МОДЕЛІ ОЦІНКИ ЕФЕКТИВНОСТІ МЕТОДІВ ПОШУКУ КЛЮЧОВИХ ТЕРМІНІВ У КОНТЕНТІ НАВЧАЛЬНИХ МАТЕРІАЛІВ

Мазурець Олександр Вікторович
старший викладач,
кафедра комп'ютерних наук та інформаційних технологій,
Хмельницький національний університет,
м. Хмельницький, Україна
exe.chong@gmail.com
Якимюк Олена Миколаївна
студент спеціальності «Комп'ютерні науки»,
Хмельницький національний університет,
м. Хмельницький, Україна
magic.fox91109@gmail.com

У сучасному суспільстві електронна інформація набуває все більшої ролі в усіх сферах життя. Це привело до значних змін і у вищій освіті. По всьому світу в інформаційних сховищах зібрані значні обсяги навчальних матеріалів з дистанційної освіти, у спеціалізованих начальних середовищах на кшталт Moodle та інших [1]. При такому розвитку інформаційних ресурсів, постає актуальність автоматизації виконання ряду задач над навчальними матеріалами — визначення їх якості, відповідності вимогам, відповідності тестовим завданням, формування рефератів та анотацій. Всі ці задачі можна вирішити шляхом формування онтології навчальних матеріалів, причому створення її нижнього рівня, ключових термінів, є найбільш складною задачею [2].

Застосування різноманітних методів аналізу текстів дозволяє зіставити окремим словам або словосполученням контенту навчальних матеріалів деякі певним чином поставлені у відповідність числові вагові значення, що вказують на міру їх важливості в досліджуваному тексті. Ці методи розрізняються за алгоритмами обрахунку цих вагових значень [3], а найбільш розповсюдженими методами аналізу текстів є частотна оцінка, оцінка TFIDF та дисперсійна оцінка. Відтак, постає проблема визначення найбільш ефективного методу для пошуку ключових термінів (слів та словосполучень) у контенті навчальних матеріалів.

Було встановлено наступні математичні моделі для оцінки методів пошуку ключових термінів у навчальних матеріалах [4]:

— Точність — відношення числа релевантних ключових термінів знайдених автоматично, до загальної кількості знайдених ключових термінів в документі;

— Повнота — це відношення числа релевантних ключових термінів знайдених автоматично, до загальної кількості релевантних ключових термінів в документі;

— F-міра (F-score) — об'єднання точності і повноти в одній усередненій величині, визначається як зважене гармонійне середнє точності і повноти.

В рамках експерименту було розроблено тестове програмне забезпечення (рис. 1), що реалізує обробку контенту навчальних матеріалів трьома розглянутими методами (частотний аналіз, аналіз TFIDF та дисперсійний аналіз) з відповідними ваговими параметрами.

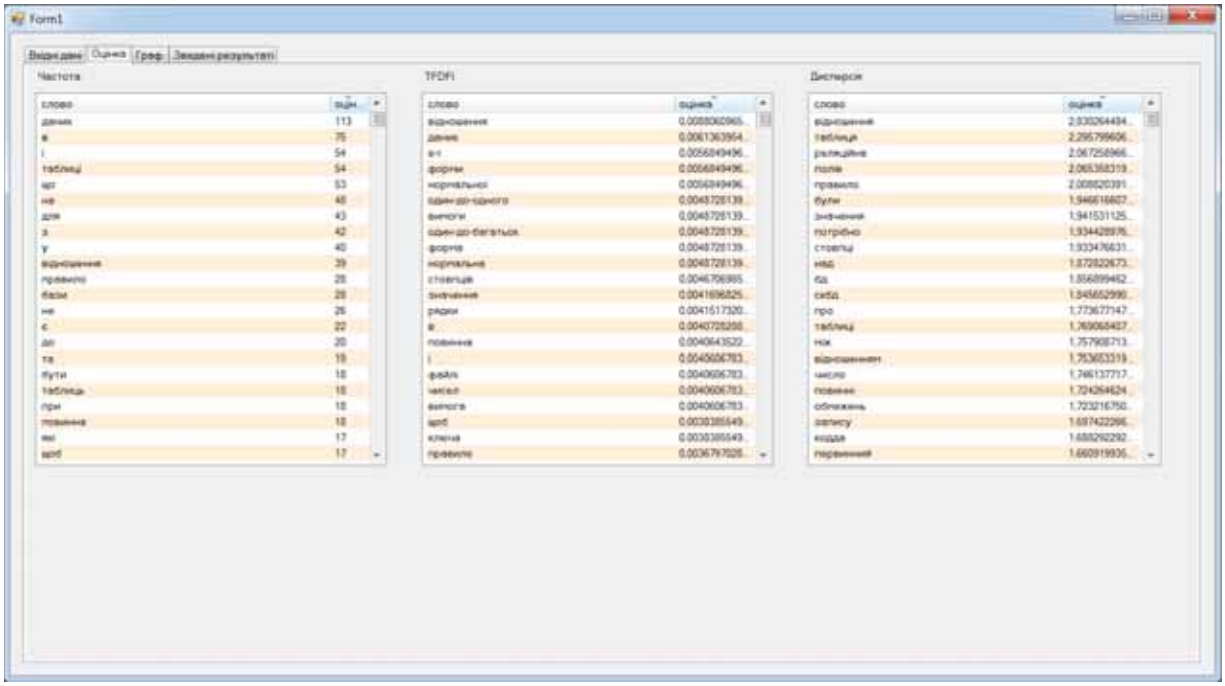

Рис. 1. Результат аналізу контенту тестовим програмним забезпеченням

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 експертом — автором відповідного навчального матеріалу. Перетин цих множин $B_k \cap B_A$ визначає ефективність відповідного методу k.

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

Ефективність наведених методів пропонується визначати за наступною формулою:

$$E_k = \frac{N_{Ak}}{N_A} \cdot 100\%, \quad (1)$$

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

В результаті тестування (на прикладі лекційного матеріалу «Реляційна модель даних» навчального курсу «Організація баз даних та знань») розробленим програмним забезпеченням отримуються три переліки ключових термінів за відповідними методами аналізу та проводиться їх порівняння у сукупності з авторським переліком.

Рис. 2. Діаграма ефективності методів обробки текстів

На основі наведених даних дослідження, за формулою (1) побудовано діаграму ефективності розглянутих методів формування переліку ключових термінів у порівнянні з авторським переліком (рис. 2). Ефективність методу частотної оцінки склала 33,3 %, методу оцінки TFIDF — 30,3 %, методу дисперсної оцінки — 84,8 %.

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

Рис. 3. Діаграма середньої ефективності методів обробки текстів

Отже, було розглянуто моделі оцінки ефективності методів пошуку ключових термінів у контенті навчальних матеріалів. За результатами проведених досліджень, метод дисперсійної оцінки продемонстрував найвищу ефективність серед досліджуваних методів, показавши при цьому мінімальну ефективність 67,7 %, максимальну — 100 %.

Список використаних джерел:

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

2. Мазурець О. В. Інформаційна технологія побудови онтологічної моделі навчального курсу для оцінювання отриманих знань / О. В. Мазурець // Матеріали III міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології». Одеса — 2014. — С. 81–83.

3. Ландэ Д. В., Снарский А. А. 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» — КПИ, Киев: 2013. — С. 158–164.

4. Полубок А. М. Інтелектуальний аналіз текстів на основі самоорганізації / А. М. Полубок [Електронний ресурс]. — Режим доступу: http://mmsa.kpi.ua/sites/default /files/abstracts/ 2017_b_sa_smdm_ polubok_am_uk_presentation.pdf

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