УДК 004.9

О.В. МАЗУРЕЦЬ
Хмельницький національний університет

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ СЕМАНТИЧНИХ ТЕРМІНІВ В ЕЛЕМЕНТАХ НАВЧАЛЬНИХ МАТЕРІАЛІВ

У статті досліджено проблему автоматизації побудови семантичної моделі навчального курсу у вигляді онтології. Встановлено, що ключовим аспектом побудови онтології навчального курсу є визначення множини ключових семантичних термінів у контенті елементів навчальних матеріалів. Запропоновано інформаційну технологію автоматизованого визначення множини ключових семантичних термінів у контенті елементів 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 електронного документу для вибору елементу для аналізу, після чого проводиться сегментація за фразами і термінами, терміни лематизуються та їх множина компактифікується. На основі автоматично лематизованого фрагменту 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 електронний документ навчального матеріалу та обраний елемент для аналізу, вихідними даними є відповідна множина ключових семантичних термінів навчального матеріалу. 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 елементів навчальних матеріалів з середніми показниками точності пошуку до 73,2% та повноти пошуку до 69,7%. 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. Встановлена ефективність запропонованої інформаційної технології сприяє її використанню для вирішення ряду актуальних задач, таких як семантична допомога при створенні тестів, автоматизація формування рефератів та анотацій до елементів навчальних матеріалів, оцінка відповідності навчальних матеріалів змістовим вимогам, оцінка відповідності наборів тестових завдань навчальним матеріалам тощо.

Ключові слова: онтологія, ключові терміни, ключові слова, електронний документ, навчальні матеріали, дисперсійна оцінка.

O. MAZURETS
Khmelnytsky National University

INFORMATION TECHNOLOGY FOR AUTOMATED DEFINITION OF SEMANTIC TERMS IN THE CONTENT OF THE ELEMENTS OF EDUCATIONAL MATERIALS

The problem of automation of the construction of semantic model of educational courses in the form of an ontology was investigated in the article. It has been established, that the key aspect of constructing of the ontology of educational courses is to make the set of key semantic terms of the contents of educational material elements. The information technology of automated determination of key semantic terms in the content of educational materials elements is considered, which is based on the search of used phrases 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 electronic 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 fragment of 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 electronic 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 program 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, average evaluated search precision metrics 73.2% and search recall 69.7%. The analysis of the results showed that the lack of programmed terms in the author's set does not always characterize the lack of considered technology. Some semantically important terms are subjectively ignored by the authors, while another category is made up of concepts in which the authors emphasize excessive attention to their secondary character within the framework of the material being taught. The established effectiveness of the proposed technology allows use it to solution a number of urgent tasks, such as semantic assistance in creating tests, automation of the creation of abstracts and annotations to the elements of educational materials, determination the conformity of educational materials to content requirements, determination the conformity of sets of test tasks to educational materials, etc.

Keywords: ontology, key terms, keywords, electronic document, educational materials, disperse evaluation.

Постановка проблеми в загальному вигляді

На сучасному етапі засобом реалізації освіти, зокрема дистанційної, є інформаційні технології. Це визначає необхідність формалізації та стандартизації навчального процесу [1]. Загальноприйнятим є підхід застосування навчальних матеріалів у вигляді електронних документів визначеної структури як інструменту навчання. Для роботи з курсами навчальних дисциплін використовуються спеціалізовані віртуальні навчаючі середовища, найбільш відомим із яких є Moodle [2]. При їх використанні, потенційна якість отриманих освітніх послуг безпосередньо визначається якістю навчальних матеріалів курсу. В умовах вузької спеціалізації курсів дисциплін, їх численності та швидкого оновлення, перспективним шляхом

оцінки якості навчальних курсів та їх елементів є автоматизація вирішення відповідного ряду задач у сучасній вищій освіти. До таких задач належать: автоматизація побудови семантичної моделі навчальних курсів, оцінка відповідності навчальних матеріалів вимогам, допомога та контроль якості при формуванні навчальних матеріалів, оцінка відповідності наборів тестових завдань навчальним матеріалам, допомога та контроль якості при формуванні тестів до навчальних матеріалів, автоматизована генерація прототипів тестових завдань, реалізація гнучких алгоритмів тестування, автоматизація формування анотацій і рефератів до елементів навчальних матеріалів тощо. Вирішення всіх цих задач може бути реалізоване через автоматизацію побудови семантичної моделі навчальних курсів та її використання у відповідних інформаційних технологіях.

Зі змістовної точки зору, базовою властивістю контенту є його семантика, яку формалізовано відображають у вигляді мережі, вузлами якої є терміни, що несуть семантичне навантаження, а дуги відображають характер зв’язку між вузлами [3]. Зв'язок між термінами навчальних матеріалів залежить від багатьох факторів (тип лекції, галузь знань, літературні здібності автора, тощо) й може змінюватися у широких межах без втрати якості викладання, тому актуальність його аналізу низька. Відповідно, аналіз саме термінів, що використовуються у навчальних матеріалах, дозволяє сформувати семантичну модель навчального курсу й вирішити наведений ряд задач.

Аналіз останніх досліджень

Задачу автоматизованої аналітичної обробки текстової інформації намагаються вирішити багато вітчизняних та іноземних авторів, серед яких можна виділити роботи Д.В. Ланде, В.Е. Снитюка, В.І. Горькової, Є.А. Борохова, Х.П. Луна, В.Є. Берзона, І.П. Севбо, В.П. Леонова, С.І. Гінді та інших. Дослідження й розробки в напрямку автоматизації обробки текстів у Європі й США привертають увагу відомих приватних фірм і державних організацій найвищого рівня. Європейський Союз вже декілька років координує ряд програм у галузі автоматичної обробки тексту, наприклад Human Language Technology Sector of the Information Society Technologies (IST) Programme [4]. Основні розробки присвячено автоматизації процесу синтаксичного аналізу.

Більшість існуючих програмних продуктів в розглядуваній області призначені переважно для SEO аналізу текстів. Наприклад, аналізатор від біржі контенту “Адвего” [5] вираховує кількість слів, кількість граматичних помилок і т.п. Сервіс також дозволяє побачити перелік ключових слів, які вираховуються за методом частотної оцінки. Текстовий аналізатор від компанії “Seozor” [6] допомагає визначити вагу слів в тексті для складання анкор-листа.

Для автоматизації пошуку ключових слів використовуються різноманітні методи аналізу текстів, таких як частотна оцінка, оцінка TFIDF та дисперсійна оцінка [7]. 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 [8]. Попередніми дослідженнями було визначено найбільш ефективним методом аналізу текстів метод дисперсійної оцінки, проте встановлено й ряд факторів, які ускладнюють його монопольне застосування для вирішення задачі автоматизованого визначення семантичних термінів в навчальних матеріалах [9]. Зокрема, малий обсяг контенту та вузька семантична направленість елементів аналізу зменшує ефективність наведених методів аналізу текстів. Тому є доцільною розробка нової інформаційної технології, яка із використанням методу дисперсійної оцінки дозволить ефективно й автоматизовано визначати семантичні терміни в навчальних матеріалах.

Постановка задачі

Метою роботи є розробка інформаційної технології автоматизованого визначення множини ключових семантичних термінів у електронних документах навчальних матеріалів й визначення її ефективності.

Викладення основних матеріалів дослідження

Семантика начального матеріалу виражається його логічною структурою (наприклад: Дисципліна / Розділ / Тема) та поняттями, що розглядаються в ньому. Очевидно, що ієрархія змістовних блоків визначає верхній рівень вертикальної онтології відповідної навчальної дисципліни [3]. Якщо розглядати ієрархію змістовних блоків начального матеріалу як рівні вертикальної онтології відповідної навчальної дисципліни, то, ключові терміни є найнижчим рівнем онтології начального матеріалу навчальної дисципліни (рис. 1). Така формалізація досліджуваного процесу визначає структуру моделі онтології начального матеріалу.

Множини ключових термінів кожного елементу ієрархії змістовних блоків навчального матеріалу можуть мати довільну кількість елементів й у сукупності формують загальну множину ключових термінів навчального матеріалу. За такої моделі, онтологія навчального матеріалу може бути методом виявлення сенсу навчального матеріалу. Пошук множин ключових семантичних термінів у навчальних матеріалах необхідний для всіх елементів ієрархії змістовних блоків, тому інформаційна технологія має використовуватись не тільки для електронного документу загалом, а я для його елементів.

Загальну схему інформаційної технології автоматизованого множини ключових семантичних термінів у електронних документах навчальних матеріалів відображено на рис. 2.

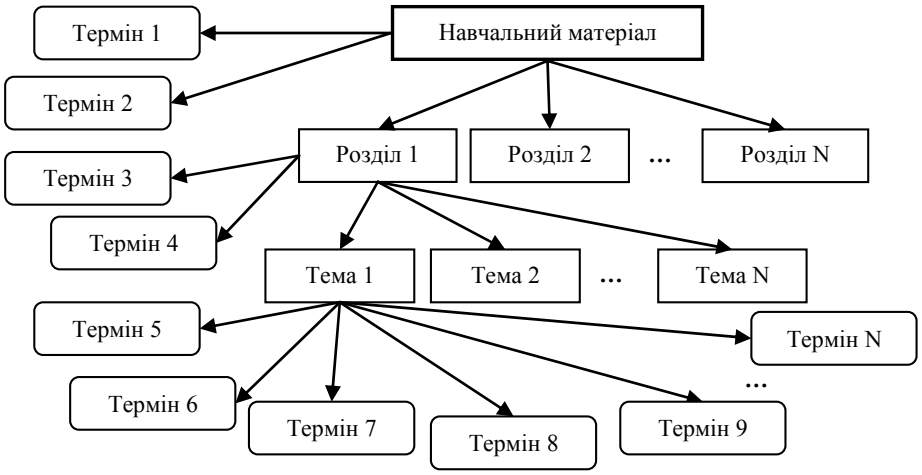


Рис. 1. Вихідна модель онтології навчального матеріалу

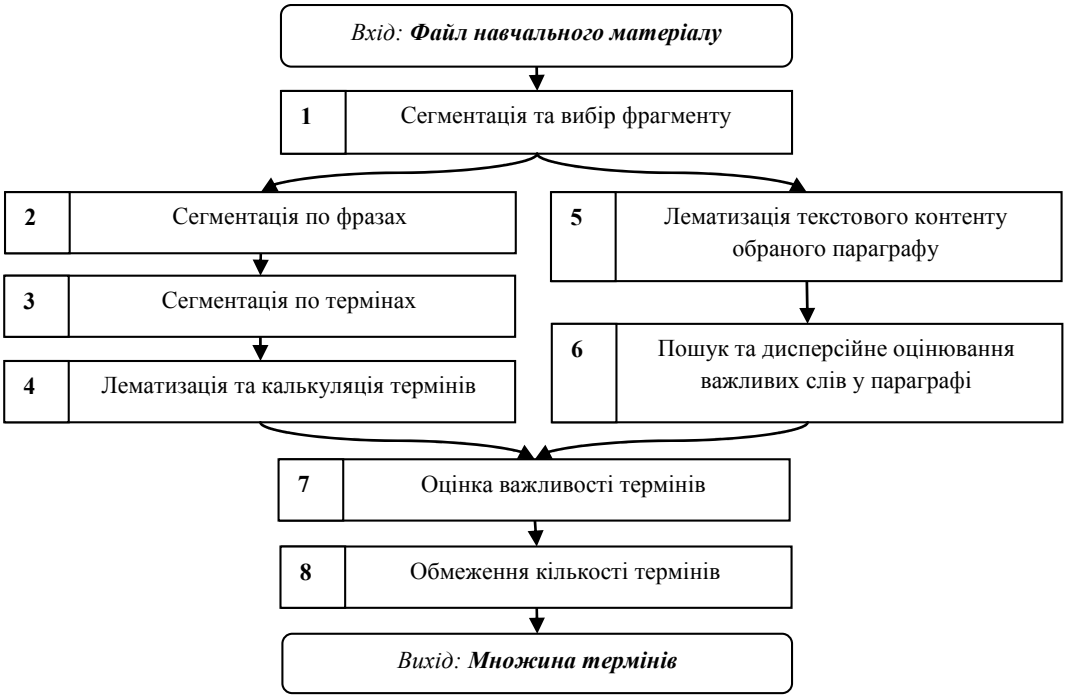


Рис. 2. Схема інформаційної технології автоматизованого множини ключових семантичних термінів

Сегментація та вибір фрагменту для аналізу (Блок 1) полягає в аналізі структури електронного документу. Зважаючи на існуючі загальноприйняті вимоги до структури навчальних матеріалів навчальних дисциплін, можна зробити висновок про відповідність системи заголовків навчальних матеріалів як електронних документів верхнім рівням семантичної структури навчального матеріалу. Наприклад, назви дисциплін відповідатимуть елементам стандартного стилю «Heading 1», назви розділів – «Heading 2», назви тем – «Heading 3» тощо.

Блок 2 (*Сегментація за фразами*) призначений для розбиття фрагменту контенту електронного документу, що обробляється, на менші фрагменти – фрази. Під фразою розуміється семантично цілісний вузол, що виокремлений стилістичним форматуванням тексту чи розділовими знаками, й локалізує місцезнаходження окремих термінів. Відповідно до об'єктної моделі документу MS Office, TextRange є найнижчим рівнем структури документу, що визначає фрагмент тексту однакового стилю в межах Paragraph [10]. Відтак до множини фраз включаються неперервні впорядковані послідовності слів, що не виходять за межі контейнерів електронного документу TextRange та не перериваються розділовими знаками. Одержання в результаті виконання блоку множини фраз дає можливість в подальшому обробляти на предмет пошуку термінів кожну з фраз окремо.

Блок 3 (*Сегментація за термінами*) забезпечує формування множини всіх можливих термінів, що присутні у досліджуваному контенті. До множини термінів навчального матеріалу M_T включаються всі можливі неперервні впорядковані послідовності слів, які не виходять за межі фраз та відповідають умові:

$$M_T = \begin{Bmatrix} \langle x_1 \rangle | x_1 \in M_I \\ \langle x_1, x_2 \rangle | x_1 \in M_I \cup M_П, x_2 \in M_I \cup M_П, \langle x_1, x_2 \rangle \cup M_I \neq \varnothing \\ \langle x_1, x_2, x_3 \rangle | x_1 \in M_I \cup M_П, x_3 \in M_I \cup M_П, \langle x_1, x_2, x_3 \rangle \cup M_I \neq \varnothing \\ \langle x_1, x_2, x_3, x_4 \rangle | x_1 \in M_I \cup M_П, x_4 \in M_I \cup M_П, \langle x_1, x_2, x_3, x_4 \rangle \cup M_I \neq \varnothing \\ \langle x_1, x_2, x_3, x_4, x_5 \rangle | x_1 \in M_I \cup M_П, x_5 \in M_I \cup M_П, \langle x_1, x_2, x_3, x_4, x_5 \rangle \cup M_I \neq \varnothing \\ \langle x_1, x_2, x_3, x_4, x_5, x_6 \rangle | x_1 \in M_I \cup M_П, x_6 \in M_I \cup M_П, \langle x_1, x_2, x_3, x_4, x_5, x_6 \rangle \cup M_I \neq \varnothing \end{Bmatrix} \quad (1)$$

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 електронного документу навчального матеріалу.

Блок 4 (*Лематизація та калькуляція термінів*) дозволяє на основі множини термінів M_T сформувати множину лемо-незалежних термінів M_{T1}, а також співставити кожному із них кількість появ 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 спіставляється 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. Одержана в результаті множина лемо-незалежних термінів M_{T1} містить терміни, що використовуються у навчальному матеріалі з кількісним показником використання, але не визначає важливість даних термінів.

Блок 5 (*Лематизація текстового контенту обраного параграфу*) переводить текст визначеного фрагменту контенту електронного 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, 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 [9].

Вихідними даними блоку є впорядкована множина слів, кожному з яких співставлена оцінка його дисперсії, що розглядається як оцінка важливості даного слова у досліджуваному фрагменті електронного документу.

Блок 7 (*Оцінка важливості термінів*) вхідними даними має множину лемо-незалежних термінів M_{T1} із співставленою кожному з них кількістю появ 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:

$$v_n = \sum_{i=1}^{x_n} \frac{K_n \sigma_n}{k_n}, \quad (2)$$

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 електронного 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_{TI} із співставленими кожному з них кількістю появ у тексті та значенням оцінки важливості, впорядкована за спаданням номінального значення оцінки важливості терміна.

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_{TI} з найбільшими значеннями оцінки важливості, а їх

кількість випливає із визначення відомого показника з семантичної обробки текстів, щільності ключових слів [11]. Щільність ключових слів P_{txt} 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:

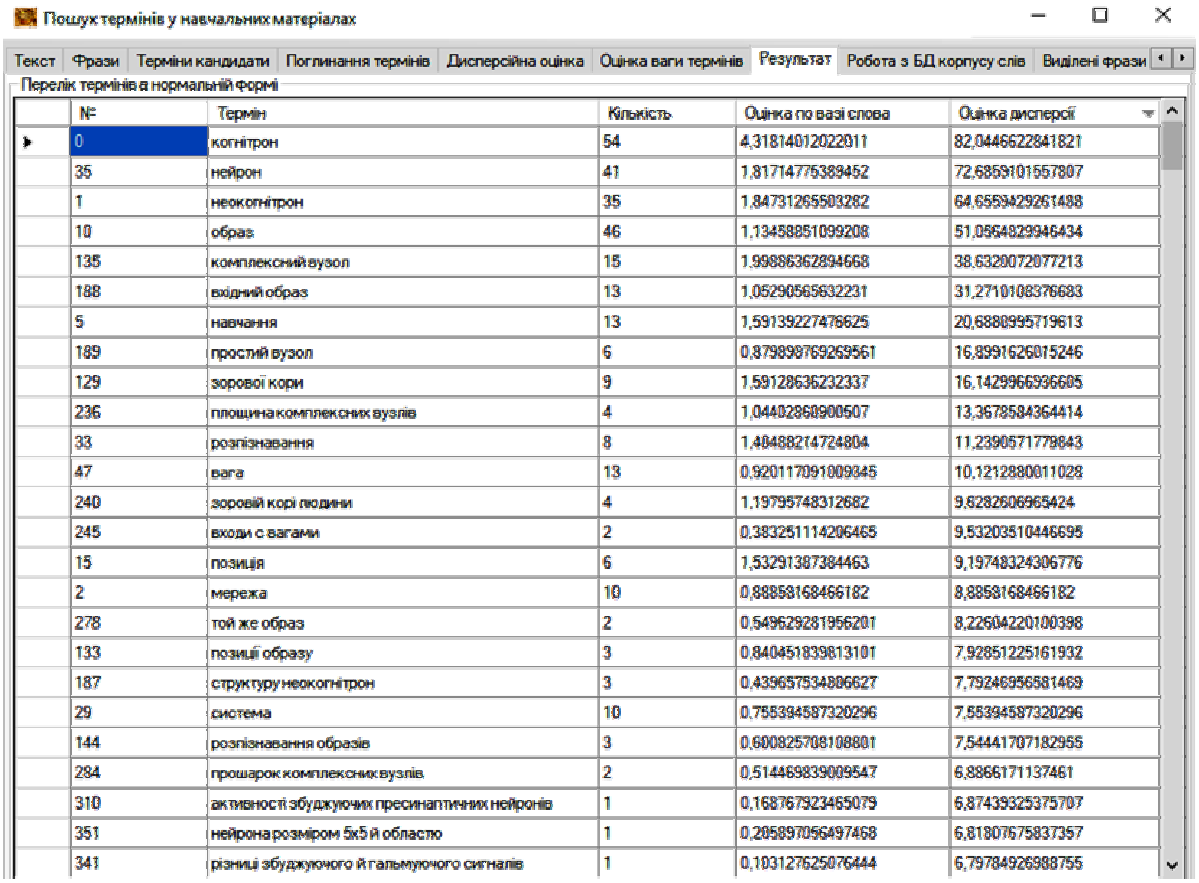
$$\sum_{i=1}^{n} \frac{K_n x_n}{X_{txt}} \le P_{txt}, \tag{3}$$

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}.

Отже, розглянута інформаційна технологія автоматизованого визначення семантичних термінів в елементах навчальних матеріалів дозволяє на основі електронного документу навчального матеріалу автоматизовано отримувати відповідний перелік ключових термінів визначеного елементу навчального матеріалу.

Практичне дослідження ефективності

Запропонована інформаційна технологія автоматизованого визначення семантичних термінів в елементах навчальних матеріалів була реалізована в тестовому програмному продукті. Вхідними даними для системи є електронний документ навчального матеріалу, а вихідними даними є множина ключових термінів, відповідна досліджуваному фрагменту контенту електронного документу навчального матеріалу. Створений програмний продукт на основі введених даних у вигляді файлу навчального матеріалу автоматично моделює структуру електронного документу для вибору елементу для аналізу, після чого проводиться сегментація по фразах та термінах, терміни лематизуються й їх множина компактифікується, паралельно 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 вище розглянутої математичної моделі. Зокрема, на рисунку 3 показано приклад обробки теми «Нейромережі когнітрон та неокогнітрон» навчального матеріалу дисципліни «Методи та системи штучного інтелекту».



№	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	38,6320072977213
188	вхідний образ	13	1,05290565632231	31,2710108376683
5	навчання	13	1,59139227476625	20,6880995719613
189	простий вузол	6	0,879899769269561	16,8991626815246
129	зорової кори	9	1,59128636232337	16,1429966936605
236	площина комплексних вузлів	4	1,04402860900507	13,3678584364414
33	розпізнавання	8	1,48488214724804	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	0,8858168466182	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	0,690825708108801	7,54441707182955
284	прошарок комплексних вузлів	2	0,514469839009547	6,8866171137461
310	активності збуджуючих пресинаптичних нейронів	1	0,168767923465079	6,87439325375707
351	нейрона розміром 5х5 й областю	1	0,205897056497468	6,81807675837357
341	різниці збуджуючого й гальмуючого сигналів	1	0,103127625076444	6,79784926988755

Рис. 3. Одержання множини ключових термінів розробленим програмним продуктом

Структуру програмного продукту у вигляді діаграми класів відображено на рисунку 4. Відповідно до діаграми, пошук ключових термінів та робота з ключовими термінами реалізовані у класі WordCombination. Клас IMainForm забезпечує інтерфейс для головної форми взаємодії з користувачем, а MainForm є класом користувацької форми. Для збереження і роботи з комбінаціями слів (словосполученням) використовується клас Combination. Клас WigthCombination наслідує клас Combination і

розширює його можливостями обрахунку ваги словосполучення. IWorkWithServer реалізує інтерфейс для роботи з базою даних, а WorkWithServer забезпечує роботи з базою даних Microsoft SQL Server, який використовується як для збереження даних роботи, так і для використання бази даних корпусу слів української мови. Клас PresenterWork використовується для взаємодії графічної частина й логіки програми. WigthWord є класом для обрахунку ваги слова в контексті досліджуваного фрагменту тексту. Для зберігання множини слів і всіх пов'язаних із ним даних використовується клас Word. Клас SelectTerm зберігає терміни, виділені в тексті і тип виділення для подальшого аналізу важливості термінів. Section – клас, який приймає текст в межах певного параграфа й організовує подальшу обробку даного фрагменту. Для первинного аналізу тексту і розбиття його на параграфи (контент, прив'язаний до одного елементу заголовку Heading, використовується клас ProcessText).

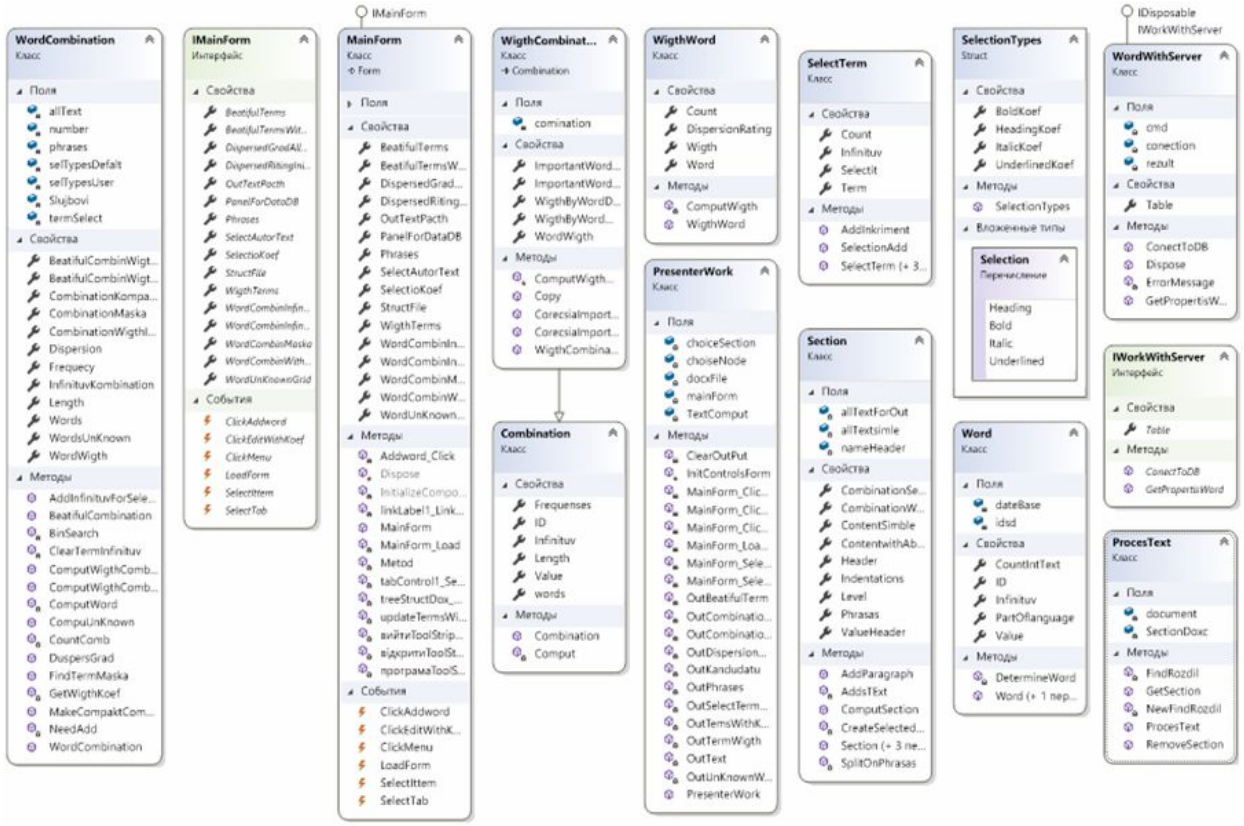


Рис. 4. Діаграма класів програмного продукту для автоматизованого визначення ключових термінів

Ефективність практичного застосування розглянутої інформаційної технології автоматизованого визначення семантичних термінів в елементах навчальних матеріалів може бути визначена шляхом оцінки результатів використання відповідного програмного продукту за показниками точності та повноти [12].

Точність пошуку P (Precision, відношення кількості релевантних ключових термінів, знайдених автоматично, до загальної кількості знайдених ключових термінів в досліджуваному тексті) та повнота пошуку R (Recall, відношення кількості релевантних ключових термінів, знайдених автоматично, до загальної кількості релевантних ключових термінів в досліджуваному тексті) обчислюються за наступними формулами:

$$P = \frac{\left|M_{TK}^{E} \cap M_{TK}\right|}{\left|M_{TK}\right|},\ R = \frac{\left|M_{TK}^{E} \cap M_{TK}\right|}{\left|M_{TK}^{E}\right|}, \tag{4}$$

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

Середня точність пошуку $\overline{P}$ та середня повнота пошуку $\overline{R}$ визначаються наступним чином:

$$\overline{P} = \frac{\sum_{i=1}^{k} P_k}{k},\ \overline{R} = \frac{\sum_{i=1}^{k} R_k}{k}, \tag{5}$$

де k – кількість навчальних матеріалів у тестовій вибірці.

Для визначення ефективності практичного застосування інформаційної технології автоматизованого визначення семантичних термінів в елементах навчальних матеріалів, тестовим програмним продуктом було оброблено тестову вибірку з 50 файлів навчальних курсів. Наприклад, у результаті тестування розглянутого

на прикладі рис. 3 навчального матеріалу за показника щільності ключових слів 7% було отримано наступне:

- до множини ключових термінів автоматично було віднесено наступний перелік термінів: когнітрон, неокогнітрон, нейрон, комплексний вузол, простий вузол, образ, вхідний образ, навчання;
- до множини ключових термінів експертом було віднесено наступний перелік термінів: когнітрон, неокогнітрон, нейрон, збуджуючий нейрон, гальмуючий нейрон, комплексний вузол, простий вузол.

Відповідно до математичних моделей (4), даному випадку точність пошуку склала 0,625, а повнота пошуку склала 0,714. Відповідно до (5), середня точність пошуку для дослідженої вибірки з 50 файлів навчальних курсів склала 0,732, а середня 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.

Висновки

У статті було досліджено проблему автоматизації побудови семантичної моделі навчального курсу у вигляді онтології. Встановлено, що ключовим аспектом побудови онтології навчального курсу є визначення множини ключових семантичних термінів у контенті елементів навчальних матеріалів.

Запропоновано інформаційну технологію автоматизованого визначення множини ключових семантичних термінів у контенті елементів 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 електронного документу для вибору елементу для аналізу, після чого проводиться сегментація по фразах і термінах, терміни лематизуються та їх множина компактифікується. На основі автоматично лематизованого фрагменту 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 електронний документ навчального матеріалу та обраний елемент для аналізу, вихідними даними є відповідна множина ключових семантичних термінів навчального матеріалу.

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 елементів навчальних матеріалів з середніми показниками точності пошуку до 73,2% та повноти пошуку до 69,7%. 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.

Встановлена ефективність запропонованої інформаційної технології сприяє її використанню для розв'язання 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.

Література

1. Нові інформаційні технології в освіті [Електронний ресурс]. – Режим доступу : http://it-tehnolog.com/statti/novi-informatsiyni-tehnologiyi-navchannya/.

2. Moodle – Open-source learning platform [Електронний ресурс]. – Режим доступу : https://moodle.org/

3. Мазурець О. В. Онтологічний підхід до побудови семантичної моделі навчальних матеріалів / О.В. Мазурець // Вісник Хмельницького національного університету. Серія: Технічні науки. – 2017. – № 6. – С. 223–229.

4. Human Language Technology Sector of the Information Society Technologies (IST) Programme [Електронний ресурс]. – Режим доступу : http://www.linglink.lu.

5. SEO-аналізатор «Адвего» [Електронний ресурс]. – Режим доступу : http://advego.ru/text/seo/.

6. Семантичний онлайн-аналізатор тексту «Seozor» [Електронний ресурс]. – Режим доступу : http://seozor.ru/tools/analyzer.php.

7. Ventura J. New Techniques for Relevant Word Ranking and Extraction / J. Ventura, J. Silva // Proceedings of the aritficial intelligence 13th Portuguese conference on Progress in artificial intelligence, EPIA'07. – Berlin : Springer-Verlag, Berlin, Heidelberg, 2007. – P. 691–702.

8. Ландэ Д. В. Компактифицированный горизонтальный граф видимости для сети слов / Д. В. Ландэ, А. А. Снарский // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения». – Киев : КПИ, 2013. – С. 158–164.

9. Бармак О. В. Методи автоматизації визначення семантичних термінів у навчальних матеріалах / О.В. Бармак, О. В. Мазурець // Вісник Хмельницького національного університету. Сер.: Технічні науки. – 2015. – № 2(223). – С. 209–213.

10. Мазурець О. В. Використання спеціалізованих програмних розширень для автоматизації роботи з цифровими документами навчальних матеріалів / О. В. Мазурець, О. В. Ковальчук, В. О. Слободзян // Вісник Хмельницького національного університету. Серія: Технічні науки. – 2018. – № 1. – С. 61–69.

11. Ключові слова. iGroup Україна [Електронний ресурс]. – Режим доступу : http://igroup.com.ua/seo-articles/keywords/

12. Manning C. Introduction to Information Retrieval / C. Manning, P. Raghavan, H. Schutze. – Cambridge University Press, 2008. – 482 p.

References

1. IT-TEHNOLOG (2018) New Information Technologies in Education. [Online] Available from: http://it-tehnolog.com/statti/novi-informatsiyni-tehnologiyi-navchannya/ [Accessed: 15 February 2018]

2. MOODLE (2018) Moodle – Open-source learning platform. [Online] Available from: https://moodle.org/ [Accessed: 15 February 2018]

3. 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.

4. LINGLINK.LU (2018) Human Language Technology Sector of the Information Society Technologies (IST) Programme. [Online] Available from: http://www.linglink.lu [Accessed: 15 February 2018]

5. ADVEGO (2018) SEO-analazer "Advego". [Online] Available from: http://advego.ru/text/seo/ [Accessed: 15 February 2018]

6. SEOZOR (2018) Semantic online-analazer of texts "Seozor". [Online] Available from: http://seozor.ru/tools/analyzer.php [Accessed: 15 February 2018]

7. 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.

8. LANDE, D. V. & SNARSKIY, A. A. (2013) Kompaktificirovanniy Gorizontalniy Graf Vidimosti dlya Seti Slov / D.V. Lande, A. A. Snarskiy // Trudi Mejdunarodnoy Nauchnoy Konferencii «Intellektualniy Analiz Informacii IAI-2013. Znania I Rassujdenia». p 158-164.

9. 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.

10. MAZURETS, O. V., KOVALCHYK, O. V. & SLOBODZIAN, V. O. (2018) Using specialized software packeges for automation of work with digital documents of educational materials // Herald of Khmelnytskyi national university. Technical Sciences, Issue 1, 2018 (257). p. 61-69.

11. IGROUP UKRAINE (2018) Keywords. [Online] Available from: http://igroup.com.ua/seo-articles/keywords/ [Accessed: 15 February 2018]

12. MANNING, C., RAGHAVAN, P. & SCHUTZE, H. (2008) Introduction to Information Retrieval. Cambridge University Press.

Рецензія/Peer review : 21.03.2018 р. Надрукована/Printed :18.05.2018 р.
Стаття рецензована редакційною колегією