelligent web mastering - 3*, pp. 121-130, 2011. doi:10.1007/978-3-642-18029-3_13

[21] W. Abilhoa, and L. Castro, "A keyword extraction method from twitter messages represented as graphs", *Applied Mathematics and Computation, №240*, pp. 308-325, 2014. doi:10.1016/j.amc.2014.04.090

[22] Z. Zhou, X. Zou, X. Lv, and J. Hu, "Research on Weighted Complex Network Based Keywords Extraction", *Lecture Notes in Computer Science Chinese Lexical Semantics, №8229*, pp. 442-452, 2013. doi:10.1007/978-3-642-45185-0_47

[23] X. Wan, and J. Xiao, "Single Document Keyphrase Extraction Using Neighborhood Knowledge", *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pp. 855-860, 2008.

[24] Е.Г. Абрамов, "Подбор ключевых слов для научной статьи", *Научная периодика: проблемы и решения, № 1(2)б*, pp. 35-40, 2011.

[25] Как составить список ключевых слов?, [Электронный ресурс]. Режим доступа: http://blog.creativeconomy.ru/2009/04/02/kak-sostavit-spisok-klyuchevyx-slov/.

[26] J. Borge-Holthoefer, and A. Arenas, "Semantic Networks: Structure and Dynamics", *Entropy, Vol. 12, № 5*, pp. 1264-1302, 2010. doi: 10.3390/e12051264

[27] A. Masucci, and G. Rodgers, "Differences Between Normal And Shuffled Texts: Structural Properties Of Weighted Networks", *Advances in Complex Systems, Vol. 12, № 01*, pp. 113-129, 2009. doi: 10.1142/s0219525909002039

[28] И.Е. Воронина, А.А. Кретов, и О.С. Титова, "Программные средства выявления семантического поля слов", *Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии, № 2*, с. 111-122, 2008.

[29] В.Т. Титов, *Частная квантитативная лексикология романских языков: Монография,* Воронеж: Изд-во Воронеж. гос. ун-та, 552 с. 2004.

[30] А.А. Кретов, "Метод формального выделения тематически нейтральной лексики (на примере старославянских текстов)", *Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии, № 1*, с. 81-90, 2007.

[31] Функциональный подход к выделению ключевых слов: методика и реализация, [Электронный ресурс]. Режим доступа: http://www.vestnik.vsu.ru/pdf/analiz/2009/01/2009-01-10.pdf

[32] И.Е. Воронина, *Компьютерное моделирование лингвистических объектов: монография,* Воронеж: Издательско- полиграфический центр ВГУ, 177 с. 2007.

[33] Е.В. Крештель, А.А. Кретов, и И.Е. Воронина, "Алгоритмы выделения ключевых слов в текстах на естественных языках", *Информатика: проблемы, методология, технологии : матер. 3-й регион. науч.-метод. конфер.* Воронеж. Изд-во ВГУ, с. 35, 2001.

[34] Программные средства выявления семантического поля слов, [Электронный ресурс]. Режим доступа: http://www.vestnik.vsu.ru/pdf/analiz/2008/02/2008_02_19.pdf

[35] D. Bourigault, "Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases", *Proceedings of the 14th conference on Computational linguistics-Volume 3. Association for Computational Linguistics.* Nantes, France, pp. 977-981, 1992.

[36] П.И. Браславский, и Е.А. Соколов, "Автоматическое извлечение терминологии с использованием поисковых машин Интернета", *Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог*. М.: РГГУ, с. 89-94, 2007.

[37] M. Baroni, "BootCaT: Bootstrapping Corpora and Terms from the Web", *Proceedings of LREC. Lisbon: ELDA*, pp. 1313-1316, 2004.

[38] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms:. the C-value/NC-value method", *International Journal on Digital Libraries, Vol. 3*, pp. 115-130, 2000.

[39] Б.В. Добров, Н.В. Лукашевич, и С.В. Сыромятников, "Формирование базы терминологических словосочетаний по текстам предметной области", *Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды пятой Всероссийской научной конференции,* с. 201-210, 2003.

[40] П.И. Браславский, и Е.А. Соколов, "Сравнение четырех методов автоматического извлечения двухсловных терминов из текста", *Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. Диалог'2006*. - М.: РГГУ, с. 88-94, 2006.

[41] С.Д. Шелов, "Терминоведение: семь вопросов и семь ответов по семантике термина", *Информационные процессы и системы, №2,* с. 1-12, 2001.

[42] Синтаксический анализ. Проект АОТ, [Электронный ресурс]. Режим доступа: http://www.aot.ru/docs/synan.html

[43] П.И. Браславский, и Е.А. Соколов, "Сравнение пяти методов извлечения терминов произвольной длины", *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог»,* с. 67-74, 2008.

[44] О.В. Пескова, "Автоматическое формирование рубрикатора полнотекстовых документо", *Электронные библиотеки: перспективные методы и технологии, электронные коллекции,* с. 139-148, 2008.

[45] О.В. Бісікало, та О.В. Яхимович, "Автоматизоване визначення лексичних технологій з тезаурусу технічного спрямування", *Оптико-електронні інформаційно-енергетичні технології*, № 1 (31), с. 26-38, 2016. ISSN 1681-7893.

[46] O. Medelyan, and I. Witten, "Thesaurus based automatic keyphrase indexing", *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries,* pp. 296-297, 2006. doi:10.1145/1141753.1141819

[47] J. Bezdek, and N. Pal, "Some New Indexes of Cluster Validity", *IEEE Transactions On Systems, Man And Cybernetics, Vol. 28, no. 3,* pp. 301-315, 1998.

[48] M. Halkidi, V. Batistakis, and M. Vazirgiannis, "On Clustering Validation", *Journal of Intelligent Information Systems, Vol. 17,* pp. 107-145, 2001.

[49] В.Б. Барахнин, и Д.А. Ткачев, "Кластеризация текстовых документов на основе составных ключевых термов", *Вестник НГУ. Серия: Информационные технологии, № 8(2),* с. 5-14, 2010.

[50] А.М. Федотов, и В.Б. Барахнин, "К вопросу о поиске документов «по аналогии»", *Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии, Т. 7, № 4*, с. 3-14, 2009.

[51] В.Б. Барахнин, В.А. Нехаева, и А.М. Федотов, "О задании меры сходства для кластеризации текстовых документов", *Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии, Т. 6, № 1,* с. 3-9, 2008.

[52] E. Martin, "Microblogging - more than fun?", *Proceedings of IADIS Mobile Learning Conference. - Inmaculada Arnedillo Sánchez and Pedro Isaías ed., Portugal,* pp. 155-159, 2008.

[53] D. Herman, M. Janh, and M. Ryan, "The Routledge Encyclopedia of Narrative Theory", *London, Routledge,* 997 p, 2005.

[54] M. Böhringer, "Really Social Syndication: A Conceptual View on Microblogging", *Sprouts: Working Papers on Information Systems, № 9(31),* pp. 147-157, 2009.

[55] D. Karger, and D. Quan, "What would it mean to blog on the semantic web?", *Web Semantics: Science, Services and Agents on the World Wide Web. - Selected Papers from the International Semantic Web Conference, Hiroshima, Japan, №3,* pp. 143-147, 2004.

[56] P. Turney, "Learning to extract keyphrases from text", *Technical report, National Research Council, Institute for Informational Technology,* pp. 143-147, 1999.

[57] P. Chen, and S. Lin, "Automatic keyword prediction using Google similarity distance", *Expert Systems with Applications, №37,* pp. 1928-1938, 2010. doi:10.1016/j.eswa.2009.07.016

[58] F. Sebastiani, "Machine learning in automated text categorization", *ACM computing surveys (CSUR), №34,* pp. 1-47, 2002.

[59] Y. Hacohen, Z. Gross, and A. Masa, "Automatic Extraction and Learning of Keyphrases from Scientific Articles", *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science,* pp. 657-669, 2005. doi:10.1007/978-3-540-30586-6_74

[60] M. Krapivin, A. Autayeu, M. Marchese, E. Blanzieri, and N. Segata, "Keyphrases Extraction from Scientific Documents: Improving Machine Learning Approaches with Natural Language Processing", *International Conference on Asian Digital Libraries Lecture Notes in Computer Science, №6102,* pp. 102-111, 2010. doi:10.1007/978-3-642-13654-2_12

[61] M. Bekavac, and J. Šnajder, "Gpkex: Genetically programmed keyphrase extraction from croatian texts", *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing,* pp. 43-47, 2013.

[62] R. Ahel, B. Dalbelo Bašic, and J. Šnajder, "Automatic keyphrase extraction from Croatian newspaper articles", *The Future of Information Sciences, Digital Resources and Knowledge Sharing,* pp. 207-218, 2009.

[63] D. Turdakov, "Word sense disambiguation methods", *Programming and Computer Software, Vol. 36, № 6,* pp. 309-326, 2010.

[64] D. Turdakov, and S. Kuznetsov, "Automatic word sense disambiguation based on document networks", *Programming and Computer Software, Vol. 36, № 1,* pp. 11-18, 2010.

[65] D. Lizorkin, P. Velikhov, M. Grinev, and D. Turdakov, "Accuracy estimate and optimization techniques for SimRank computation", *The International Journal on Very Large Data Bases archive, Vol. 19, № 1,* pp. 15-19, 2010.

[66] D. Lizorkin, P. Velikhov, M. Grinev, and D. Turdakov, "Accuracy Estimate and Optimization Techniques for SimRank Computation", *Proceedings of the VLDB Endowment, Vol. 1, № 1,* pp. 12-17, 2008.

[67] M. Grinev, D. Lizorkin, and D. Turdakov, "Effective Extraction of Thematically Grouped Key Terms From Text", *Proc. of the AAAI 2009 Spring Symposium on Social Semantic Web,* pp. 39-44, 2009.

[68] D. Turdakov, and D. Lizorkin, "HMM Expanded to Multiple Interleaved Chains as a Model for Word Sense Disambiguation", *The 23rd Pacific Asia Conference on Language, Information and Computations*, pp. 549-559, 2009.

[69] M. Grineva, M. Grinev, and D. Lizorkin, "Extracting Key Terms From Noisy and Multitheme Documents", *WWW2009: 18th International World Wide Web Conference*, pp. 521-533, 2009.

[70] A. Guo, and Y. Tao, "Research and Improvement of Feature Words Weight Based on TFIDF Algorithm", *IEEE Information Technology, Networking, Electronic and Automation Control Conference (May 2016),* 2016.
doi:10.1109/itnec.2016.7560393

[71] M. Grineva, M. Grinev, A. Boldakov, L. Novak, A. Syssoev, and D. Lizorkin, "Sifting Micro-blogging Stream for Events of User Interest", *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 327-333, 2009.

[72] J. Reed, Y. Jiao, T. Potok, B. Klump, M. Elmore, and A. Hurson, "TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams", *Proc. Machine Learning and Applications*, pp. 258-263, 2006.

[73] R. Mihalcea, and A. Csomai, "Wikify!: linking documents to encyclopedic knowledge", *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management,* pp. 233-242, 2007.

[74] G. Salton, "The SMART Retrieval System - experiments in automatic", *Prentice-Hall, Inc., Englewood Cliffs*, pp. 111-121, 1971.

[75] Z. Dejin, and M. Rosson, "How and why people Twitter: the role that micro blogging plays in informal communication at work", *Proceedings of the ACM 2009 international conference on Supporting group work*, pp. 31-40, 2009.

[76] P. McFedries, "All A-Twitter", IEEE Spectrum, № 84, pp. 14-18, 2007.

[77] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities", *Proc. WebKDD/SNA-KDD '07. ACM Press*, pp. 47-50, 2007.

[78] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about twitter", *Proc. WOSP '08. - ACM Press*, pp. 33-36, 2008.

[79] C. Honeycutt, and S. Herring, "Beyond microblogging: Conversation and collaboration via Twitter", *Proc. HICSS '09. - IEEE Press*, pp. 6-9, 2009.

[80] M. Naaman, J. Boase, and C. Lai, "Is it really about me? Message content in social awareness streams", *Proc. CSCW 2010*, pp. 58-63, 2010.

[81] B. Huberman, D. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope", *First Monday, Vol. 14, №1,* pp. 14-21, 2008.

[82] А.В. Коршунов, "Извлечение ключевых терминов из сообщений микроблогов с помощью Википедии", *Труды Института системного программирования РАН, №20,* с. 102-115, 2011.

[83] A. Özgür, J. Hur, and Y. He, "The Interaction Network Ontology-Supported Modeling and Mining of Complex Interactions Represented with Multiple Keywords in Biomedical Literature.", *BioData Mining 9, no. 1 (December 2016)*, doi:10.1186/s13040-016-0118-0

[84] W. Wong, W. Liu, and M. Bennamoun, "Ontology Learning from Text", *ACM Computing Surveys 44, no. 4 (August 1, 2012),* pp. 1-36, 2012. doi:10.1145/2333112.2333115

[85] D. Korobkin, S. Fomenkov, and S. Kolesnikov, "Method of ontology-based extraction of physical effect description", *Vestnik Komp'iuternykh i Informatsionnykh Tekhnologii (2015),* pp. 28-35, 2015. doi:10.14489/vkit.2015.02.pp.028-035

[86] Л.Л. Волкова, "Приложения теории тесного мира в компьютерной лингвистике", *3-я Международная научно-практическая конференция «Модель подготовки специалистов новой формации, адаптированных к инновационному развитию отраслей»: сборник трудов. - Душанбе,* с. 172-155, 2012.

[87] Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, и Е.В. Ягунова, *Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие*, МИЭМ, 272 с. 2011.

[88] Бесплатный онлайн-генератор ключевых слов с текста [Электронный ресурс]. Режим доступа: http://seotool.by/analiz/seo/keywordstext.php

[89] Генератор ключевых слов с текста., [Электронный ресурс]. Режим доступа: http://www.rise-top.com/keywordstext.php

[90] R. Mihalcea, and A. Csomai, "Wikify!: linking documents to encyclopedic knowledge", *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. Lisbon,* pp. 233-242, 2007. doi: http://doi.org/10.1145/1321440.1321475

[91] J. Wu, and A. Agogino, "Automating keyphrase extraction with multi-objective genetic algorithms", *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the IEEE,* 2004.

[92] Y. Zhang, E. Milios, and N. Zincir-Heywood, "A comparison of keyword-and keyterm-based methods for automatic web site summarization", *AAAI04 Workshop on Adaptive Text Extraction and Mining,* pp. 15-20, 2004.

[93] С.В. Кулешов, А.А. Зайцева, и В.С. Марков, "Ассоциативно-онтологический подход к обработке текстов на естественном языке", *Интеллектуальные технологии на транспорте, № 4*, с. 40-45, 2015.

[94] О.В. Яхимович, "Формалізація задачі визначення ключових слів тексту", *XLVIII науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2019, с. 1136-1138.

[95] A set of Unicode character values, [Online]. Available: http://www.unicode.org/reports/tr44/#General_Category_Values

[96] Universal Dependencies (UD): Introduction, [Online]. Available: http://universaldependencies.org/introduction.html

[97] S. Sonawane, and P. Kulkarni, "Graph based Representation and Analysis of Text Document: A Survey of Techniques", *International Journal of Computer Applications, Vol. 96, № 19*, pp. 1-8, 2014. doi: 10.5120/16899-6972

[98] Universal Dependency Relations, [Online]. Available: http://universaldependencies.org/u/dep/

[99] C. Manning, and M. de Marneffe, Stanford typed dependencies manual, [Online]. Available: https://nlp.stanford.edu/software/dependencies_manual.pdf

[100] Stanford dependency hierarchy, [Online]. Available: https://nlp-ml.io/jg/software/pac/standep.html

[101] О.В. Бісікало, та О.В. Яхимович, "Знаходження ключових слів англомовного тексту за допомогою інструментальних засобів пакету DKPro Core", *Інформаційні технології та комп'ютерна інженерія*, № 2(34), с. 10-14, 2015. ISSN 1999-9941.

[102] О.В. Бісікало, "Концептуальна модель системи образного аналізу і синтезу природно-мовних конструкцій", *Математичні машини і системи, № 2,* с. 184-187, 2013.

[103] О.В. Бісікало, *Формальні методи образного аналізу та синтезу природно-мовних конструкцій : монографія,* Вінниця : ВНТУ, 316 с. 2013.

[104] С. Турчиняк, "Критерії відбору лексичного матеріалу та процес засвоєння лексичних одиниць", *V Міжнародна науково-практична інтернет-конференція «Проблеми та перспективи розвитку науки на початку третього тисячоліття у країнах СНД».* - Переяслав-Хмельницький, 17 - 19 листопада 2012, с. 211-213 2012.

[105] Слово як лексична одиниця, [Електронний ресурс]. Режим доступу: http://school-world.com.ua/lib/slovo-yak-leksichna-odinicya/

[106] Determiner, [Online]. Available: http://universaldependencies.org/u/dep/det.html

[107] O. Bisikalo, and A. Yahimovich. *Keyword search based on lexical relationships in the text*. Beau Bassin-Rose Hill, Mauritius: Lap Lambert Academic Publishing, 2019. ISBN 978-620-0-00314-0.

[108] Welo, E. "Null Anaphora", Encyclopedia of Ancient Greek Language and Linguistics (2013), [Online]. Available: https://referenceworks.brillonline.com/entries/encyclopedia-of-ancient-greek-language-and-linguistics/*COM_00000254

[109] Fixed multiword expression, [Online]. Available: http://universaldependencies.org/u/dep/fixed.html

[110] Punctuation, [Online]. Available: http://universaldependencies.org/u/dep/punct.html

[111] Root, [Online]. Available: http://universaldependencies.org/u/dep/root.html

[112] O. Bisikalo, and A. Yahimovich. *Lexical relationships-based keywords selection in english texts*. Vinnytsya, Ukraine: VNTU, 2020.

[113] A. Taylor, M. Marcus, and B. Santorini, The Penn Treebank: An Overview", [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.8216&rep=rep1&type=pdf

[114] Penn Treebank II Constituent Tags: Word level, [Online]. Available: http://www.surdeanu.info/mihai/teaching/ista555-fall13/readings/PennTreebankConstituents.html#Word

[115] Alphabetical list of part-of-speech tags used in the Penn Treebank Project, [Online]. Available: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

[116] О.В. Бісікало, та О.В. Яхимович, "Метод визначення ключових слів англомовного тексту на основі DKPro Core", *Технологический аудит и резервы производства: Информационные технологии*, Том 1, № 2(21), с. 26-30, 2015. ISSN 2226-3780.

[117] Интегрированные пакеты, [Электронный ресурс]. Режим доступа: https://nlpub.ru/%D0%9E%D0%B1%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B0_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%B0#.D0.98.D0.BD.D1.82.D0.B5.D0.B3.D1.80.D0.B8.D1.80.D0.BE.D0.B2.D0.B0.D0.BD.D0.BD.D1.8B.D0.B5_.D0.BF.D0.B0.D0.BA.D0.B5.D1.82.D1.8B

[118] Natural Language Toolkit, [Online]. Available: http://www.nltk.org/

[119] О.В. Бісікало, О.В. Яхимович, А.І. Лісовенко, та Траченко С.С., "Моделювання процесів побудови парадигматичних зв'язків між словоформами на основі вимірювання текстової інформації", *Вимірювання, контроль та діагностика в технічних системах (ВКДТС-2015)*, Вінниця, 2015, с. 119-121.

[120] A collection of software components for natural language processing (NLP) based on the Apache UIMA framework, [Online]. Available: https://dkpro.github.io/dkpro-core/

[121] Stanford NLP, [Online]. Available: https://nlp.stanford.edu/software/index.shtml

[122] SharpNLP - open source natural language processing tools, [Online]. Available: https://archive.codeplex.com/?p=sharpnlp

[123] MeTA: ModErn Text Analysis, [Online]. Available: https://meta-toolkit.org/

[124] Apache OpenNLP, [Online]. Available: https://nlpub.ru/Apache_OpenNLP

[125] О.В. Яхимович, "Визначення ключових слів з тексту повідомлень мікроблогів", *XLV науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2016, с. 1119-1121.

[126] О.В. Яхимович, "Колізія при знаходженні ключових слів", *XLVI науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2017, с. 1312-1314.

[127] L. Burgareli, "Variability management in software product lines using adaptive object and reflection. Journal of Aerospace Technology and Management, V. 1, № 2.", *Journal of Aerospace Technology and Management, V. 1, № 2. (2009, Jul.-Dec.)* [Online]. Available: http://www.jatm.com.br/papers/vol1_n2/JATMv1n2_thesis_abstracts.pdf

[128] N. Astrakhantsev, "Automatic Term Acquisition from Domain-Specific Text Collection by Using Wikipedia.", *Proceedings of the Institute for System Programming of RAS 26, no. 4*, pp. 7-20, 2014. doi:10.15514/ispras-2014-26(4)-1

[129] Л.А. Гращенко, "О модельном стоп-словаре", *Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук, № 1 (150)*, с. 40-46, 2013.

[130] О.В. Яхимович, "Зменшення вербального шуму при визначенні ключових слів", *XLVII науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2018, с. 1463-1465.

[131] Natural Language Processing: Integration of Automatic and Manual Analysis, [Online]. Available: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf

[132] О.В. Бісікало, Р.Н. Квєтний, С.Г. Кривогубченко, Л.Є Азарова, та О.В. Яхимович, "Синтез інтегрованої бази знань природно-мовного контенту", ВНТУ, Вінниця, Д/б № 0114U003462, 2015.

[133] О.В.Бісікало, Р.Н. Квєтний, С.Г. Кривогубченко, А.І. Лісовенко, та О.В. Яхимович, "Розв’язання семантико-залежних задач обробки природно-мовних об’єктів на основі бази знань", ВНТУ, Вінниця, Д/б № 0114U003462, 2016.

[134] О.В. Бісікало, О.В. Яхимович, А.І. Лісовенко, та Траченко С.С., "Підтримка діалогу з навчальним контентом", *Адаптивні технології управління навчанням: матеріали першої міжнародної конференції*, Одеса, 2015, с. 97-100.

[135] О. В. Бісікало, А. І. Лісовенко, О. В. Яхимович, та В. В. Шолота, "Спосіб автоматичного пошуку ключових слів з використанням технології DKPro Core", *МПК G06F 17/21, G06F 17/27, G06F 17/28. № и 2019 00016*, Черв. 25, 2019.

[136] Р. Розенталь, "Почему Java, а не что-то другое?", [Электронный ресурс]. Режим доступа: http://www.fainaidea.com/jeto-interesno-znat/pochemu-java-a-ne-chto-to-drugoe-130156.html

[137] Почему Java так популярна?, [Электронный ресурс]. Режим доступа: https://vertex-academy.com/tutorials/ru/pochemu-yazyk-java-tak-populyaren/

[138] 12 причин длительного доминирования Java, [Электронный ресурс]. Режим доступа: https://habr.com/post/201612/

[139] Cloud Natural Language, [Online]. Available: https://cloud.google.com/natural-language/

[140] Natural Language Framework, [Online]. Available: https://developer.apple.com/documentation/naturallanguage

[141] T. Casino, "Apple's Natural Language Processing (NLP) API", [Online]. Available: https://willowtreeapps.com/ideas/apples-natural-language-processing-nlp-api

[142] О.В. Яхимович, "Застосування інструментальних засобів пакету DKPRO CORE для визначення ключових слів англомовного тексту", *XLIV науково-технічній конференції підрозділів ВНТУ: факультет комп'ютерних систем та автоматики*, Вінниця, 2015, с. 7.

[143] Л.О. Терещенко, та І.І. Матієнко-Зубенко, "Інформаційні системи і технології обліку:", Навч. Посібник. К.: КНЕУ, 2003.

[144] С.Д. Штовба, О.В. Штовба, О.В. Яхимович та М.В. Петричко, "Вплив синтаксичних зв'язків у реченнях на якість ідентифікації токсичних коментарів в соціальній мережі", *Наукові праці ВНТУ*, № 4, с. 1-8, 2019. DOI: https://doi.org/10.31649/2307-5376-2019-4-35-42.

[145] Darmstadt Knowledge Processing Repository Based on UIMA, [Online]. Available: https://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2007/gldv-uima-ukp.pdf

[146] О.В. Яхимович, "Визначення ключових слів англомовного тексту з використанням технології DKPRO CORE", *Молодь в технічних науках: дослідження, проблеми, перспективи (МТН-2015) : Матеріали міжнародної Інтернет-конференції*, Вінниця, 2015, с. 72-74. ISBN 978-966-924-027-9.

[147] О.В. Бісікало, А.І. Лісовенко, О.В. Яхимович, та С.С. Траченко, "Визначення змістовних ознак тексту на основі аналізу зв'язків між лексичними одиницями", *Вісник НТУ «ХПІ». Серія "Механіко-технологічні системи та комплекси"*, № 21 (1130), с. 83-89, 2015. ISSN 2411-2798.

[148] Address by President of the Russian Federation, [Online]. Available: http://eng.kremlin.ru/transcripts/6402

[149] Address by President of the Russian Federation, [Online]. Available: http://eng.kremlin.ru/news/6889

[150] English grammatical relations, [Online]. Available: http://universaldependencies.org/en/dep/

[151] K. Bougé, "Lists of stop words", [Online]. Available: https://sites.google.com/site/kevinbouge/stopwords-lists

[152] О.В. Бісікало, О.В. Яхимович, та Я.В. Яхимович, "Розробка методу фільтрації вербального шуму в процесі пошуку ключових слів англомовного

тексту", *Технологический аудит и резервы производства: Информационные технологии*, № 6(44), с. 33–41, 2018. ISSN 2226-3780.

[153] L. Cameron, "Participation, archival activism and learning to learn", [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305748814001030

[154] D. Heitman, "Workingman's Poet", *Humanities №34*, 2013 p. 28, 2013.

[155] A new pattern for historical geography: working with enthusiast communities and public history, [Online]. Available: http://ac.els-cdn.com/S0305748814001029/1-s2.0-S0305748814001029-main.pdf?_tid=d45ec9e6-ba7b-11e4-b562-00000aab0f01&acdnat=1424600353_4-8bb4ef54fffbc3b800698d175c3c052

[156] O.V. Bisikalo, W. Wójcik, O.V. Yahimovich, and S. Smailova, "Method of determining of keywords in English texts based on the DKPro Core", *Proceedings of SPIE 10031, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016*, 100314T, 2016. DOI:10.1117/12.2249225.

[157] The Movies Dataset, [Online]. Available: https://www.kaggle.com/rounakbanik/the-movies-dataset