УДК 004.9

ДОСЛІДЖЕННЯ ПРАКТИЧНОЇ ЕФЕКТИВНОСТІ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ СЕМАНТИЧНИХ ТЕРМІНІВ НАВЧАЛЬНИХ МАТЕРІАЛІВ

Ковальчук О.В., Мазурець О.В.
Україна, Хмельницький національний університет
E-mail: losha.kovalchyk1998@gmail.com

Базовим засобом реалізації дистанційної освіти є інформаційні технології. Це визначає необхідність формалізації та стандартизації навчального процесу. Загальноприйнятим є підхід застосування навчальних матеріалів у вигляді цифрових документів визначеної структури як інструменту навчання. Для роботи із курсами навчальних дисциплін використовуються спеціалізовані віртуальні навчаючі середовища. При їх використанні, потенційна якість отриманих освітніх послуг безпосередньо визначається якістю навчальних матеріалів курсу. В умовах вузької спеціалізації курсів дисциплін, їх значної численності та інтенсивного оновлення, перспективним шляхом оцінки якості навчальних курсів та їх елементів є автоматизація вирішення відповідного ряду задач у сучасній вищій освіті.

З семантичної 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 [1]. Відтак, аналіз термінів, що використовуються у навчальних матеріалах, дозволяє побудувати семантичну модель навчального курсу й вирішити ряд похідних задач.

Метою роботи є висвітлення загальних аспектів інформаційної технології автоматизованого визначення множин ключових семантичних термінів у електронних документах навчальних матеріалів й дослідження її ефективності.

Семантика навчального матеріалу виражається його логічною структурою (наприклад: Дисципліна / Розділ / Тема) та поняттями, що розглядаються в ньому. 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 визначення множини ключових семантичних термінів у електронних документах навчальних матеріалів відображено на рисунку 1.

Рис. 1 – Загальна схема інформаційної технології автоматизованого множини ключових семантичних термінів

В результаті аналізу системи заголовків навчальних матеріалів як електронних документів забезпечується сегментація контенту навчальних матеріалів та вибір фрагменту для аналізу. Після цього проводиться розбиття фрагменту контенту електронного документу, на менші фрагменти – фрази. Під фразою розуміється семантично цілісний вузол, що виокремлений форматуванням тексту чи розділовими знаками, й локалізує місцезнаходження окремих термінів. Одержання в результаті виконання блоку множини фраз дає можливість в подальшому обробляти на предмет пошуку термінів кожну з фраз окремо. Наступним кроком проводиться 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:

$$M_T = \left\{ \left\langle x_1, x_2, x_3, x_4, x_5, x_6 \right\rangle \middle| \begin{array}{l} x_1 \in M_I \bigcup M_{\Pi} \\ x_2, x_3, x_4, x_5, x_6 \in M_M \\ \left\langle x_1, x_2, x_3, x_4, x_5, x_6 \right\rangle \bigcup M_I \neq \varnothing \end{array} \right\} \quad (1)$$

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 $M_{ПЙ}$), $M_M = M_I \bigcup M_{ПК} \bigcup M_C \bigcup M_Ч \bigcup M_{ПЙ} \bigcup \varnothing$. 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.

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. Після чого одержана множина обробляється й компактифікується таким чином, що всі ідентичні повторення термінів видаляються, а кожному терміну співставляється 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. 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.

Лематизація текстового контенту переводить текст електронного документу навчального матеріалу, що аналізується, до відповідної послідовності слів у інфінітивному стані. Вони дозволяють проводити подальше оцінювання дисперсії 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 [2]. Пошук та дисперсійне оцінювання важливих слів у параграфі призначені для оцінки важливості кожного

слова в досліджуваному тексті, що проводиться з використанням методу дисперсійного оцінювання [3].

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

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

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.

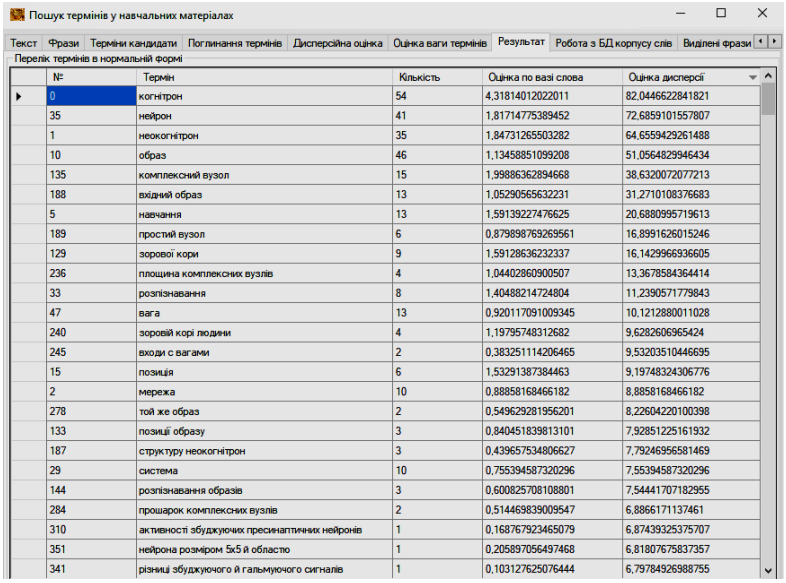
Множина ключових термінів формується на основі лемо-незалежних термінів із множини M_{T1} з найбільшими значеннями оцінки важливості, а їх кількість випливає із визначення відомого показника з семантичної обробки текстів, щільності ключових слів P_{txt}.

Розглянута інформаційна технологія автоматизованого визначення семантичних термінів в елементах навчальних матеріалів дозволяє на основі електронного документу навчального матеріалу автоматизовано отримувати відповідний перелік ключових термінів навчального матеріалу.

Запропонована інформаційна технологія автоматизованого визначення семантичних термінів в елементах навчальних матеріалів була реалізована в дослідницькому програмному продукті. Структурно програмний продукт складається з ряду класів. Так, 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.

Вхідними даними для системи є електронний документ навчального матеріалу, а вихідними даними є відповідна множина ключових термінів. Зокрема, на рисунку 2 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	38,6320072077213
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. 2 – Приклад роботи розробленого програмного продукту

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 [4].

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:

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

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. Так, у результаті тестування розглянутого на прикладі рисунку 2 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.

Таким чином, було 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. 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 [1].

Література

1. Мазурець О. В. Онтологічний підхід до побудови семантичної моделі навчальних матеріалів / О. В. Мазурець // Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2017, №6. – С.223-229.

2. Бармак О. В., Мазурець О. В. Інформаційна технологія автоматизованого визначення термінів у навчальних матеріалах / О. В. Бармак, О. В. Мазурець // Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах» – Хмельницький, 2015. – №2. – С.94–102.

3. Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. Keyword detection in natural languages and DNA / M. Ortuño, P. Carpena, P. Bernaola, E. Muñoz, A. M. Somoza // Europhys. Lett, 2002. – 57(5). – P. 759-764.

4. Powers D. M. W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation / D. M. W. Powers // Journal of Machine Learning Technologies, Vol. 2, No. 1. (2011), p. 37-63.