УДК 004.912

ПРАКТИЧНЕ ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ СЕМАНТИЧНИХ ТЕРМІНІВ У КОНТЕНТІ НАВЧАЛЬНИХ МАТЕРІАЛІВ

Ю.В. Крак, О.В. Бармак, О.В. Мазурець

Запропоновано інформаційну технологію, що ґрунтується на дисперсійній оцінці важливості слів, яка дозволяє з достатньо високою ефективністю визначати семантичні терміни в контенті навчальних матеріалів. Розглянуто фактори, що ускладнюють ефективне визначення семантичних термінів у навчальних матеріалах. Встановлена ефективність запропонованої технології сприяє її використанню для вирішення ряду актуальних задач, таких як оцінка відповідності навчальних матеріалів змістовим вимогам, оцінка відповідності наборів тестових завдань навчальним матеріалам, семантична допомога при створенні тестів, автоматизована генерація множин ключових слів та створення анотацій.
Ключові слова: інформаційна технологія, семантичні терміни, навчальні матеріали, анотація.

Предложена информационная технология, основанная на дисперсионной оценке важности слов, которая позволяет с достаточно высокой эффективностью определять семантические термины в контенте учебных материалов. Рассмотрены факторы, затрудняющие эффективное определение семантических терминов в учебных материалах. Установленная эффективность предложенной технологии способствует ее использованию для решения ряда актуальных задач, таких как оценка соответствия учебных материалов содержательным требованиям, оценка соответствия наборов тестовых заданий учебным материалам, семантическая помощь при создании тестов, автоматизированная генерация множеств ключевых слов и создания аннотаций.
Ключевые слова: информационная технология, семантические сроки, учебные материалы, аннотация.

The information technology on base of the disperse evaluation, which with enough high efficiency allows automated define the semantic terms in content of educational materials article is given. The factors that hinder effective analysis of educational materials have been considered. High efficiency offered technologies gives possible of its using in row of the problems, such as estimation of the correspondence of educational materials to requirements, estimation of the correspondence of set test tasks to educational materials, semantic help of making tests, automated keyword list and abstract generation.
Key words: information technology, semantic terms, training materials, summary.

Вступ та постановка задачі

Засобом реалізації дистанційної освіти є інформаційні технології [1], що визначає необхідність суттєвої формалізації та стандартизації навчального процесу [2]. Загальноприйнятим є підхід [3, 4] застосування навчальних матеріалів у вигляді цифрових документів визначеної структури як інструменту навчання, й тестів як інструмента контролю рівня отриманих знань [5].

Для розробки й використання курсів навчальних дисциплін за наведеним підходом використовуються спеціалізовані віртуальні навчаючі середовища, найбільш відомим із яких є Moodle [6]. При їх використанні, потенційна якість отриманих освітніх послуг безпосередньо визначається відповідністю навчальних матеріалів курсу вимогам стандартів освіти (робочим планам, структурі навчального плану тощо), й тестів – навчальним матеріалам [7].

Структурна відповідність навчальних матеріалів вимогам стандартів може бути оцінена шляхом аналізу структури відповідних цифрових документів. Задача ж оцінки семантичної відповідності в рамках визначених структурних одиниць контенту залишається актуальною [8].

Зі змістовної точки зору, ключовою властивістю контенту є його семантика, яку формалізовано відображають у вигляді семантичної мережі, вузлами якої є терміни, що несуть смислове навантаження, а дуги відображають характер зв'язку між вузлами [9]. Зв'язок між термінами навчальних матеріалів залежить від багатьох факторів (галузь знань, тип лекції, літературні здібності автора, тощо) й може змінюватися у широких межах без втрати якості викладання, що знижує актуальність його аналізу. Тому переважно аналіз саме термінів, що використовуються у навчальних матеріалах, дозволяє визначити якість цих навчальних матеріалів та їх відповідність вимогам.

Оскільки тести є засобом перевірки якості засвоєння сенсу навчальних матеріалів й ставлять на меті задачу перевірки якості засвоєння термінів як складових семантичних одиниць навчальних матеріалів, то визначення семантичних термінів у навчальних матеріалах може забезпечити допомогу та контроль при розробці наборів тестових завдань. Отже, автоматизація визначення семантичних термінів у навчальних матеріалах є перспективною задачею інформаційних технологій у сучасній освіті.

Для автоматизації пошуку ключових слів використовуються різноманітні методи аналізу текстів, таких як частотна оцінка [10], оцінка TFIDF [11] та дисперсійна оцінка [12]. Ці методи дозволяють співставити окремим словам або словосполученням тексту деякі певним чином поставлені у відповідність числові вагові значення, що вказують на міру їх важливості в досліджуваному тексті. Попередніми дослідженнями було визначено найбільш ефективним методом аналізу текстів метод дисперсійної оцінки, проте встановлено й ряд факторів, які унеможливлюють його монопольне застосування для вирішення розглядуваної задачі [13]. Це, зокрема, необхідність попередньої перевірки тексту на відповідність нормам ведення наукової літератури;

потреба в виявленні як ключових слів, так і з ключових словосполучень; удосконалення алгоритмів пошуку ключових слів і словосполучень з використанням методу дисперсійної оцінки, та інші. Тому є доцільною розробка нової інформаційної технології, яка із використанням методу дисперсійної оцінки дозволить ефективно й автоматизовано визначати семантичні терміни в навчальних матеріалах.

Мета даної роботи – розробка інформаційної технології автоматизованого визначення семантичних термінів у контенті навчальних матеріалів й дослідження її ефективності за допомогою відповідного програмного забезпечення.

Основні результати

Задача автоматизації визначення семантичних термінів у контенті навчальних матеріалів складається з ряду етапів перетворення інформації. Вхідними даними є контент навчальних матеріалів або його визначена частина; вихідними даними є ранжована множина семантичних термінів навчальних матеріалів.

Інформаційна технологія автоматизованого визначення термінів у навчальних матеріалах. На рис. 1 показано функціональну діаграму (виконану за стандартом IDEF0 [14]), яка ілюструє послідовність дій при формуванні множини термінів у контенті навчальних матеріалів.

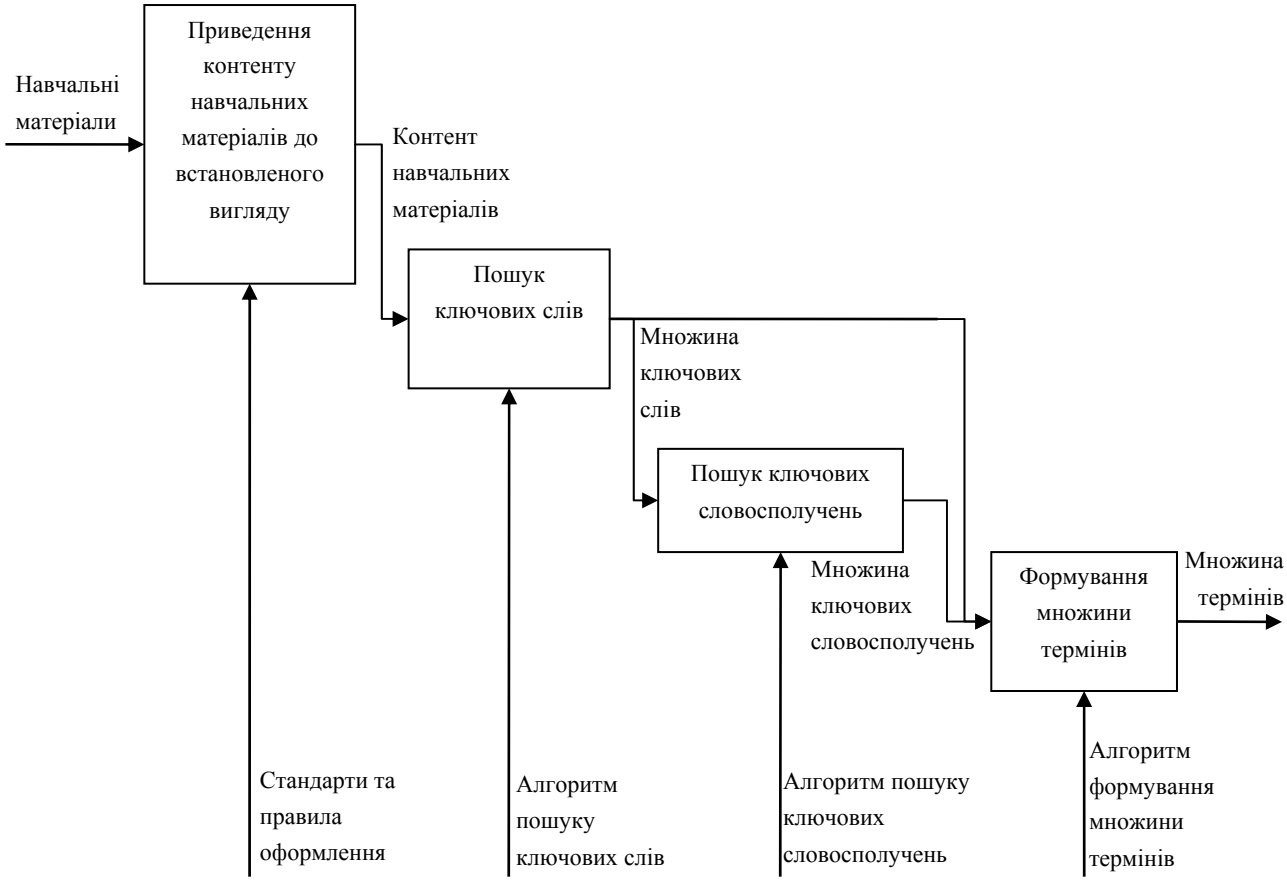

Рис. 1. Діаграма етапів пошуку семантичних термінів у контенті навчальних матеріалів

Кожному етапу з наведених відповідає деяка послідовність перетворення даних, що в сукупності формують інформаційну технологію автоматизованого визначення семантичних термінів, подану на рис. 2.

Попередня технічна обробка тексту (Блок 1) полягає в усуненні неоднозначного іменування термінів (наприклад, «СКБД», «Система керування БД», «Система керування базами даних») та обробці розділових знаків.

Блок 2 (*пошук ключових слів у контенті навчального матеріалу)* проводиться з використанням методу дисперсійного оцінювання, який показав свою достатню ефективність у рамках попередніх досліджень [10].

Дисперсійна оцінка є оцінкою дискримінантної сили слів й дозволяє відділити із загальної множини широковживаних у тексті слів слова, що розташовані рівномірно. Якщо деяке слово A в тексті, що складається з N слів, позначене як A_k^n, де індекс k – номер появи даного слова в тесті, а n – позиція даного слова в тексті, то інтервал між послідовними появами слова при таких позначеннях буде величина

$$\Delta A_k^m = A_{k+1}^m - A_k^n = m - n ,$$

де на *m*-ій і *n*-ій позиціях в тексті знаходиться слово *A*, яке зустрілось *k*+1-ий і *k*-ий рази. Тоді дисперсійна оцінка розраховується наступним чином [11]:

$$\sigma = \frac{\sqrt{(\Delta A^2) - (\Delta A)^2}}{(\Delta A)}, \tag{1}$$

де (ΔA) – середнє значення послідовності $\Delta A_1, \Delta A_2, \Delta A_k$; (ΔA^2) – послідовності A_1^2, A_2^2, A_k^2; K – кількість появи слова *A* в тексті.

Рис. 2. Схема інформаційної технології автоматизованого визначення семантичних термінів у контенті навчальних матеріалів

Для *формування вихідної множини слів* (Блок 3) слова у множині, сформованої у результаті дисперсійного оцінювання, сортуються за зменшенням значення дисперсійної оцінки й обмежуються за кількісним порогом. Термінами можуть виступати як слова, так і словосполучення й абревіатури. Якщо абревіатури з технічної точки зору розглядаються як слова, то словосполучення вимагають окремого алгоритму ідентифікації.

Словосполучення є стійкими сукупностями важливих слів, що згруповані у визначеній послідовності та у такій комбінації неодноразово присутні в розглядуваному контенті. Для *формування вихідної множини словосполучень* (Блок 4) проводиться пошук неперервних скупчень ключових слів протягом тексту й знайдені зразки фіксуються у масиві словосполучень за алгоритмом, показаним на рис. 3. Отриманий масив

словосполучень сортується за частотою вживання, після чого з нього видаляються неключові словосполучення.

Розглянутим чином, на основі множин ключових слів та ключових словосполучень формується узагальнена множина термінів, до якої входять словосполучення й ті слова, значення дисперсійної оцінки яких суттєво перевищує значення дисперсійної оцінки зв'язаних із цим словом колокацій (словосполучень). Об'єднана множина термінів сортується за значеннями їх дисперсійної оцінки.

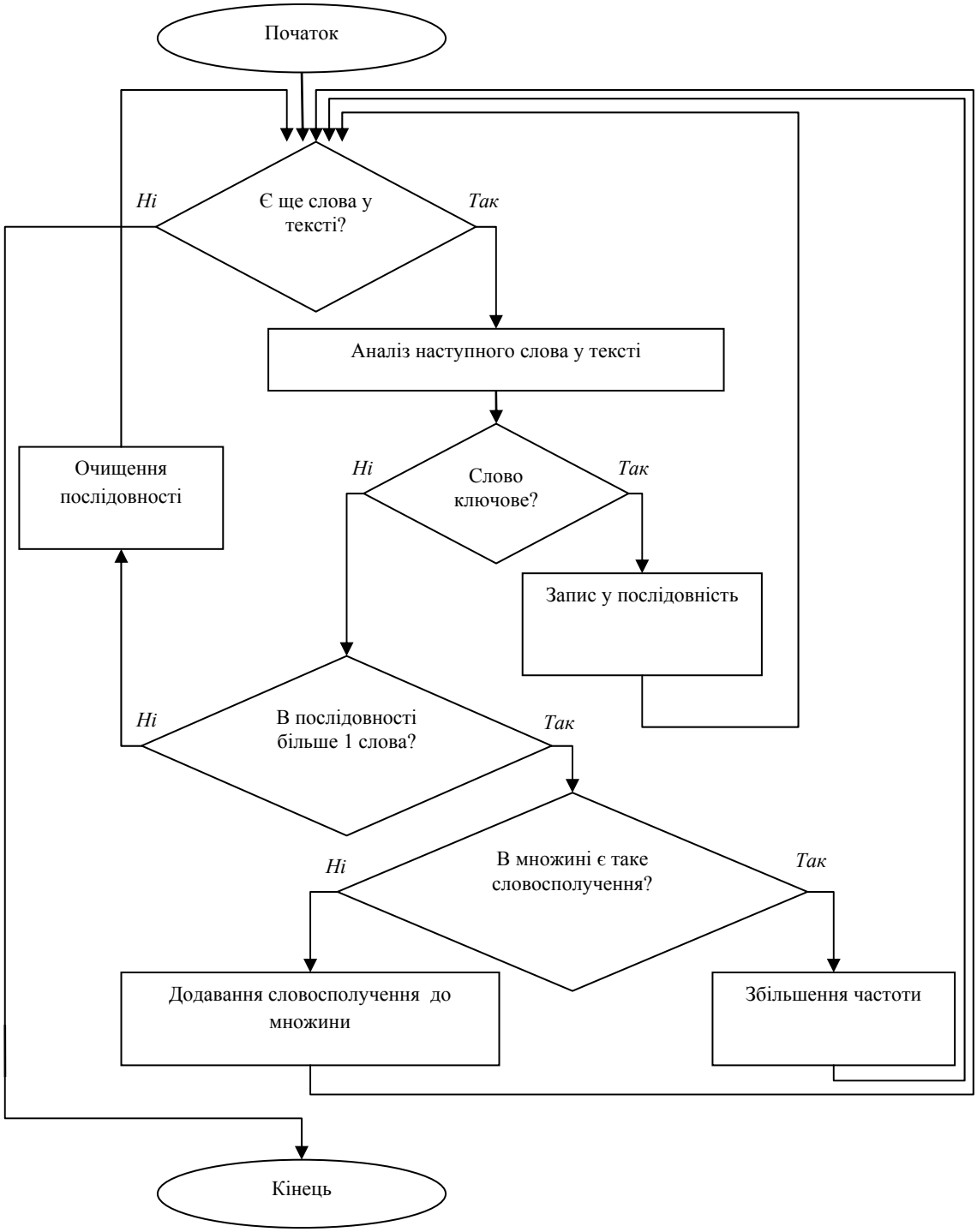

Рис. 3. Алгоритм пошуку сталих словосполучень у контенті навчальних матеріалів

Реалізація інформаційної технології. З метою перевірки ефективності розробленої інформаційної технології було проведене порівняння результату автоматизованого визначення ключових семантичних термінів зі списком, сформованим експертом (автором відповідних матеріалів).

Для автоматизованого формування множини ключових термінів авторами було розроблене тестове програмне забезпечення, що реалізує обробку контенту навчальних матеріалів викладеним вище чином. Структура програми складається з 6 модулів, кожен з яких у свою чергу складається з декількох класів (рис. 4).

MainForm – клас головної форма, що призначений для керування усім процесом знаходження ключових слів, словосполучень та об'єднання двох переліків у перелік ключових термінів.

Модуль, що відповідає за реалізацію основної логіки програми, складається з декількох класів (рис. 5), основні з яких наступні: Document – клас для збереження і обробки тексту документу; також клас реалізує перетворення ключових слів та ключових словосполучень і їх об'єднання; DispersiveEstimate – клас, що реалізує процес знаходження ключових слів; DBWorker – клас, що реалізує процес лематизації слів й інші функції, що пов'язані із базою даних.

Для зручного використання даних модулями програми, використовуються класи-представлення AnalysisWord та WordsRelations (рис. 6). Розроблена структура класів дозволяє реалізувати основні функції системи, такі як формування переліку ключових слів, формування переліку словосполучень та об'єднання їх у перелік ключових термінів.

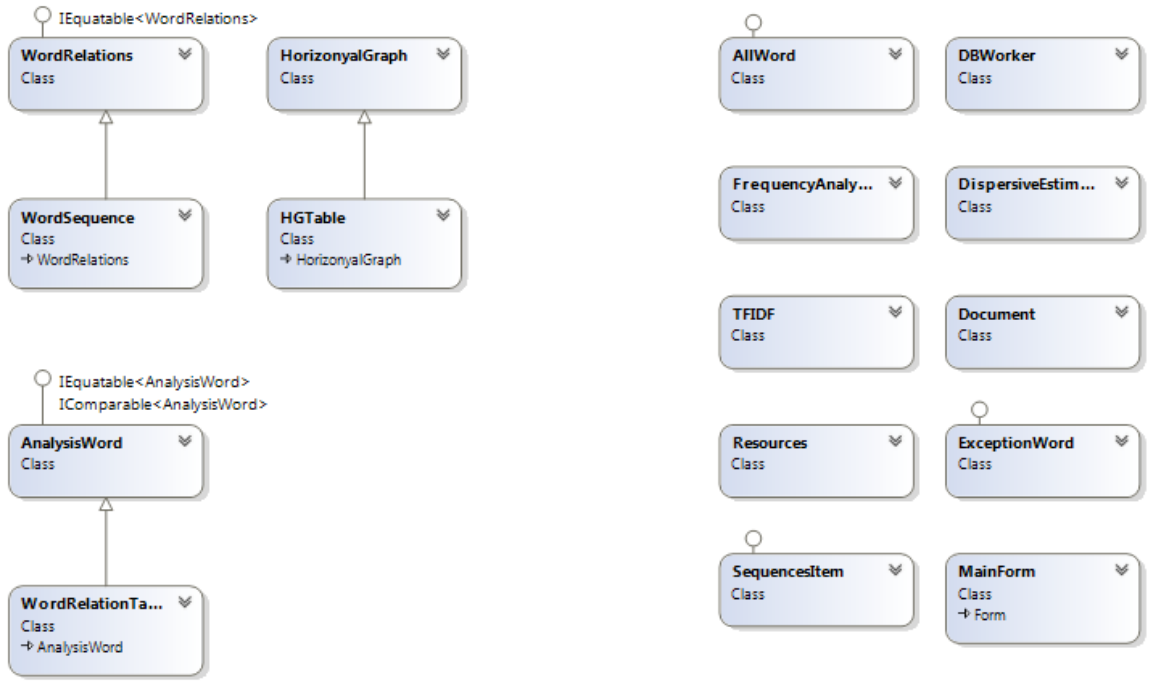

Рис. 4. Діаграма класів застосування з автоматизованого визначення ключових семантичних термінів

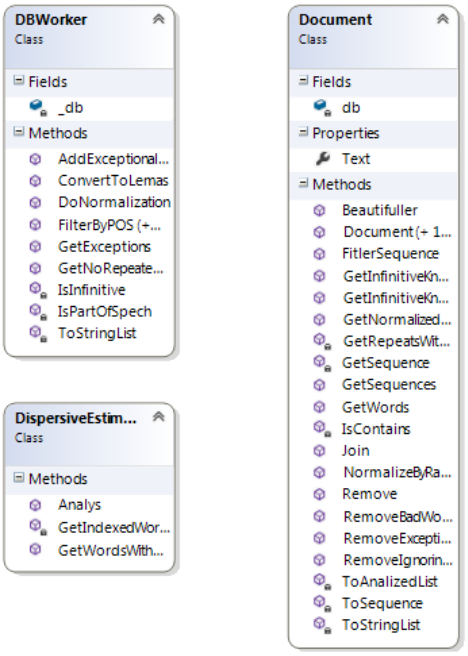

Рис. 5. Класи логіки програми

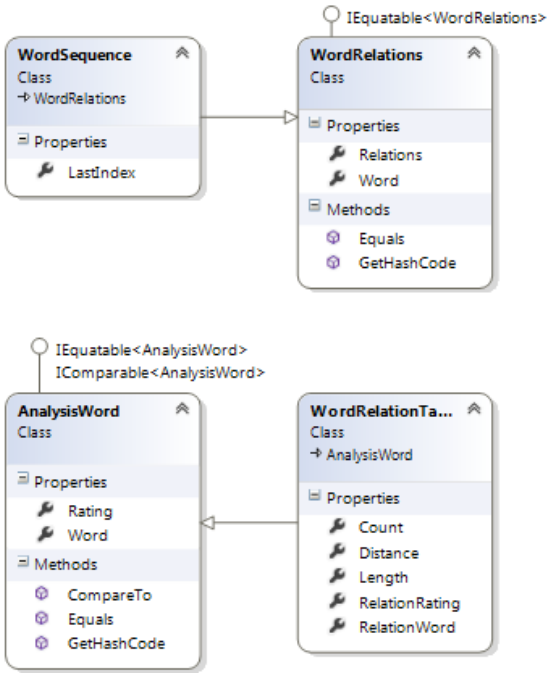

Рис. 6. Класи-представлення

Розроблений програмний продукт на основі введених даних у вигляді тексту документу (рис. 7) проводить його попередню обробку аналіз відповідно до розробленої інформаційної технології.

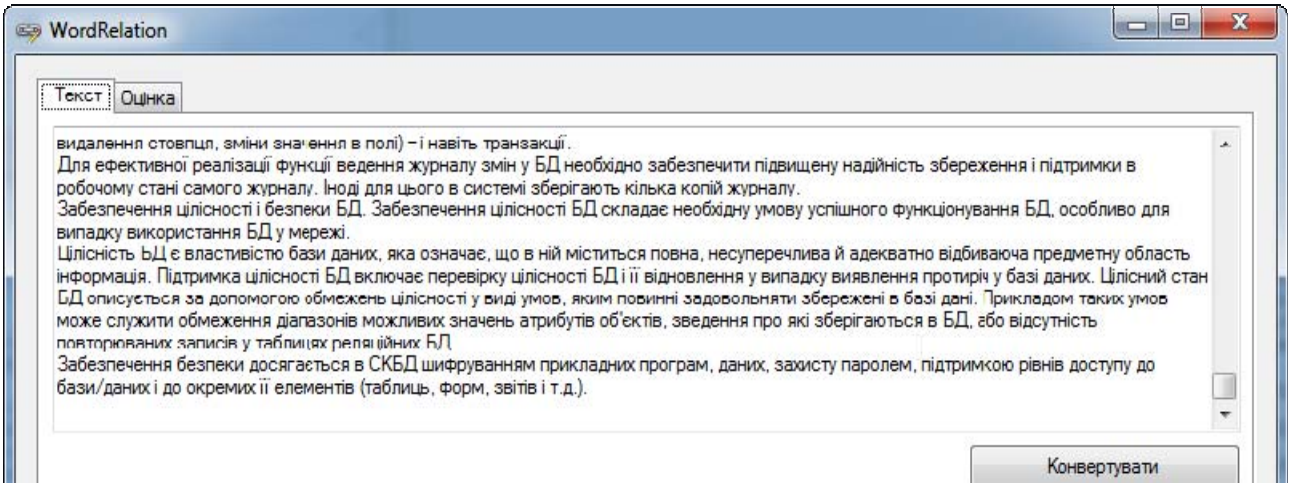

Рис. 7. Завантаження та попередня обробка тексту програмою

За наведеним алгоритмом формуються множина ключових слів та множина ключових словосполучень, а на їх основі формується узагальнена множина термінів (рис. 8).

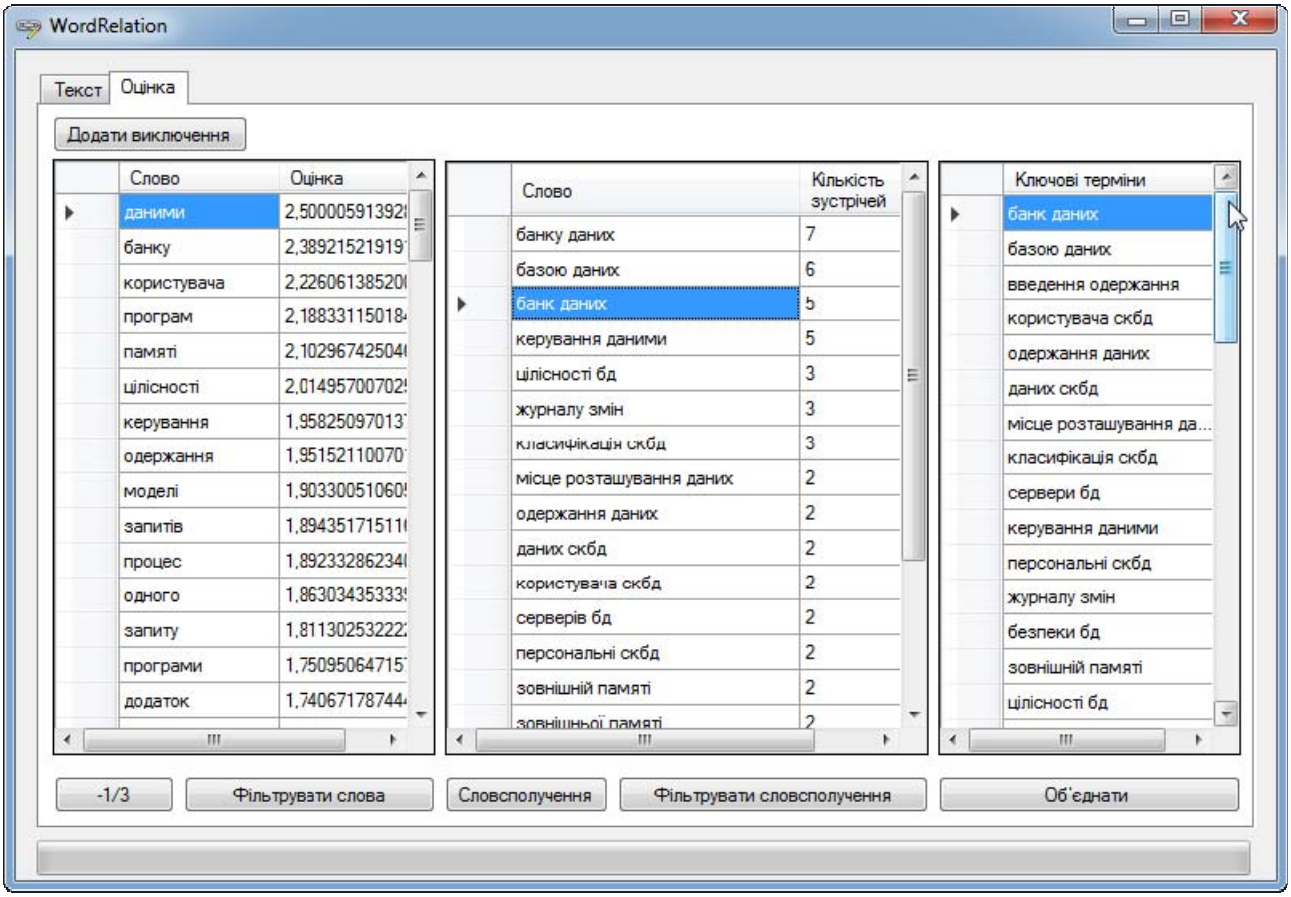

Рис. 8. Формування результуючої множина ключових термінів тексту

Експериментальні результати. В процесі обробки контенту кожного навчального матеріалу множина ключових слів, отримана за допомогою тестового програмного забезпечення, обмежуються за кількісним порогом й формує множину B_A. В подальшому ця множини порівнюються із множиною ключових термінів

B_E, яку сформовано експертом. Перетин $B_П$ цих множин $B_A \bigcap B_E$ визначає ефективність автоматизованого визначення ключових семантичних термінів у відповідному навчальному матеріалі.

Практична ефективність технології автоматизованого визначення термінів для кожного з матеріалів *k* визначається за формулою:

$$E_k = \frac{N_П}{N_E} \cdot 100\%, \tag{2}$$

де $N_П$ – кількість термінів у експертній (B_E) та сформованій автоматично (B_A) множині термінів, що співпали ($B_A \bigcap B_E$); N_E – кількість термінів у множині термінів B_E, сформованій експертом (автором).

Відповідно, середня практична ефективність розробленої інформаційної технології визначається за наступною формулою:

$$\overline{E} = \frac{\sum_{i=1}^{k} E_{Пk}}{k}, \tag{3}$$

де $E_{Пk}$ – ефективність технології визначення термінів для *k*-го матеріалу; *k* – кількість навчальних матеріалів у тестовій вибірці.

У рамках досліджень було оброблено вибірку з 30 лекцій із різних навчальних курсів. Наприклад, у результаті тестування лекційного матеріалу «Основні поняття й архітектура систем баз даних» навчального курсу «Організація баз даних та знань» розробленим програмним забезпеченням було отримано множину ключових термінів та проведено її порівняння з авторською множиною. Деякі результати порівняння наведено у таблиці. В даному випадку ефективність методу склала 80 %. Середня ж практична ефективність застосування розробленої інформаційної технології склала 87,3 %.

Таблиця. Фрагмент порівняльної таблиці аналізу множин термінів

№ п/п	**Термін**	**Дисперсійна оцінка**	**Визначено автоматично**	**Визначено автором**
1.	Банк даних	2,498457	+	+
2.	База даних	2,389215	+	+
3.	Класифікація СКБД	2,189637	+	+
4.	Сервер БД	2,095059	+	
5.	Цілісність БД	2,014957	+	+
6.	Запит	1,951622	+	+
7.	Багатокористувацькі СКБД	-		+
8.	Персональні СКБД	1,863380	+	+
9.	Система керування базою даних	1,811303	+	+
10.	Безпека БД	1,721403	+	+

Аналіз отриманих результатів виявив, що відсутність програмно визначених термінів у множині автора не завжди характеризує недолік запропонованої технології. Деякі семантично важливі терміни автори суб'єктивно ігнорують, в той час як іншу категорію складають поняття, на яких автори акцентують надмірну увагу попри їх другорядність у рамках матеріалу, що викладається. Тому, для різносторонньої оцінки результатів дослідження розглядалися не тільки терміни із множини експерта, що не були знайдені програмно, а й автоматично знайдені терміни, які не увійшли до множини експерта. Отже, як показано на

рис. 9, для вибірки з 30 навчальних лекцій, для яких проводився аналіз, із об'єднаних множин 87,3 % термінів співпало. Решту 7,5 % склали терміни експерта, що не мали відповідників у згенерованій множині, й 5,2 % – терміни, що входили до автоматично згенерованої множини, проте не були відзначені в множині термінів експерта.

Рис. 9. Діаграма середньої ефективності пошуку ключових термінів

Таким чином, практичне тестування розробленої інформаційної технології автоматизованого визначення семантичних термінів у контенті навчальних матеріалів виявило її середню ефективність 87,3 %, показавши при цьому мінімальну ефективність 65,1 % та максимальну – 100 %.

Висновки

Запропонована інформаційна технологія дозволяє з достатньою ефективністю автоматизовано визначати семантичні терміни в контенті навчальних матеріалів. Проведені за допомогою розробленого авторами тестового програмного забезпечення дослідження підтвердили здатність запропонованого підходу з ефективністю 87,3 % автоматизовано визначати ключові слова та словосполучення у контенті навчальних матеріалів.

Подальші дослідження спрямовані на пошук виключних випадків та вдосконалення розглянутої інформаційної технології для покращення результатів при їх обробці.

1. *Нові* інформаційні технології в освіті [Електронний ресурс]. – 2015. – Режим доступу: http://it-tehnolog.com/statti/novi-informatsiyni-tehnologiyi-navchannya/.
2. *Концепція* якості освіти. – [Електронний ресурс]. – 2015. – Режим доступу: http://osvita.ua/.
3. *University* of the People. – [Електронний ресурс]. – 2015. – Режим доступу: http://www.uopeople.org/.
4. *Факультет* заочно-дистанційного навчання, післядипломної освіти та довузівської підготовки ХНУ. – [Електронний ресурс]. – 2015. – Режим доступу: http://dn.tup.km.ua/.
5. *Аванесов В.С.* Композиция тестовых заданий. – М.: АДЕПТ, 1998. – 217 с.
6. *Moodle* – Open-source learning platform. – [Електронний ресурс]. – 2015. – https://moodle.org/
7. *Кірей К.О., Кірей Л.О.* До проблеми стандартизації термінології освітніх інформаційно-телекомунікаційних технологій // Вісник Черкаського університету. Сер.: Педагогічні науки. – Черкаси, 2009. – Вип. 146. – С. 27–29.
8. *Снитюк В.Е., Юрченко К.Н.* Интеллектуальное управление оцениванием знаний. − Черкассы, 2013. – 262 с.
9. *IDEF5* – Ontology Description Capture Method. – [Електронний ресурс]. – 2015. – Режим доступу: http://www.idef.com/IDEF5.htm.
10. *Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M.* Keyword detection in natural languages and DNA // Europhys. Lett, 2002. – 57(5). – P. 759–764.
11. *Ventura J., Silva, J.* New Techniques for Relevant Word Ranking and Extraction // Proceedings of the aritficial intelligence 13th Portuguese conference on Progress in artificial intelligence, EPIA'07. – Berlin: Springer-Verlag, Berlin, Heidelberg, 2007. – P. 691–702.
12. *Ландэ Д.В., Снарский А.А.* Компактифицированный горизонтальный граф видимости для сети слов // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения» – КПИ, Киев: 2013. – С. 158–164.
13. *Бармак О.В., Мазурець О.В.* Методи автоматизації визначення семантичних термінів у навчальних матеріалах // Вісник Хмельницького національного університету. Сер.: Технічні науки. Хмельницький. – 2015. – № 2(223). – С. 209–213.
14. *Марка Д., МакГоуэн К.* Методология структурного анализа и проектирования. Пер. с англ. – М.: МетаТехнология, 1993. – 240с.

References

1. New Information Technologies in Education. URL: http://it-tehnolog.com/statti/novi-informatsiyni-tehnologiyi-navchannya/.
2. Koncepcia Yacosti Osvityu. URL: http://osvita.ua/.
3. University of the People. URL: http://www.uopeople.org/.
4. Facultet Zaochno-Distanciynogo Navchannya, Pislyadiplomnoi Osvity ta Dovuzivskoi Pidgotovki KhNU. URL: http://dn.tup.km.ua/.
5. Avanesov V.S. Kompozicia Testovih Zadaniy. – M., Centr Testirovanya, 2002.
6. Moodle – Open-source learning platform. URL: https://moodle.org/.
7. Do Problemi Standartizacii Terminologii Osvitnih Informaciyno-Telekomunikaciynih Tehnologiy / K.O. Kirey, L.O. Kirey // Visnik Cherkaskogo Universitetu / Cherkaskiy Nacionalnogo Universitet im. Bogdana Khmeinitskogo. – Cherkasy, 2009. Ser.: Pedagogichni Nauki, Vip. 146. – P. 27–29.
8. Snituk V.E., Yurchenko K.N. Intelektualnoe Upravlenie Ocenivaniem Znaiy. − Cherkassy, 2013. – 262 p.
9. IDEF5 – Ontology Description Capture Method. URL: http://www.idef.com/IDEF5.htm.
10. Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. Keyword detection in natural languages and DNA // Europhys. Lett, 2002. – 57(5). – P. 759–764.
11. Ventura, J. & Silva, J. (2007). New Techniques for Relevant Word Ranking and Extraction. In Proceedings of 13th Portuguese Conference on Artificial Intelligence, Springer-Verlag, P. 691–702.
12. Lande D.V., Snarskiy A.A. Kompaktificirovanniy Gorizontalniy Graf Vidimosti dlya Seti Slov // Trudi Mejdunarodnoy Nauchnoy Konferencii «Intellektualniy Analiz Informacii IAI-2013. Znania I Rassujdenia» – KPI, Kiev: 2013. – P. 158–164.
13. O.Barmak, O.Mazurets, Methods of Automation of Definition of Semantic Terms in Educational Materials // Naukoviy jurnal "Visnik Khmelnitskogo Nacionalnogo Universitetu" ser.: Tehnichni nauki. Khmelnitsky, 2015, №2(223). – P. 209–213.
14. David Marka, Kliment McGouen, Metodologia Structurnogo Analiza i Proectirovania. Per. s angl. M.: 1993, 240 p.

Про авторів:

[1]*Крак Юрій Васильович,*
доктор фізико-математичних наук, професор,
завідувач кафедри Теоретичної кібернетики.
Кількість наукових публікацій в українських виданнях – 397.
Кількість наукових публікацій в іноземних журналах – 53.
Індекс Гірша – 2.
http://orcid.org/0000-0002-8043-0785,

[2]*Бармак Олександр Володимирович,*
доктор технічних наук, професор,
професор кафедри Комп'ютерних наук та інформаційних технологій.
Кількість наукових публікацій в українських виданнях – 88.
Кількість наукових публікацій в іноземних журналах – 12.
Індекс Гірша – 2.
http://orcid.org/0000-0003-0739-9678,

[2]*Мазурець Олександр Вікторович,*
старший викладач кафедри Комп'ютерних наук та інформаційних технологій.
Кількість наукових публікацій в українських виданнях – 60.

Місце роботи авторів:

[1]Київський національний університет імені Тараса Шевченка,
01601, Київ, вул. Володимирська, 60.
E-mail: krak@unicyb.kiev.ua,
yuri.krak@gmail.com,

[2]Хмельницький національний університет МОН України,
29016, Хмельницький, вул. Інститутська, 11.
E-mail: alexander.barmak@gmail.com,
exe.chong@gmail.com