УДК 004.4

МАТЕМАТИЧНІ МОДЕЛІ ДЛЯ ВИЗНАЧЕННЯ СЕМАНТИЧНИХ ТЕРМІНІВ У КОНТЕНТІ НАВЧАЛЬНИХ МАТЕРІАЛІВ

Кондаков О.В., Мазурець О.В., Скрипник Т.К.
Україна, Хмельницький національний університет
E-mail: exe.chong@gmail.com

Поширення інформаційних технологій та розвиток глобальної мережі та телекомунікацій привели до значних змін у вищій освіті. Одними із проявів цих змін стало виникнення дистанційної форми освіти й розвиток спеціалізованих навчальних середовищ, найбільш розповсюдженим із яких наразі є Moodle [1]. Застосування спеціалізованих навчальних середовищ й електронних навчальних курсів вимагає вирішення ряду задач автоматизації, зокрема: 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. Одним із способів вирішення задачі оцінки семантичної відповідності є аналіз термінологічної бази навчальних матеріалів. Тому задача автоматизації визначення семантичних термінів у навчальних матеріалах є актуальною задачею сучасної освіти.

Термінами можуть бути як ключові слова, так і ключові словосполучення. Ключові словосполучення можуть містити довільну кількість слів, і бути семантичними мережами малої ємності. В рамках аналізу їх склад доцільно спрощувати до двох слів. При цьому в процесі пошуку як мінімум одне із цих слів можна розглядати як термін в межах навчальних матеріалів. Тож питання автоматизації пошуку ключових слів у контенті навчальних матеріалів є першочерговою задачею в процесі вирішення розглядуваної проблеми.

Метою роботи є дослідження сучасних відомих методів аналізу текстів для оцінки їх ефективності й придатності до

використання у задачі автоматизації пошуку ключових семантичних термінів у контенті навчальних матеріалів.

Застосування різноманітних методів аналізу текстів дозволяє зіставити окремим словам або словосполученням тексту деякі певним чином поставлені у відповідність числові вагові значення, що вказують на міру їх важливості в досліджуваному тексті. Ці методи розрізняються за алгоритмами обрахунку вказаних вагових значень [2]. Найбільш розповсюдженими методами аналізу текстів є частотна оцінка, оцінка TFIDF та дисперсійна оцінка.

Частотна оцінка TF (term frequency) є частотою згадувань певного слова і у тексті, що розглядається, й обчислюється наступним чином [3]:

$$Tf_i = \frac{n(i)}{\sum_k n_{ik}}, \qquad (1)$$

де *n(i)* – кількість згадувань слова *i* у тексті, $\sum_k n_{ik}$ – загальна кількість слів у тексті.

Оцінка TFIDF є добутком частоти згадувань слова у тексті *Tf* (term frequency) та зворотної документарної частоти слова *Idf* (inverse document frequency) [4]:

$$TfIdf = Tf * Idf,\ Tf_i = \frac{n(i)}{\sum_k n_{ik}},\ Idf_i = \log\frac{D}{d_i}, \qquad (2)$$

де *D* – кількість фрагментів, на які розбивається текст при аналізі; d_i – кількість фрагментів, у яких дане слово присутнє.

Дисперсійна оцінка DE за змістом близька до оцінки TFIDF, та є оцінкою дискримінантної сили слів. Вона дозволяє відділити із загального переліку широковживаних у тексті слів слова, що розташовані рівномірно. Якщо деяке слово *A* в тексті, що складається з *N* слів, позначене як A_k^n, де індекс *k* – номер появи даного слова в тесті, а *n* – позиція даного слова в тексті, то інтервал між послідовними появами слова при таких позначеннях буде величина $\Delta A_k^m = A_{k+1}^m - A_k^n = m - n$, де на *m*-ій і *n*-ій позиціях в тесті знаходиться слово *A*, яке зустрілось *k*+1-ий і *k*-ий рази. Тоді дисперсійна оцінка розраховується наступним чином [5]:

$$\sigma = \frac{\sqrt{(\Delta A^2) - (\Delta A)^2}}{(\Delta A)} \tag{3}$$

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.

Для проведення експериментів за наведеною вище схемою було розроблено тестове програмне забезпечення, що реалізує обробку контенту навчальних матеріалів трьома розглянутими методами (частотний аналіз, аналіз TFIDF та дисперсійний аналіз) з відповідними ваговими параметрами.

В процесі обробки контенту переліки ключових слів, отримані за відповідними методами, обмежуються за кількісним порогом й формують множини B_1, B_2, B_3. В подальшому ці множини порівнюються із множиною B_A, утвореною переліком ключових термінів, який сформовано автором навчального матеріалу. Перетин цих множин $B_k \bigcap B_A$ визначає ефективність відповідного методу k.

Максимальна область перетину авторського переліку зі сформованими застосунком переліками $B_k \bigcap B_A \to \max$ визначає найбільш ефективний метод автоматизації пошуку ключових семантичних термінів у контенті навчальних матеріалів.

Ефективність наведених методів пропонується визначати за наступною формулою:

$$E_k = \frac{N_{Ak}}{N_A} \cdot 100\%, \tag{4}$$

де N_{Ak} – кількість термінів у авторському (B_A) та сформованому за k-им методом (B_k) переліками термінів, що співпали ($B_k \bigcap B_A$); N_A – кількість термінів у переліку термінів B_k, сформованому експертом (автором).

В результаті тестування (на прикладі лекційного матеріалу «Введення у реляційну модель даних» навчального курсу «Вступ до реляційних баз даних» [2]) розробленим програмним забезпеченням отримуються три переліки ключових термінів за відповідними методами аналізу та проводиться їх порівняння у сукупності з авторським переліком. Деякі результати порівняння наведено у табл. 1.

На основі наведених даних дослідження, за формулою (4) побудовано діаграму ефективності розглянутих методів формування переліку ключових термінів у порівнянні з авторським переліком. В

даному випадку ефективність методу частотної оцінки склала 33,3%, методу оцінки TFIDF – 30,3%, методу дисперсійної оцінки – 84,8%.

Таблиця 1

Фрагмент порівняльної таблиці аналізу термінів

№ п/п	Термін	Визначено автором	Аналіз TF	Аналіз TFIDF	Аналіз DE
1.	реляційна база даних	+	+		+
2.	тип даних	+	+	+	+
3.	домен	+		+	+
4.	реляційна модель даних	+			+
5.	обмеження цілісності	+		+	+
6.	заголовок відношення	+			+
7.	значення відношення	+	+		+
8.	перша нормальна форма	+			+
9.	модель даних	+	+		+
10.	СКБД	+	+		+
11.	реляційне числення	+			
12.	цілісність сутності	+		+	+
13.	булевий тип	+			+
14.	зовнішній ключ	+		+	+
15.	SQL	+			
16.	заголовок відношення				+
17.	унікальність значень			+	

Аналогічним чином було досліджено 30 лекцій із різних навчальних курсів й обраховано середню ефективність кожного із методів. Середня ефективність методу частотної оцінки склала 27,1%, методу оцінки TFIDF – 45,5% та методу дисперсійної оцінки – 88,3% (рис.1).

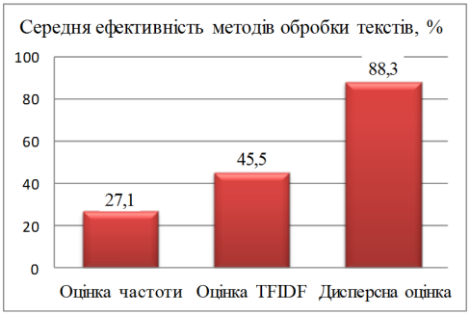

Рис. 1 – Діаграма середньої ефективності методів обробки текстів

Таким чином, метод дисперсійної оцінки продемонстрував найвищу ефективність серед досліджуваних методів, показавши при цьому мінімальну ефективність 67,7%, максимальну – 100%.

Результат застосування частотного аналізу свідчить, що цей метод надає велику вагу не тільки ключовим словам, а й словам із максимальною частотою – сполучникам, прийменникам і часткам, що відіграють велику роль для зв'язності тексту, проте не несуть навантаження з точки зору семантичної структури.

Метод TFIDF дозволяє дещо відсіяти слова, що використовуються для зв'язування тексту, через їх велике значення розповсюдженості у контенті, але значна вага надається словам, важливість яких є обмеженою в рамках локальних елементів контенту. Тому даний метод використовується переважно для аналізу масивів незв'язних текстів, й продемонстрував низьку ефективність при аналізі контенту навчальних матеріалів.

Отриманий результат аналізу контенту лекції методом дисперсійного оцінювання дозволив визначити перелік слів, найбільш близький до переліку, сформованого експертом (автором курсу).

Таким чином, у результаті дослідження методів аналізу текстів було встановлено, що найбільшу ефективність в вирішенні задачі автоматизації пошуку ключових слів у контенті навчальних матеріалів досягнуто методом дисперсійної оцінки.

Література

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

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

3. 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.

4. Ventura, J. & Silva, J. (2007). New Techniques for Relevant Word Ranking and Extraction. In Proceedings of 13th Portuguese Conference on Artificial Intelligence, Springer-Verlag, pp. 691-702.

5. 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.