ДОСЛІДЖЕННЯ ТОЧНОСТІ ТА ПОВНОТИ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ СЕМАНТИЧНИХ ТЕРМІНІВ У НАВЧАЛЬНИХ МАТЕРІАЛАХ

Бармак Олександр Володимирович

доктор технічних наук, професор Кафедри комп'ютерних наук та інформаційних технологій Хмельницького національного університету, alexander.barmak@gmail.com

Мазурець Олександр Вікторович

старший викладач Кафедри комп'ютерних наук та інформаційних технологій Хмельницького національного університету, exe.chong@gmail.com

Актуальність теми На сучасному етапі засобом реалізації освіти, зокрема дистанційної, є інформаційні технології, що визначає необхідність формалізації та стандартизації навчального процесу. 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, при використанні яких потенційна якість отриманих освітніх послуг безпосередньо визначається якістю навчальних матеріалів курсу.

Для вирішення ряду проблем в області автоматизації роботи з інформаційними та тестовими електронними навчальними матеріалами запропоновано використання інформаційної технології автоматизованого визначення множини ключових семантичних термінів у контенті елементів навчальних матеріалів [1, 2], що базується на пошуку використаних фраз у тексті та дисперсійній оцінці важливості слів. Для оцінки ефективності визначення множин ключових термінів запропоновано обрахунок показників точності та повноти пошуку [3].

Виклад основного матеріалу Загальну схему інформаційної технології автоматизованого визначення множини ключових семантичних термінів у електронних документах навчальних матеріалів наведено на рисунку 1.

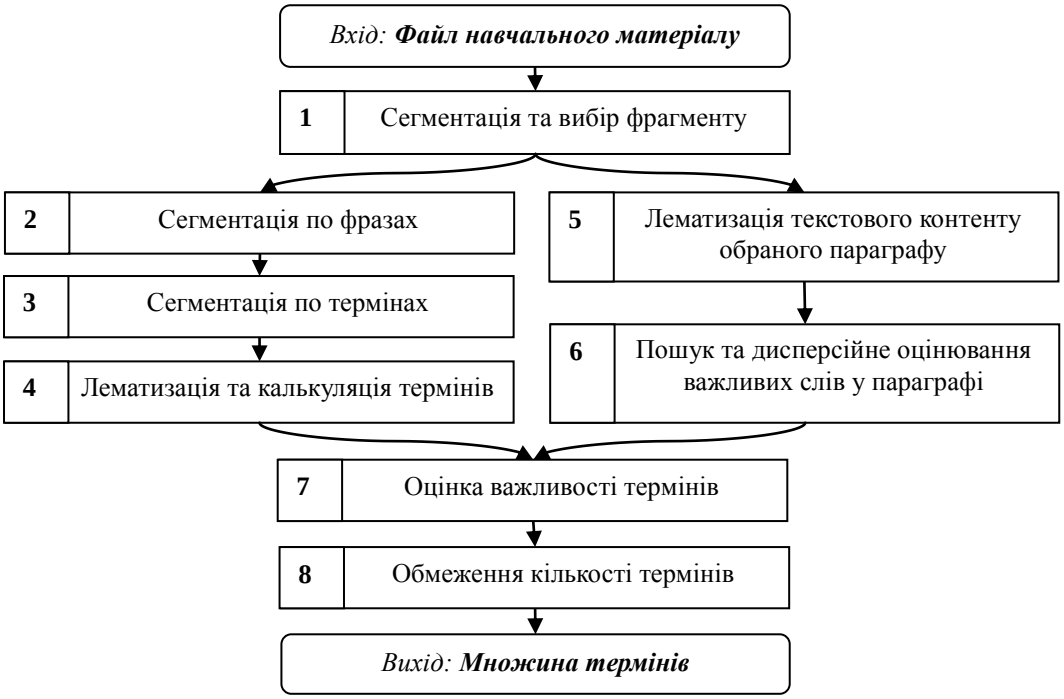
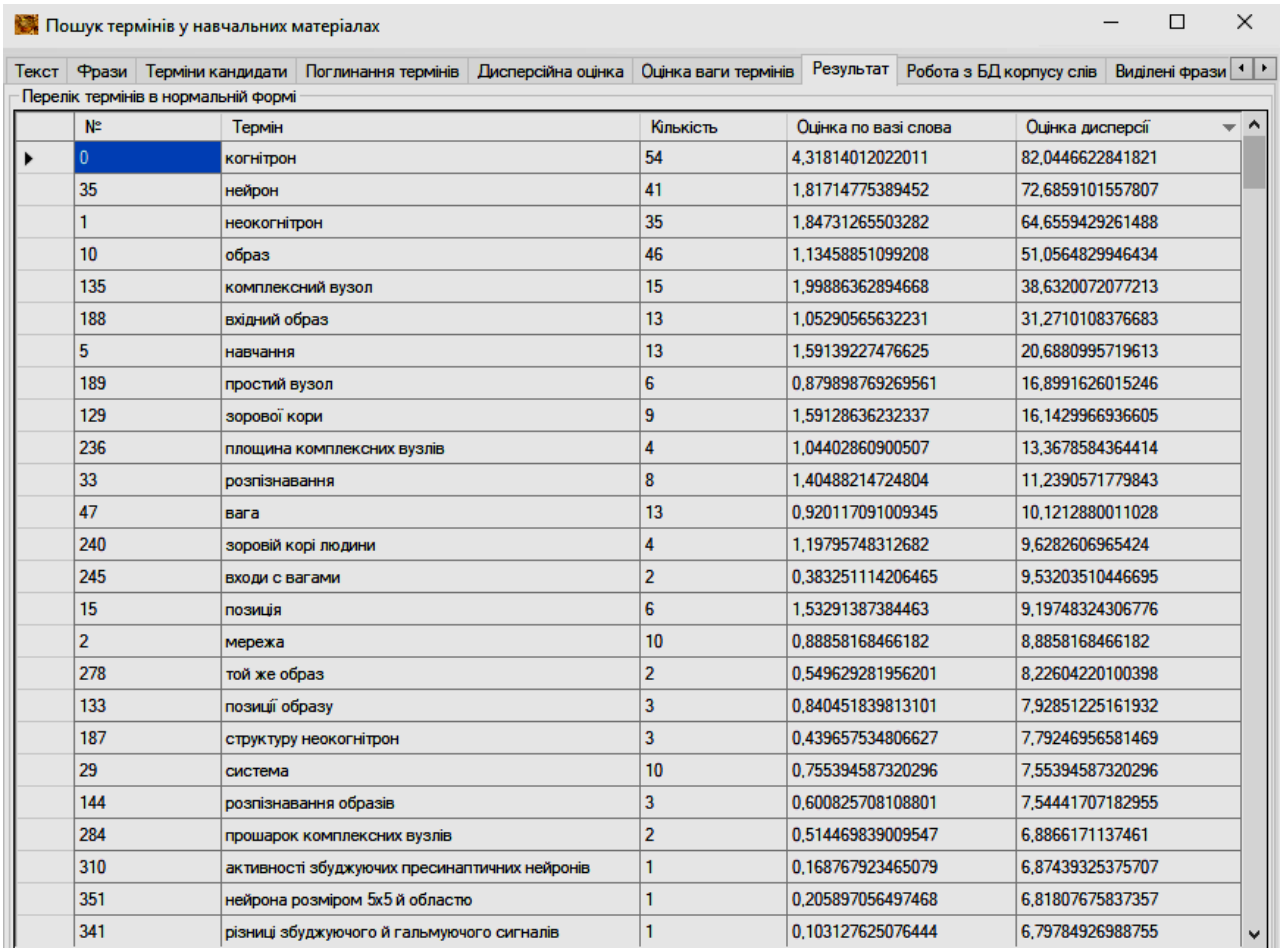


Рис. 1. Схема інформаційної технології автоматизованого множини ключових семантичних термінів

Вхідними даними інформаційної технології є електронний документ навчального матеріалу, а вихідними даними є множина ключових термінів, відповідна досліджуваному фрагменту контенту електронного документу навчального матеріалу. Інформаційна технологія 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 [1, 2].

Запропонована інформаційна технологія автоматизованого визначення семантичних термінів в елементах навчальних матеріалів була реалізована в тестовому програмному продукті. Зокрема, на рисунку 2 показано приклад обробки теми «Нейронні мережі 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. Одержання множини ключових термінів розробленою системою

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) є відношенням кількості релевантних ключових термінів, знайдених автоматично, до загальної кількості знайдених ключових термінів в досліджуваному тексті й обчислюються за формулою [3]:

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

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

Повнота пошуку R (Recall) – це відношення кількості релевантних ключових термінів, знайдених автоматично, до загальної кількості релевантних ключових термінів в досліджуваному тексті, обчислюються за формулою [3]:

$$R = \frac{\left| M_{TK}^{E} \cap M_{TK} \right|}{\left| M_{TK}^{E} \right|}.$$

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

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

де k – кількість навчальних матеріалів у тестовій вибірці.

Для визначення ефективності практичного застосування інформаційної технології, тестовим програмним продуктом було оброблено тестову вибірку з 50 файлів навчальних курсів. Так, в 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.*

Згідно вищенаведених математичних моделей, у даному випадку точність пошуку склала 0,625, а повнота пошуку склала 0,714. Відповідно, 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 Розглянуто інформаційну технологію автоматизованого визначення множини ключових семантичних термінів у контенті елементів

навчальних матеріалів й відповідний їй тестовий програмний продукт.

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%.

Встановлена ефективність запропонованої інформаційної технології сприяє її використанню для вирішення ряду актуальних задач, таких як оцінка відповідності навчальних матеріалів змістовим вимогам, автоматизація формування рефератів та анотацій до елементів навчальних матеріалів, оцінка відповідності наборів тестових завдань навчальним матеріалам, семантична допомога при створенні тестів тощо.

Література

1. Мазурець О. В. Інформаційна технологія автоматизованого визначення семантичних термінів в елементах навчальних матеріалів / О. В. Мазурець // Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2018, №3. – С.223-230.

2. Крак Ю. В. Практична реалізація інформаційної технології автоматизованого визначення множини семантичних термінів в контенті навчальних матеріалів / Ю. В. Крак, О. В. Бармак, О. В. Мазурець // Науковий журнал «Проблеми програмування». Київ, 2018, №2-3. – С.245-254.

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

4. Крак Ю. В. Практичне дослідження ефективності інформаційної технології автоматизованого визначення семантичних термінів в контенті навчальних матеріалів / Ю. В. Крак, О. В. Бармак, О. В. Мазурець // Прикладне програмне забезпечення. Збірник наукових праць за матеріалами десятої міжнародної науково-практичної конференції по програмуванню «УкрПРОГ’2016». 24 – 26 травня 2016 року; Київ – 2016. – С.237-245.