УДК 004.912

ПРАКТИЧНА РЕАЛІЗАЦІЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ МНОЖИНИ СЕМАНТИЧНИХ ТЕРМІНІВ В КОНТЕНТІ НАВЧАЛЬНИХ МАТЕРІАЛІВ

Ю.В. Крак, О.В. Бармак, О.В. Мазурець

Досліджено проблему автоматизації пошуку ключових термінів у контенті навчальних матеріалів. Розглянуто інформаційну технологію автоматизованого визначення множини ключових семантичних термінів у контенті навчальних матеріалів, що ґрунтується на пошуку використаних фраз у тексті та дисперсійній оцінці важливості слів. Відповідно до даної інформаційної технології, на основі введених даних у вигляді файлу навчального матеріалу автоматизовано формується структура цифрового документу для вибору елементу для аналізу, після чого проводиться сегментація по фразах і термінах, терміни лематизуються та їх множина компактифікується. На основі автоматично лематизованого тексту проводиться пошук та дисперсійне оцінювання важливості слів у обраному фрагменті, після чого оцінюється важливість термінів, а їх кількість обмежується відповідно до коефіцієнту щільності ключових слів. Вхідними даними інформаційної технології є цифровий документ навчального матеріалу, вихідними даними є відповідна множина ключових семантичних термінів навчального матеріалу. Також описано результати аналізу закономірностей існуючих множин ключових семантичних термінів.
Розглянуто тестовий програмний продукт, що дозволяє автоматизовано визначати множину ключових семантичних термінів за даною інформаційною технологією. Проведені дослідження підтвердили можливість ефективно формувати множини ключових семантичних термінів навчальних матеріалів з показниками точності пошуку до 92,9 % та повноти пошуку до 100,0 %. Розглянуто практичні особливості використання спеціалізованого розширення при роботі з електронними документами. Викладено 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, автоматизація формування рефератів та анотацій до елементів навчальних матеріалів тощо.
Подальші дослідження спрямовані на аналіз впливу на показники ефективності технології взаємозв'язку між кількістю ключових семантичних термінів в результуючій множині та значенням коефіцієнту щільності ключових слів та вдосконалення розглянутої інформаційної технології для покращення результатів.
Ключові слова: цифровий документ, навчальні матеріали, ключові терміни, дисперсійна оцінка.

Исследовано проблему автоматизации поиска ключевых терминов в контенте обучающих материалов. Рассмотрено информационную технологию автоматизированного определения множества ключевых семантических терминов в контенте обучающих материалов, основанную на поиске использованных фраз в тексте и дисперсионной оценке важности слов. Согласно данной информационной технологии, на основе введенных данных в виде файла обучающего материала автоматизировано формируется структура цифрового документа для выбора элемента для анализа, после чего проводится сегментация по фразам и терминам, термины лемматизируются и их множество компактифицируется. На основе автоматически лематизированного текста производится поиск и дисперсионная оценка важности слов в выбранном фрагменте, после чего оценивается важность терминов, а их количество ограничивается в соответствии с коэффициентом плотности ключевых слов. Входными данными информационной технологии является цифровой документ обучающего материала, выходными данными является соответствующее множество ключевых семантических терминов обучающего материала. Также описаны результаты анализа закономерностей существующих множеств ключевых семантических терминов.
Рассмотрен тестовый программный продукт, позволяющий автоматизировано определять множество ключевых семантических терминов по данной информационной технологии. Проведенные исследования подтвердили возможность эффективно формировать множества ключевых семантических терминов обучающих материалов с показателями точности поиска до 92,9 % и полноты поиска до 100,0 %. Рассмотрены практические особенности использования специализированного расширения при работе с электронными документами. Изложены факторы, затрудняющие эффективное определение семантических терминов в учебных материалах. Определенная эффективность предложенной технологии способствует ее использованию для решения ряда актуальных задач, таких как оценка соответствия обучающих материалов требованиям, оценка соответствия наборов тестовых заданий обучающим материалам, помощь при создании тестов, автоматизация формирования рефератов и аннотаций к элементам обучающих материалов и прочие.
Дальнейшие исследования направлены на анализ влияния на показатели эффективности технологии взаимосвязи между количеством ключевых семантических терминов в результирующем множестве и значением коэффициента плотности ключевых слов и совершенствования рассмотренной информационной технологии для улучшения результатов.
Ключевые слова: цифровой документ, обучающие материалы, ключевые термины, дисперсионная оценка.

The problem of automation of key terms search in the content of educational materials is investigated. The information technology of automated determination of a set of key semantic terms in the content of educational materials is considered, which is based on the search of used phrases in the text and the disperse evaluation of words importance. In accordance with this information technology, on the basis of the data entered as an educational material file, the structure of a digital document is automatically formed to select an element for analysis, after which segmentation is performed by phrases and terms, the terms are lemmatized and set of them is compactified. On the basis of automatically lemmatized text, a search and disperse evaluation of the importance of words in the chosen fragment is performed, after which the terms importance is calculated, and their number is limited by the value of the keyword density ratio. Input data of information technology is a digital document of educational material, the output data is the corresponding set of key semantic terms of the educational material. The results of the analysis of the regularities of the existing sets of key semantic terms are also described.
The test software that allows to automate the determination of sets of key semantic terms using this information technology is considered. Conducted investigations confirmed the possibility of effectively forming the set of key semantic terms of educational materials, evaluated search precision metrics up to 92.9 % and search recall up to 100.0 %. The practical features of the use of specialized extension for working with electronic documents are considered. The factors that complicate effective search of semantic

terms in educational materials are described. The established effectiveness of the proposed technology allows use it to solution a number of urgent tasks, such as determination the conformity of educational materials to content requirements, determination the conformity of sets of test tasks to educational materials, semantic assistance in creating tests, automation of the creation of abstracts and annotations to the elements of educational materials, etc.
Further researches are aimed at analyzing the impact on the effectiveness of the technology of the relationship between the number of key semantic terms in the resulting set and the value of the keyword density ratio and improve of the information technology considered to improve the results.
Key word: digital document, key terms, educational materials, disperse evaluation.

Вступ та постановка задачі

Опис інформаційної технології. На сучасному етапі у галузі сучасної вищої освіти для розробки й використання курсів навчальних дисциплін використовуються спеціалізовані віртуальні навчаючі середовища, наприклад, Moodle. При їх використанні, потенційна якість отриманих освітніх послуг прямо залежить від якості навчальних матеріалів [1]. В умовах вузької спеціалізації курсів навчальних дисциплін, їх чисельності та інтенсивного оновлення, єдиним шляхом оцінки якості навчальних курсів та їх елементів є автоматизація вирішення відповідного ряду задач у галузі сучасної вищої освіти. До таких задач належать: оцінка відповідності навчальних матеріалів вимогам, оцінка відповідності наборів тестових завдань навчальним матеріалам, автоматизована генерація прототипів тестових завдань, допомога та контроль якості при формуванні навчальних матеріалів, допомога та контроль якості при формуванні тестів до навчальних матеріалів, реалізація гнучких алгоритмів тестування, автоматизація формування рефератів та анотацій до елементів навчальних матеріалів тощо.

Загальноприйнятим є підхід до застосування навчальних матеріалів у вигляді цифрових документів визначеної структури як інструменту навчання. Проте в усіх наведених випадках для досягнення відповідних результатів використовується не власне цифровий документ чи його контент, а його семантична модель. Формалізація побудови такої семантичної моделі забезпечується через застосування онтології як методу формального опису знань, що містяться в навчальних матеріалах [2]. Модель онтології навчального матеріалу може складатися з ключових слів, ключових термінів, структури навчального матеріалу, атрибутів ключових слів та ключових термінів, що визначають їх властивості та забезпечують прив'язку до елементів структури навчального матеріалу. За такої моделі, онтологія навчального матеріалу є засобом як для виявлення сенсу навчального матеріалу так і для вирішення наведеного ряду практичних задач.

Основними етапами побудови онтології навчального матеріалу є пошук ключових термінів у контенті навчального матеріалу та побудова його логічної структури. Вхідними даними є електронний документ навчального матеріалу, тому для автоматизації виконання наведених етапів потрібна програмна обробка відповідних цифрових файлів (зазвичай формату .docx). Проблему автоматизації побудови логічної структури навчального матеріалу (наприклад: Дисципліна / Розділ / Тема) пропонується вирішувати шляхом визначення ієрархії змістовних блоків у цифровому документі за стилями текстового редактора (Заголовок 1 / Заголовок 2 / Заголовок 3), таким чином формуючи верхній рівень вертикальної онтології відповідної навчальної дисципліни. Проблему пошуку ключових термінів у контенті навчального матеріалу пропонується вирішувати шляхом використання відповідної інформаційної технології, що забезпечить формування нижнього рівня онтології навчальної дисципліни.

Характерною особливістю елементів навчальних матеріалів, що використовуються для аналізу в процесі пошуку ключових термінів, є достатньо малий обсяг контенту. Малий обсяг контенту та вузька семантична направленість елементів аналізу зменшує ефективність застосування розповсюджених методів аналізу текстів, таких як частотна оцінка TF, оцінка TFIDF та дисперсійна оцінка DE [3]. Це обумовлює потребу в розробці спеціалізованої інформаційної технології, призначеної для автоматизованого визначення ключових термінів у контенті навчальних матеріалів.

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

Основні результати

При автоматизованому визначенні множини семантичних термінів у контенті навчальних матеріалів *вхідними* даними є контент навчального матеріалу або його визначена частина у вигляді файлу .docx довільної ієрархії елементів; *вихідними* даними є множина семантичних термінів навчального матеріалу; процес автоматизованого визначення множини семантичних термінів складається з ряду етапів перетворення інформації.

За результатами аналізу понад 1300 елементів навчальних матеріалів із визначеними експертом (укладачем) репрезентативними множинами ключових термінів, встановлено, що всі елементи наведених множин M_T відповідають наступним закономірностям:

- кількість слів у терміні $n = 1...6$;
- якщо термін є словом ($n = 1$), то воно входить до множини іменників M_I ;

– якщо термін є словосполученням ($n>1$), то до його складу входять елементи множини M_M . До складу множини M_M входять множини семантично значущих елементів (іменників M_I та прикметників $M_{ПК}$) та семантично зв'язуючих елементів (сполучників M_C , часток $M_Ч$ та прийменників $M_{ПЙ}$);

– якщо $n>1$, то до складу словосполучення входить принаймні один елемент із множини іменників M_I ;

– якщо $n>1$, то першим ($k=1$) та останнім ($k=n$) словом є елементи множини семантично значущих елементів $M_I \cup M_{ПК}$;

– якщо $n>1$, то між елементами словосполучення відсутні розділові знаки (окрім дефісу всередині складних іменників, який є частиною слова);

– всі елементи (символи, слова) одного терміна в тексті мають однакові стильові властивості, відповідно в структурі цифрового документу не виходять за межі контейнеру TextRange.

В результаті використання розроблюваної інформаційної технології ставиться за мету отримання множин термінів M_T , які відповідають наведеним закономірностям.

На рис. 1 подано схему інформаційної технології автоматизованого визначення множини семантичних термінів у контенті навчальних матеріалів, що висвітлює послідовність етапів перетворення даних для досягнення кінцевої мети.

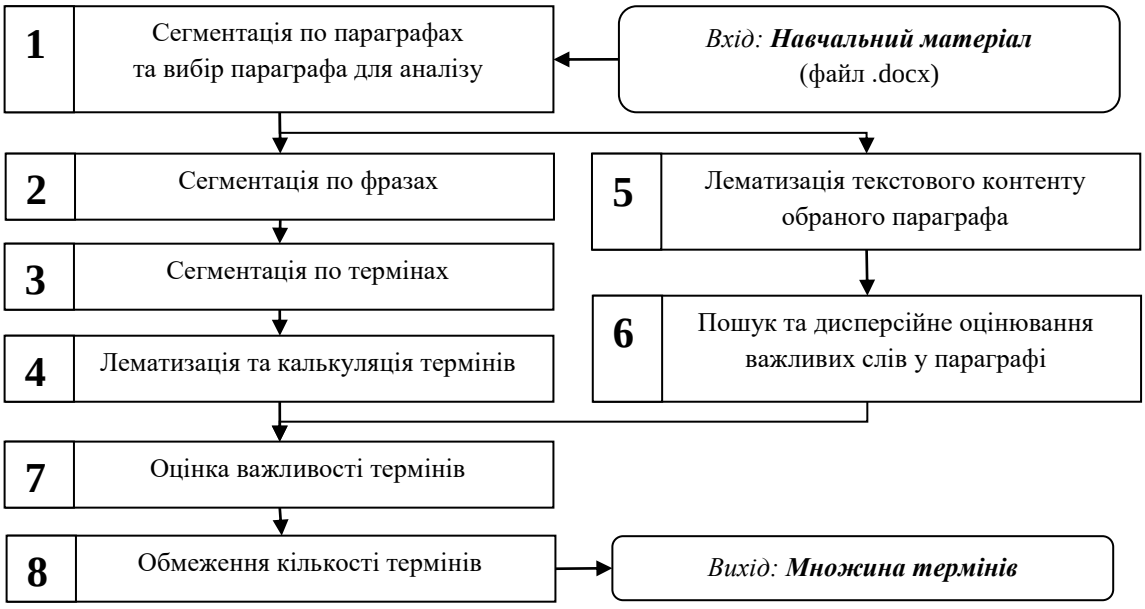


Рис.1. Схема інформаційної технології автоматизованого визначення множини семантичних термінів у контенті навчальних матеріалів

Сегментація по параграфах та вибір параграфу для аналізу (Блок 1) полягає в аналізі структури цифрового документу. Зважаючи на існуючі загальноприйняті вимоги до структури навчальних матеріалів навчальних дисциплін (зокрема: Назва дисципліни / Розділ / Тема), можна зробити висновок про природню відповідність ієрархічної системи заголовків навчальних матеріалів як електронних документів верхнім рівням семантичної структури навчального матеріалу дисципліни. Наприклад, назви дисциплін відповідатимуть елементам стандартного стилю «Heading 1», назви розділів – «Heading 2», назви тем – «Heading 3» тощо (табл. 1). Таким чином, структура навчальних матеріалів як цифрових документів регламентується мовами розмітки цифрових документів й реалізується через систему заголовків. Оскільки обсяг охоплення визначеним навчальним матеріалом відповідної навчальної дисципліни та глибина формування ієрархії наперед невідомі, є доцільним використання рекурсивних конструкцій даталогічних моделей для реляційного збереження даних (назва та підпорядкованість) верхніх рівнів семантичної структури навчальних матеріалів. На рис. 2 модель Headings включає елементи: ID (унікальний ідентифікатор – порядковий номер запису), Name (назва елементу ієрархії навчального матеріалу), Level (цифра рівню ієрархії навчального матеріалу – наприклад, для назви дисципліни Headings(Level)=1), Sequence (цифра, що визначає послідовність даного елементу серед елементів такого ж рівня в межах одного надрівня), Heading_ID (код рівня-«батька», для кореневого Headings(ID)=Headings(Heading_ID), посилання на надрівень).

Таблиця 1. Приклад відповідності верхніх рівнів семантичної структури навчальних матеріалів стандартним стилям цифрових документів

Порядок в ієрархії	**Рівень онтології навчальних матеріалів**	**Назва стандартного стилю цифрового документу**
1	Навчальна дисципліна	Heading 1
2	Розділ	Heading 2
3	Тема	Heading 3

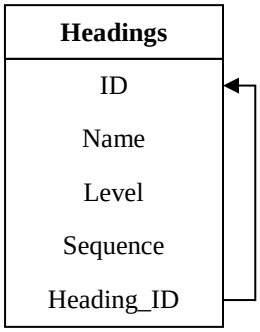


Рис. 2. Модель для збереження даних структури навчальних матеріалів

Вихідними даними Блоку 1 є визначений фрагмент контенту цифрового документу навчального матеріалу, над яким буде проводитись подальша обробка.

Блок 2 (*Сегментація по фразах*) проводиться з метою розбиття фрагменту контенту цифрового документу, що обробляється, на менші фрагменти – фрази. Під фразою мається на увазі семантично цілісний вузол, що виокремлений стилістичним форматуванням тексту чи розділовими знаками, й локалізує місцезнаходження окремих термінів. Відповідно до об'єктної моделі документу, MS Office використовує розділи (Section), щоб вказати частини документа, що мають відмінне форматування. Об'єкти Section містяться в об'єкті Document (рис. 3), в колекції Selections. Розділи (Section) містять в собі менші елементи структури – абзаци (Paragraph). TextRange є найнижчим рівнем структури документу, що визначає фрагмент тексту однакового стилю в межах Paragraph.

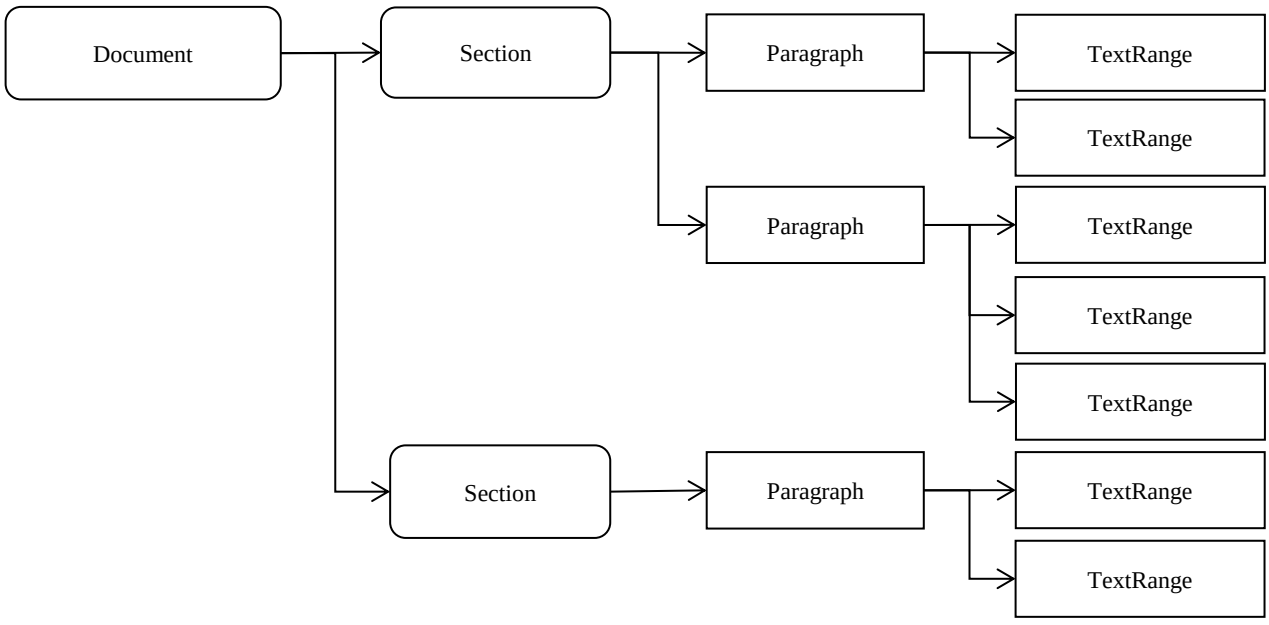


Рис. 3. Загальна структура об'єктної моделі документу MS Office

Так технічно до множини фраз включаються неперервні впорядковані послідовності слів, що не виходять за межі контейнерів цифрового документу TextRange та не перериваються розділовими знаками. Одержання в результаті виконання блоку множини фраз дозволяє в подальшому опрацьовувати на предмет пошуку термінів кожну з фраз поокремо.

Блок 3 (*Сегментація по термінах*) ставить за мету формування множини всіх можливих термінів, що присутні у досліджуваному контенті.

Таким чином, до множини термінів навчального матеріалу M_T включаються всі можливі неперервні впорядковані послідовності слів, що не виходять за межі фраз та відповідають умові:

$$M_T = \{\langle x_1, x_2, x_3, x_4, x_5, x_6\rangle|_1 \in M_I \cup M_{\Pi}, x_2 \in M_M, x_3 \in M_M, x_4 \in M_M, x_5 \in M_M, x_6 \in M_M,$$

$$\langle x_1, x_2, x_3, x_4, x_5, x_6\rangle \cap M_I \neq \varnothing\},$$

де M_M – множина семантично значущих елементів (іменників M_I та прикметників $M_{ПК}$) та семантично зв'язуючих елементів (сполучників M_C, часток $M_Ч$ та прийменників $M_{ПЙ}$),

$$M_M = M_I \cup M_{ПК} \cup M_C \cup M_Ч \cup M_{ПЙ} \cup \varnothing.$$

Сегментація по термінах проводиться з використанням бази даних корпусу слів української мови та в якості вихідних даних формує множину термінів M_T, що містяться в оброблюваному фрагменті цифрового документу навчального матеріалу.

Блок 4 (*Лематизація та калькуляція термінів*) дозволяє на основі множини термінів M_T сформувати множину лемо-незалежних термінів M_{T1} і співставити кожному з них кількість зустрічань у досліджуваному тексті. Для цього спершу проводиться лематизація кожного слова у кожній фразі в множині M_T. Під лематизацією мається на увазі приведення слів до, інфінітивного стану – наприклад, іменники переводяться у називний відмінок однини. Після чого одержана множина обробляється й компактифікується таким чином, що всі ідентичні повторення термінів видаляються, а кожному терміну співставляться величина K_n, що відображає встановлену кількість появ даного терміну n у вхідній множині M_T.

Оскільки на етапі формування множини термінів M_T до неї додавались усі можливі варіанти термінів в межах фраз без поглинання більшими словосполученнями менших, в даному блоці проводиться аналіз необхідності такого поглинання. Якщо в множині M_{T1} існує термін n_1 (K_{n1} – кількість появ терміну n_1 в множині M_{T1}), що є впорядкованою множиною з x_1 слів, та термін n_2 (K_{n2} – кількість появ терміну n_2 в множині M_{T1}), що є впорядкованою множиною з x_2 слів, причому n_1 є підмножиною n_2 й $x_1 < x_2$, то при вірності виразу $2x_1 > x_2$ термін видаляється з результуючої множини. З метою спрощення подальшої обробки із одержаної множини M_{T1} доцільно також видаляти всі терміни, в яких $K_n = 1$, оскільки однократне використання терміну виключає факт цілеспрямованого розгляду відповідного поняття в структурній одиниці навчального матеріалу.

Отримана в результаті множина лемо-незалежних термінів M_{T1} містить терміни, що використовуються у навчальному матеріалі з кількісним показником використання, проте не визначає важливість даних термінів.

Блок 5 (*Лематизація текстового контенту обраного параграфу*) переводить текст визначеного фрагменту контенту цифрового документу навчального матеріалу, що аналізується, до відповідної послідовності слів у інфінітивному стані, що є вихідними даними цього блоку. Вони дозволяють проводити подальше оцінювання дисперсії слів.

Блок 6 (*Пошук та дисперсійне оцінювання важливих слів у параграфі*) призначений для оцінки важливості кожного слова в досліджуваному тексті, що проводиться з використанням методу дисперсійного оцінювання [4], який є оцінкою дискримінантної сили слів. Метод дисперсійного оцінювання дозволяє відділити із загальної множини широковживаних у тексті слів слова, що розташовані рівномірно й показав свою високу ефективність у попередніх дослідженнях [5].

Відповідно до існуючої математичної моделі [6], 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. Таким чином, дисперсійна оцінка розраховується за формулою

$$\sigma = \sqrt{(\Delta A^2) - (\Delta A)^2} \Big/ (\Delta A),$$

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.

Вхідними даними блоку є лематизований текстового контент визначеного фрагменту контенту цифрового документу навчального матеріалу, вихідними даними – впорядкована множина слів, кожному з яких співставлена оцінка його дисперсії, що позиціонується як оцінка важливості даного слова у досліджуваному фрагменті цифрового документу.

Блок 7 (*Оцінка важливості термінів*) вхідними даними має множину лемо-незалежних термінів M_{T1} із співставленою кожному з них кількістю зустрічань у досліджуваному тексті та впорядковану множину слів із співставленою кожному з них оцінкою його важливості (дисперсії) у досліджуваному тексті.

Оцінка важливості v_n кожного терміна n із множини M_{T1} обчислюється за формулою:

$$v_n = \sum_{i=1}^{x_n} \frac{K_n \sigma_n}{k_n}, \qquad (1)$$

де K_n – кількість появ терміну n в множині M_{T1}; k_n – кількість появ i-го слова терміну n в лематизованому текстовому контенті визначеного фрагменту цифрового документу; σ_n – дисперсійна оцінка для i-го слова терміну n; x_n – кількість слів у терміні n.

Вихідними даними блоку є множина лемо-незалежних термінів M_{T1} із співставленими кожному з них кількістю зустрічань у досліджуваному тексті та значенням оцінки важливості, впорядкована за спаданням номінального значення оцінки важливості.

Блок 8 (*Обмеження кількості термінів*) призначений для формування множини ключових термінів за вхідними даними – множиною лемо-незалежних термінів M_{T1}. Множина ключових термінів формується на основі лемо-незалежних термінів із множини M_{T1} з найбільшими значеннями оцінки важливості, а їх кількість випливає із визначення відомого показника з семантичної обробки текстів, щільності ключових слів [7]. Щільність ключових слів є відношенням кількості слів ключових термінів в тексті до загальної кількості слів у тексті й для навчальних матеріалів становить 6–8 %. Відповідно, до порожньої результуючої множини ключових термінів M_{TK} додаються терміни з множини M_{T1} з найбільшими значеннями оцінки важливості доти, доки справджується рівність:

$$\sum_{i=1}^{n} \frac{K_n x_n}{X_{txt}} \le 0{,}07, \qquad (2)$$

де K_n – кількість появ терміну n в множині M_{T1}; x_n – кількість слів у терміні n; X_{txt} – загальна кількість слів у тексті; n – поточна кількість термінів у множині M_{TK}.

Вихідними даними блоку й відповідно інформаційної технології є множина M_{TK} ключових термінів, відповідна досліджуваному фрагменту контенту цифрового документу навчального матеріалу.

Таким чином, запропонована інформаційна технологія автоматизованого визначення множини семантичних термінів у контенті навчальних матеріалів дозволяє на основі цифрового документу навчального матеріалу автоматизовано отримувати відповідну множину ключових термінів.

Реалізація інформаційної технології. З метою перевірки ефективності розробленої інформаційної технології автоматизованого визначення множини семантичних термінів у контенті навчальних матеріалів було проведене порівняння автоматизовано сформованої множини ключових семантичних термінів із множиною автора (експерта) для тестової вибірки цифрових документів навчальних матеріалів.

Відповідно до запропонованої інформаційної технології, було розроблене тестове програмне забезпечення, що реалізує обробку контенту цифрових документів навчальних матеріалів викладеним вище чином.

Цифрові файли навчальних матеріалів .docx організовані за допомогою відкритого формату XML, в якому зберігаються документи як колекції окремих файлів і папок в стиснутому пакеті. Для реалізації програмної обробки цифрових документів є доцільним використання спеціалізованих програмних комплексів, що надають об'єктно-орієнтований інструментарій для програмної роботи з контентом відповідних файлів, наприклад Microsoft.Office.Interap.Word.dll, DocumentFormat.OpenXml.dll та Spire.Doc.dll. В рамках розробленого тестового програмного забезпечення було використано розширення Spire.Doc.dll [8], яке забезпечило як аналіз рівнів структури документу Heading, так і доступ до елементів контенту, зокрема TextRange (рис. 4), який є найнижчим рівнем структури документу, що визначає фрагмент тексту однакового стилю. Перенесення функцій автоматичного співставлення стилів текстових блоків їх

властивостям з рівня функціоналу програмного коду застосунка на рівень функціоналу бібліотеки дозволило спростити як роботу системи з цифровим документом, так і процес програмування.

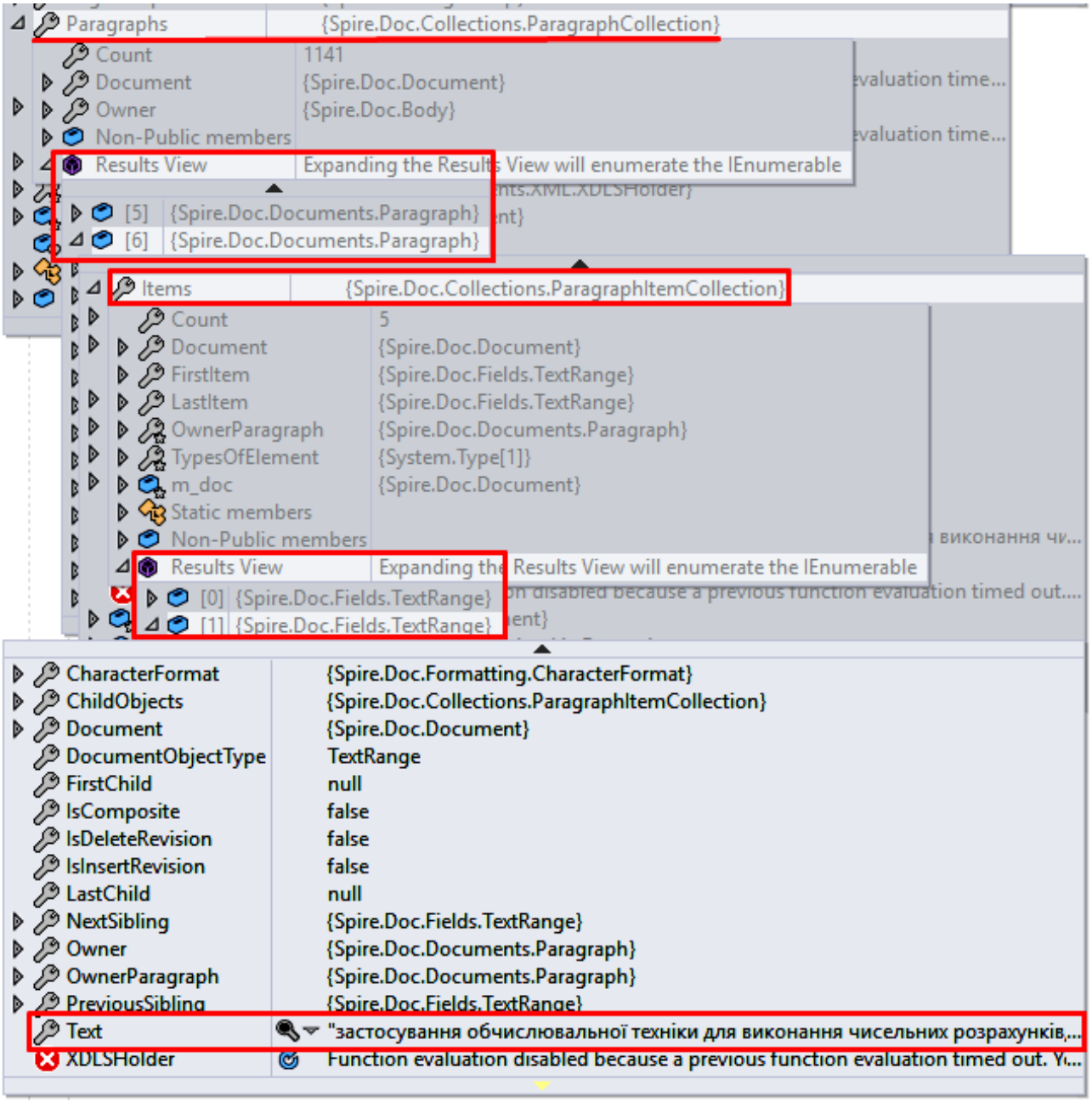
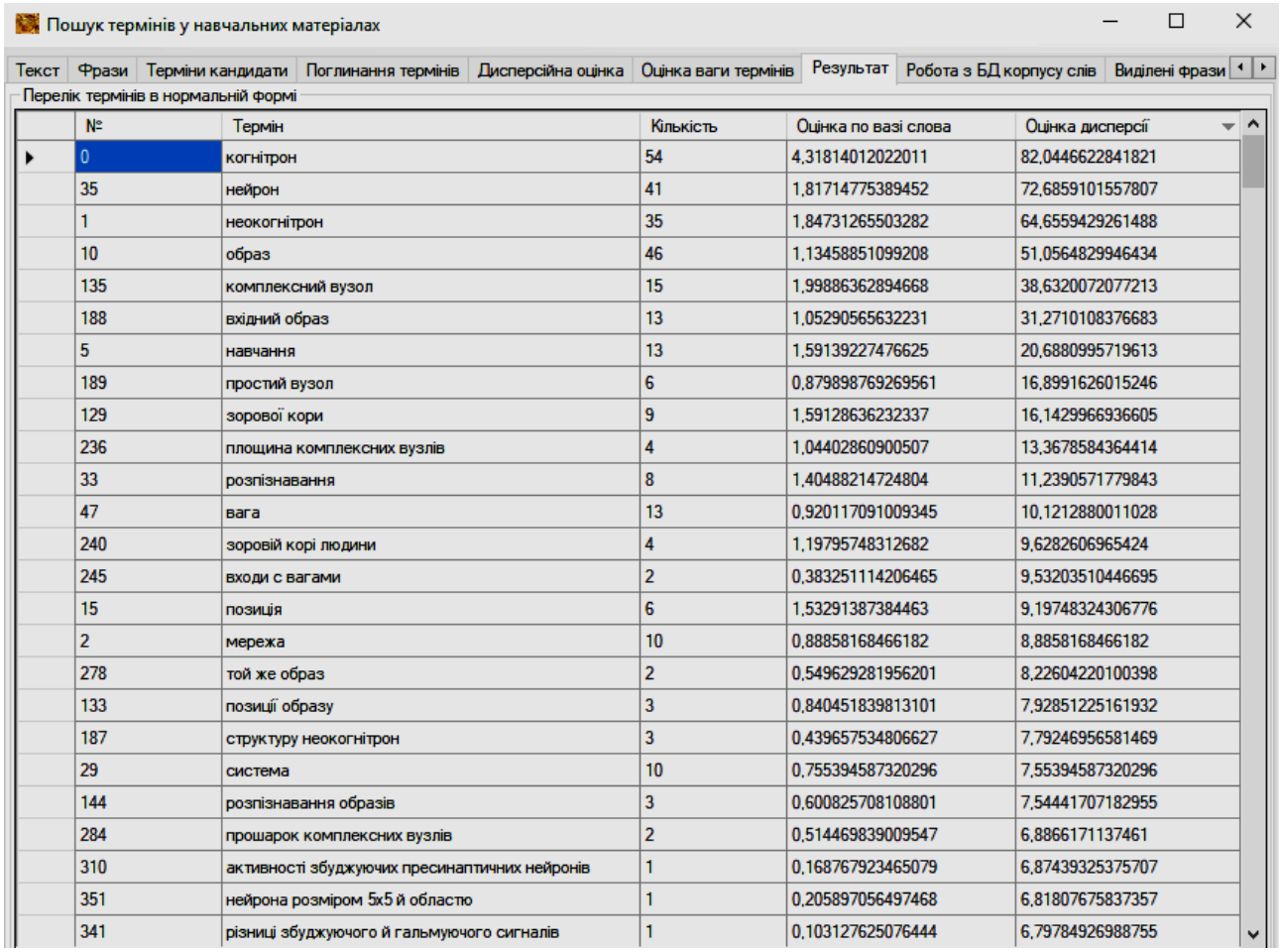


Рис. 4. Забезпечення доступу до елементів TextRange цифрового документу навчальних матеріалів за допомогою спеціалізованого розширення

Створений тестовий програмний продукт на основі введених даних у вигляді файлу навчального матеріалу автоматизовано формує структуру цифрового документу для вибору елементу для аналізу, після чого проводиться сегментація по фразах і термінах, терміни лематизуються та їх множина компактифікується, на основі автоматично лематизованого тексту проводиться пошук та дисперсійне оцінювання важливості слів у обраному фрагменті, після чого оцінюється важливість термінів, а їх кількість обмежується відповідно до вищенаведеної математичної моделі. Зокрема, на рисунку 5 показано приклад обробки теми «Нейронні мережі когнітрон та неокогнітрон» дисципліни «Методи та системи штучного інтелекту.

Кінцевим результатом роботи тестового програмного продукту є множина ключових термінів тексту. В розглянутому випадку (рис. 5) за показника щільності ключових слів 7 % до множини ключових термінів було віднесено: когнітрон, нейрон, неокогнітрон, образ, комплексний вузол, вхідний образ, навчання, простий вузол.



№	Термін	Кількість	Оцінка по вазі слова	Оцінка дисперсії
0	когнітрон	54	4,31814012022011	82,0446622841821
35	нейрон	41	1,81714775389452	72,6859101557807
1	неокогнітрон	35	1,84731265503282	64,6559429261488
10	образ	46	1,13458851099208	51,0564829946434
135	комплексний вузол	15	1,99886362894668	38,6320072077213
188	вхідний образ	13	1,05290565632231	31,2710108376683
5	навчання	13	1,59139227476625	20,6880995719613
189	простий вузол	6	0,879898769269561	16,8991626015246
129	зорової кори	9	1,59128636232337	16,1429966936605
236	площина комплексних вузлів	4	1,04402860900507	13,3678584364414
33	розпізнавання	8	1,40488214724804	11,2390571779843
47	вага	13	0,920117091009345	10,1212880011028
240	зоровій корі людини	4	1,19795748312682	9,6282606965424
245	входи с вагами	2	0,383251114206465	9,53203510446695
15	позиція	6	1,53291387384463	9,19748324306776
2	мережа	10	0,88858168466182	8,8858168466182
278	той же образ	2	0,549629281956201	8,22604220100398
133	позиції образу	3	0,840451839813101	7,92851225161932
187	структуру неокогнітрон	3	0,439657534806627	7,79246956581469
29	система	10	0,755394587320296	7,55394587320296
144	розпізнавання образів	3	0,600825708108801	7,54441707182955
284	прошарок комплексних вузлів	2	0,514469839009547	6,8866171137461
310	активності збуджуючих пресинаптичних нейронів	1	0,168767923465079	6,87439325375707
351	нейрона розміром 5x5 й областю	1	0,205897056497468	6,81807675837357
341	різниці збуджуючого й гальмуючого сигналів	1	0,103127625076444	6,79784926988755

Рис. 5. Отримання множини важливих термінів тестовим програмним продуктом

Експериментальні результати. Ефективність практичного застосування розглянутої інформаційної технології може бути оцінена шляхом використання відповідного тестового програмного продукту за показниками точності (Precision) та повноти (Recall) [9].

Точність пошуку P (відношення кількості релевантних ключових термінів, знайдених автоматично, до загальної кількості знайдених ключових термінів в досліджуваному тексті) та повнота пошуку R (відношення кількості релевантних ключових термінів, знайдених автоматично, до загальної кількості релевантних ключових термінів в досліджуваному тексті) обчислюються наступним чином:

$$P = \frac{\left| M_{TK}^{E} \cap M_{TK} \right|}{\left| M_{TK} \right|}, \; R = \frac{\left| M_{TK}^{E} \cap M_{TK} \right|}{\left| M_{TK}^{E} \right|}, \tag{3}$$

де M_{TK}^{E} – множина релевантних ключових термінів, сформована експертом; M_{TK} – множина знайдених автоматично ключових термінів.

Відповідно, середня точність пошуку $\overline{P}$ та середня повнота пошуку $\overline{R}$ визначаються за наступними формулами:

$$\overline{P} = \frac{\sum_{i=1}^{k} P_k}{k}, \; \overline{R} = \frac{\sum_{i=1}^{k} R_k}{k}, \tag{4}$$

де k – кількість навчальних матеріалів у тестовій вибірці.

З метою визначення ефективності практичного застосування розглянутої інформаційної технології, тестовим програмним продуктом було оброблено тестову вибірку з 50 файлів із різних навчальних курсів. Наприклад, у результаті тестування розглянутого вище навчального матеріалу «Нейронні мережі когнітрон та неокогнітрон» було отримано множину ключових термінів та проведено її порівняння з авторською множиною.

Результати порівняння наведено у табл. 2. В даному випадку точність пошуку склала 0,625, а повнота пошуку склала 0,714.

Середня точність пошуку склала 0,732, а повнота пошуку склала 0,697. Мінімальна точність пошуку одержана 0,512, мінімальна повнота пошуку – 0,581; максимальна точність пошуку – 0,929, максимальна повнота пошуку – 1,000.

Табл. 2. Порівняльна таблиця аналізу множин термінів

№ п/п	Ключовий термін	Визначено автором	Визначено автоматично
1.	Когнітрон	+	+
2.	Неокогнітрон	+	+
3.	Нейрон	+	+
4.	Збуджуючий нейрон	+	
5.	Гальмуючий нейрон	+	
6.	Комплексний вузол	+	+
7.	Простий вузол	+	+
8.	Образ		+
9.	Вхідний образ		+
10.	Навчання		+

Аналіз отриманих результатів виявив, що відсутність програмно визначених термінів у множині автора не завжди характеризує недолік розглядуваної 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 в рамках матеріалу, що викладається.

Висновки

Розглянута інформаційна технологія дозволяє з достатньою ефективністю автоматизовано формувати множини ключових семантичних термінів навчальних матеріалів. Розроблене відповідно до запропонованої інформаційної технології програмне забезпечення в результаті обробки вхідних даних у вигляді цифрового документу навчального матеріалу формату .docx дозволяє одержувати вихідні дані у вигляді множини ключових термінів відповідного навчального матеріалу.

Проведені за допомогою розробленого авторами тестового програмного забезпечення дослідження підтвердили можливість ефективно автоматизовано формувати множини ключових семантичних термінів навчальних матеріалів з показниками точності пошуку до 92,9 % та повноти пошуку до 100,0 %.

Подальші дослідження спрямовані на аналіз впливу на показники ефективності запропонованої інформаційної технології взаємозв'язків між кількістю ключових семантичних термінів в результуючій множині та значеннями коефіцієнту щільності ключових слів та вдосконалення розглянутої інформаційної технології для покращення результатів.

Література

1. Снитюк В. Е., Юрченко К. Н. Интеллектуальное управление оцениванием знаний. Черкассы, 2013. 262 с.
2. Мазурець О. В. Онтологічний підхід до побудови семантичної моделі навчальних матеріалів. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький. 2017. № 6. С. 223–229.
3. 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.
4. 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.
5. 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.
6. 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.
7. Ключові слова. iGroup Україна. [Електронний ресурс]. Режим доступу: http://igroup.com.ua/seo-articles/keywords/
8. Create .NET Apps With NuGet. Spire.Doc for .NET [Електронний ресурс]. Режим доступу: https://www.nuget.org/packages/Spire.Doc/
9. Manning C., Raghavan P., Schutze H. Introduction to Information Retrieval. Cambridge University Press, 2008. 482p.

References

1. Snituk V. E. & Yurchenko K. N. (2013) Intelektualnoe Upravlenie Ocenivaniem Znaniy. Cherkassy.
2. Mazurets O. V. (2017) Ontological Approach to Building a Semantic Model of Educational Materials. Herald of Khmelnytskyi national university. Technical Sciences, Issue 6, 2017 (255). P. 223–229.
3. 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.
4. Ortuño M., Carpena P., Bernaola P., Muñoz E. & Somoza A.M. (2002) Keyword detection in natural languages and DNA // Europhys. Lett, 2002. 57(5). P. 759–764.
5. Barmak O.V. & Mazurets O.V. (2015) Methods of Automation of Definition of Semantic Terms in Educational Materials // Herald of Khmelnytskyi national university. Technical Sciences, Issue 2, 2015 (223). P. 209–213.
6. Lande D.V. & Snarskiy A.A. (2013) Kompaktificirovanniy Gorizontalniy Graf Vidimosti dlya Seti Slov. Trudi Mejdunarodnoy Nauchnoy Konferencii «Intellektualniy Analiz Informacii IAI-2013. Znania I Rassujdenia». P. 158–164.
7. IGROUP UKRAINE (2018) Keywords. [Online] Available from: http://igroup.com.ua/seo-articles/keywords/ [Accessed: 12 February 2018]
8. CREATE .NET APPS WITH NUGET (2018) Spire.Doc for .NET [Online] Available from: https://www.nuget.org/packages/Spire.Doc/ [Accessed: 12 February 2018].
9. Manning, C., Raghavan, P., Schutze, H. (2008) Introduction to Information Retrieval. Cambridge University Press.

Про авторів:

[1]*Крак Юрій Васильович*,
доктор фізико-математичних наук, професор,
завідувач кафедри теоретичної кібернетики Київського національного університету імені Тараса Шевченка,
старший науковий співробітник Інституту кібернетики імені В.М. Глушкова НАН України.
Кількість друкованих праць – понад 500, в тому числі:
кількість наукових публікацій в українських фахових виданнях – 170,
кількість наукових публікацій в зарубіжних виданнях – 60.
H-індекс – 2.
http://orcid.org/0000-0002-8043-0785,

[2]*Бармак Олександр Володимирович,*
доктор технічних наук, професор,
професор кафедри Комп'ютерних наук та інформаційних технологій
Хмельницького національного університету.
Кількість друкованих праць – понад 200, в тому числі:
кількість наукових публікацій в українських фахових виданнях – 70,
кількість наукових публікацій в зарубіжних виданнях – 15.
H-індекс – 1.
http://orcid.org/0000-0003-0739-9678,

[2]*Мазурець Олександр Вікторович,*
старший викладач кафедри Комп'ютерних наук та інформаційних технологій
Хмельницького національного університету.
Кількість наукових публікацій в українських виданнях – 93.
Кількість наукових публікацій в зарубіжних виданнях – 1.
http://orcid.org/0000-0002-8900-0650,

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