УДК 004.91

Мороз О.О., Мазурець О.В.

Хмельницький національний університет, Україна

МАТЕМАТИКО-АЛГОРИТМІЧНА МОДЕЛЬ ДЛЯ ОБМЕЖЕННЯ МНОЖИНИ КЛЮЧОВИХ ТЕРМІНІВ У ЦИФРОВИХ ТЕКСТАХ

Moroz O.O., Mazurets A.V.

MATHEMATICAL-ALGORITHMIC MODEL FOR LIMITATION OF THE NUMBER OF KEY TERMS SET OF DIGITAL TEXTS

Семантичний аналіз текстів є важким математичним завданням, яке ускладнюється необхідністю обробки природної мови. Множину ключових слів цифрового текстового документу називають пошуковим образом цього документу.

Множина ключових слів тексту є найбільш семантично стиснутим результатом семантичного аналізу тексту, й якість її автоматизованого формування в рамках семантичного аналізу тексту визначається ефективністю використовуваних методів пошуку ключових слів.

Для автоматизації пошуку ключових слів використовуються різноманітні методи аналізу текстів, таких як TFIDF, частотна оцінка та дисперсійна оцінки тощо [1].

Хоча використання таких методів є ефективним, вихідними даними є множина ключових термінів із кількістю елементів, рівною кількості оригінальних слів у документі. Оскільки кожному елементу співставляється деяке певним чином поставлене у відповідність числове вагове значення, то множина може бути впорядкована за важливістю відповідних термінів, але проблему складає необхідність

обмеження вихідної множини. Очевидно, що даний етап є останнім при формування вихідної множини ключових термінів (рис. 1).

Рис. 1. Етапи формування множини ключових термінів

Запропоновано визначати рекомендовану кількість ключових термінів у цифрових текстах номінально похідною від параметру щільності ключових термінів [2]. Щільність ключових термінів є відношенням кількості ключових термінів у тексті до загальної кількості слів у даному тексті.

Відповідно, до порожньої результуючої множини ключових термінів M_{TK} додаються терміни з загальної впорядкованої множини термінів M_{T1} з найбільшими значеннями оцінки важливості доти, доки справджується рівність:

$$\sum_{i=1}^{n} \frac{K_n x_n}{X_{txt}} \le P_{txt}, \qquad (3)$$

де K_n – кількість появ терміну n у множині M_{T1}; x_n – кількість слів в терміні n; X_{txt} – загальна кількість слів у тексті; n – поточна кількість термінів у результуючій множині ключових термінів M_{TK}.

Експериментально встановлено, що якщо щільність виражається у відсотковому співвідношенні, то оптимальна щільність ключових термінів у звичайному тексті художнього стилю становить 4-6%, хоча даний параметр залежить від типу тексту й, наприклад, у навчальних матеріалах може сягати 18%. Відповідно, рекомендована кількість ключових термінів для цифрового тексту, яка технічно є порогом для відсікання неважливих елементів, має бути такою, щоб сукупна щільність ключових термінів у тексті не перевищувала заданий відсотковий поріг.

Розглянуті результати досліджень призначені для удосконалення розробленої інформаційної технології визначення ключових термінів у цифрових текстах [3], що дозволить підвищити ефективність роботи відповідних інформаційних систем шляхом підвищення рівня автоматизації.

Список літератури

1. 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.

2. Ключові слова. iGroup Україна – [Електронний ресурс]. – Режим доступу: http://igroup.com.ua/seo-articles/keywords/

3. Krak I., Barmak O., Mazurets O. The practice investigation of the information technology efficiency for automated definition of terms in the semantic content of educational materials I. Krak, O. Barmak, O. Mazurets // CEUR Workshop Proceedings. Proceedings of the 10th International Conference of Programming UkrPROG'2016. p. 237-245.