УДК 004.91

Шпичко А.В., Мазурець О.В.

Хмельницький національний університет, Україна

МЕТОДИ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ СЕМАНТИЧНИХ ТЕРМІНІВ У ЦИФРОВИХ ТЕКСТАХ

Shpychko A.V., Mazurets A.V.

METHODS OF AUTOMATED DETERMINATION OF SEMANTIC TERMS IN DIGITAL TEXTES

SEO-оптимізація є комплексом заходів по оптимізації сайту з метою поліпшення його позицій в пошукових системах й включає не лише підбір семантичних термінів у вигляді множини ключових слів, що відповідають тематиці сайту та конкретної сторінки, але і визначення частотного характеру їх вживання.

В топ пошукової видачі виходять веб-документи, тексти яких містять певні слова або фрази, що відповідають запитам користувачів. Пошуковими роботами враховується і кількість входжень ключових слів, і відповідність тематики сайту запитам та багато інших параметрів, над поліпшенням яких працюють SEO-оптимізатори.

Найбільш відомими методами пошуку ключових слів наразі є TF, TF-IDF, DE та BM25 [1]. З метою дослідження ефективності автоматизованого визначення семантичних термінів цими методами, було створено тестову програмну систему (рис. 1). Даний програмний продукт проводить аналіз тексту, знаходить ключові слова за допомогою методів TF, TF-IDF, DE і BM25, й дозволяє порівнювати одержані результати із множиною семантичних термінів, створених автором тексту.

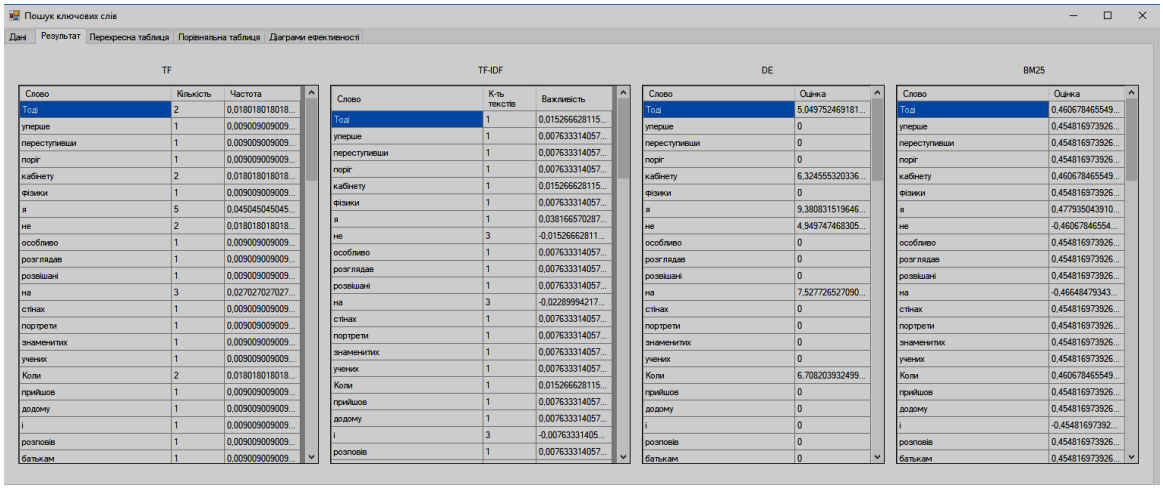

Рис. 1. Програмна система для автоматизованого визначення семантичних термінів у цифрових текстах

Ефективність автоматизованого визначення семантичних термінів для кожного з методів було визначено за допомогою показника повноти пошуку R, який виражає відношення кількості релевантних ключових слів, знайдених автоматично, до загальної кількості релевантних (визначених автором) ключових слів в тексті. Відповідно, середня повнота пошуку $\overline{R}$ для тестової вибірки із k текстів визначається наступним чином:

$$\overline{R} = \frac{\sum_{i=1}^{k} R_k}{k}, R = \frac{\left| M_{TK}^{E} \cap M_{TK} \right|}{\left| M_{TK}^{E} \right|}, \quad (1)$$

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

Для аналізу було використано k = 90 цифрових текстів, всі методи визначення ключових слів одержали однакові вибірки цифрових текстів у якості вхідних даних. Результат автоматизованого визначення

семантичних термінів для кожного з методів свідчить (рис. 2), що метод дисперсійної оцінки продемонстрував найвищу ефективність серед досліджуваних методів, показавши при цьому середню ефективність 72,8%, мінімальну – 57,5%, максимальну – 100%.

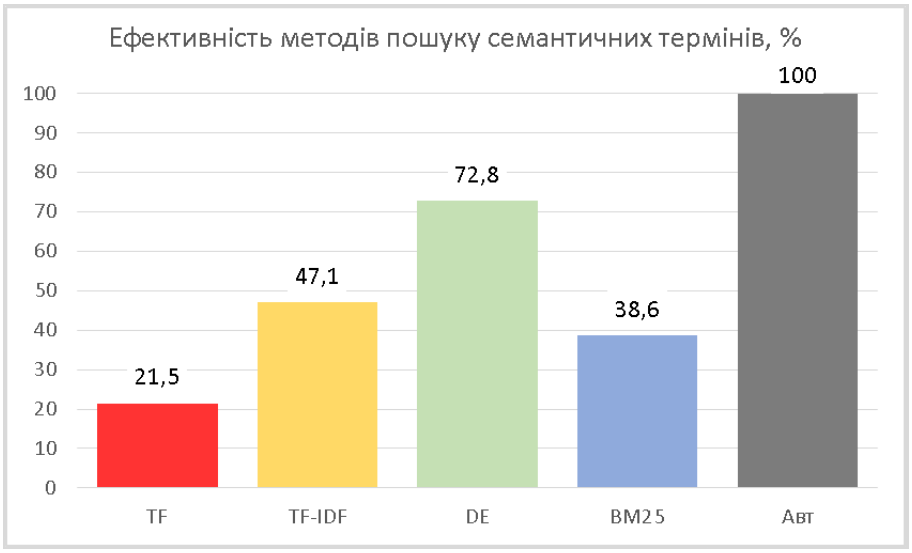

Рис. 2. Програмна система для автоматизованого визначення семантичних термінів у цифрових текстах

Отже, в результаті дослідження було встановлено, що найбільшу ефективність у вирішенні задачі автоматизації пошуку ключових слів у цифрових текстах досягнуто методом дисперсійної оцінки. Це відкриває перспективи для її використання в системах семантичного аналізу текстів.

Список літератури

1. Ланде Д. В. Методи оцінки рівня дискримінантної сили слів у текстах з правової тематики / Д. В. Ланде // Правова інформатика. Київ, 2012, №3(35). – С.3-7.