

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.

TF			TF-IDF			DE		BM25	
Слово	Кількість	Частота	Слово	Кількість текстів	Важливість	Слово	Оцінка	Слово	Оцінка
Тоді	2	0.01801801818	Тоді	1	0.015266628115	Тоді	5.049752469181	Тоді	0.460678465549
улерше	1	0.00909090909	улерше	1	0.007633314057	улерше	0	улерше	0.454816973926
переступивши	1	0.00909090909	переступивши	1	0.007633314057	переступивши	0	переступивши	0.454816973926
порг	1	0.00909090909	порг	1	0.007633314057	порг	0	порг	0.454816973926
кабінету	2	0.01801801818	кабінету	1	0.015266628115	кабінету	6.324555320336	кабінету	0.460678465549
фзвжи	1	0.00909090909	фзвжи	1	0.007633314057	фзвжи	0	фзвжи	0.454816973926
я	5	0.04545454545	я	1	0.038166570287	я	9.380831519646	я	0.477935043910
не	2	0.01801801818	не	3	-0.015266628115	не	4.949747468305	не	-0.460678465549
особливо	1	0.00909090909	особливо	1	0.007633314057	особливо	0	особливо	0.454816973926
розглядав	1	0.00909090909	розглядав	1	0.007633314057	розглядав	0	розглядав	0.454816973926
розважані	1	0.00909090909	розважані	1	0.007633314057	розважані	0	розважані	0.454816973926
на	3	0.027027027027	на	3	-0.02289994217	на	7.527726527090	на	-0.46648479343
стінах	1	0.00909090909	стінах	1	0.007633314057	стінах	0	стінах	0.454816973926
портрети	1	0.00909090909	портрети	1	0.007633314057	портрети	0	портрети	0.454816973926
значенитих	1	0.00909090909	значенитих	1	0.007633314057	значенитих	0	значенитих	0.454816973926
ученик	1	0.00909090909	ученик	1	0.007633314057	ученик	0	ученик	0.454816973926
Коли	2	0.01801801818	Коли	1	0.015266628115	Коли	6.708203932499	Коли	0.460678465549
прийшов	1	0.00909090909	прийшов	1	0.007633314057	прийшов	0	прийшов	0.454816973926
додаму	1	0.00909090909	додаму	1	0.007633314057	додаму	0	додаму	0.454816973926
і	1	0.00909090909	і	3	-0.007633314057	і	0	і	-0.454816973926
розповів	1	0.00909090909	розповів	1	0.007633314057	розповів	0	розповів	0.454816973926
батькам	1	0.00909090909	батькам	1	0.007633314057	батькам	0	батькам	0.454816973926

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 \bar{R} для тестової вибірки із k текстів визначається наступним чином:

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

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%.

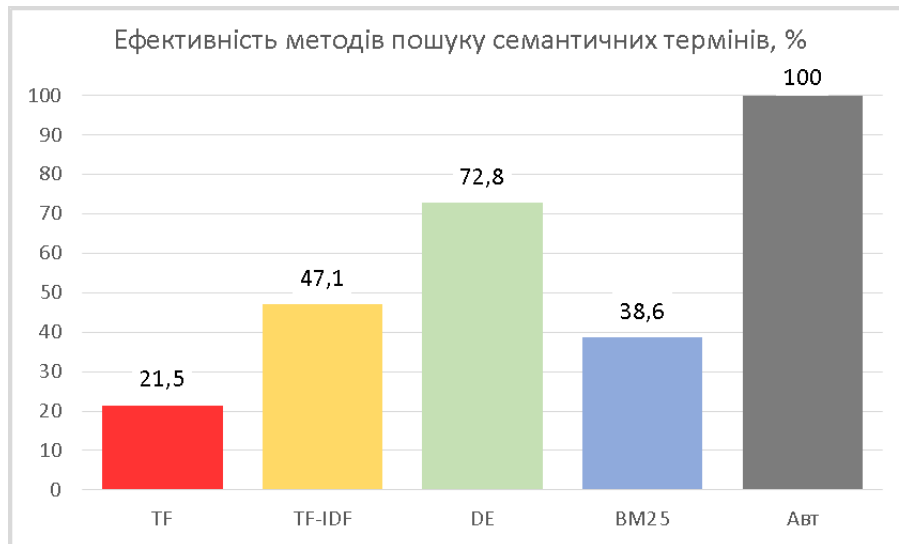


Рис. 2. Програмна система для автоматизованого визначення семантичних термінів у цифрових текстах

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.