УДК 004.8

Тимуш О.Ю., Шпичко А.В., Мазурець О.В.

Хмельницький національний університет

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ТЕМАТИЧНОГО СОРТУВАННЯ ТЕКСТОВИХ ПОВІДОМЛЕНЬ

У статті розглянуто інформаційну технологію сортування текстових повідомлень за тематикою. На основі розробленої інформаційної технології тематичного сортування текстової інформації було створено два програмних продукти: систему визначення множин ключових слів для рубрик новин та систему тематичного сортування новин. Одержані результати дослідження ефективності показали, що в переважній більшості випадків програмна система, виконана відповідно до запропонованої інформаційної технології, успішно виконала сортування новин за рубриками, й середня успішність сортування за рубриками склала 94,4%.

The article considers the information technology for thematic classification of text messages. Based on the developed information technology of thematic sorting of textual information, two software products were created: a system of definition the keywords sets for news headings and a system of thematic sorting of news. The results of the information technology efficiency investigation showed that in most cases the software system, which was made in accordance with the proposed information technology, successfully completed news sorting by headings, and the average success of sorting by headings was 94.4%.

Розвиток інформаційних технологій та глобальної мережі призвели до надання відкритого доступу пересіченому користувачу до великих обсягів інформації. Інформація, що представлена здебільшого в текстовому вигляді, не може бути сприйнята в доступних обсягах. Тому є доречним її фільтрування за певними критеріями відповідно до інтересів і вподобань клієнта. Якщо взяти за об'єкт дослідження стрічки новин, то такими критеріями можуть бути ключові слова окремих новин та тематичні рубрики, до яких вони відносяться [1]. Автоматизація такого сортування текстової інформації є ефективним інструментом, що заощаджує час користувача й підвищує якість роботи новинних агрегаторів [2], що визначає актуальний напрямок наукових досліджень.

Інформаційна технологія тематичного сортування текстової інформації призначена для одержання за вхідною інформацією у вигляді цифрового текстового контенту вихідної інформації у вигляді оцінок

приналежності даного контенту до кожної із відомих категорій. В даному випадку як область застосування розглядаються сайти новин, відповідно вхідними даними виступає цифровий текстовий контент новини, а категоріями для сортування – рубрики новин.

Вибіркою вхідних даних інформаційної технології тематичного сортування текстової інформації на прикладі новин є навчальні множини випадкових новин для кожної з рубрик (кількість множин рівна кількості рубрик) та тестова новина для аналізу приналежності до рубрик.

Першим етапом обробки даних є визначення множин ключових слів для рубрик новин (Рис. 1). На цьому етапі використовуються тільки вхідні дані у вигляді навчальних множин новин для кожної з рубрик. Також на цьому етапі проводиться обрахунок значень оцінки TFIDF для кожного оригінального слова для кожної з навчальних множин новин [3].

Після чого для кожної з одержаних множин ключових слів, що містять також значення оцінки TFIDF [4], проводиться сортування за спаданням оцінки TFIDF та обмеження кількості слів у множині за порогом в 30 слів, що визначений емпірично й може бути змінений відповідно до потужності навчальних множин новин. Вихідними даними цього етапу обробки даних є множини з 30 ключових слів для кожної з рубрик.

Другим етапом обробки даних інформаційної технології тематичного сортування текстової інформації на прикладі новин є визначення приналежності тестової новини до рубрик новин. Вхідними даними для цього етапу обробки даних є множини з 30 ключових слів для кожної з рубрик і тестова новина для аналізу приналежності до рубрик. На основі цих даних проводиться обрахунок кількостей збігів за ключовими словами рубрик.

Кількостей збігів за ключовими словами для кожної з рубрик є числовою оцінкою приналежності тестової новини до кожної з рубрик. Наступним кроком на основі одержаних даних проводиться обрахунок відсоткового значення приналежності. Вихідними даними цього заключного етапу обробки даних інформаційної технології тематичного сортування текстової інформації на прикладі новин є одержані оцінки приналежності тестової новини до кожної з актуальних рубрик новин.

Визначення приналежності тестової новини до актуальних рубрик новин проводиться шляхом порівняння множини ключових слів та множини слів контенту новини рубрикам. Для цього для кожної з рубрик спочатку визначається перетин множин ключових слів та множини слів контенту новини. Так, для *n*-ї рубрики множина слів такого перетину S_{Per}^{n} визначається наступним чином:

$$S_{Per}^{n} = S_{Cont} \bigcap S_{Words}^{n}, \quad (1)$$

де S_{Cont} – множина слів контенту новини; S_{Words}^{n} – множина ключових слів для *n*-ї рубрики.

Рисунок 1 – Загальна схема інформаційної технології

Відповідно, кількість збігів за ключовими словами L для *n*-ї рубрики рівна потужності множини слів S_{Per}^{n} перетину множини слів контенту новини та множини ключових слів для *n*-ї рубрики $L = \left| S_{Per}^{n} \right|$.

Дослідження ефективності інформаційної технології тематичного сортування текстової інформації виконано за результатами застосування розроблених тестових програмних систем, що виконані на засадах розробленої інформаційної технології (Рис. 2). Для експерименту, з використанням системи визначення множин ключових слів для рубрик

новин було сформовано 6 множин по 30 ключових слів для кожної з 6 категорій (рубрик) новин, до яких було віднесено наступні: Політика, Економіка, Наука, Туризм, Спорт, Здоров'я. В якості вхідних даних для кожної рубрики необхідно використано вибірки з 100 новин кожна. Сформовані множини ключових слів для рубрик новин було використано в роботі системи тематичного сортування новин.

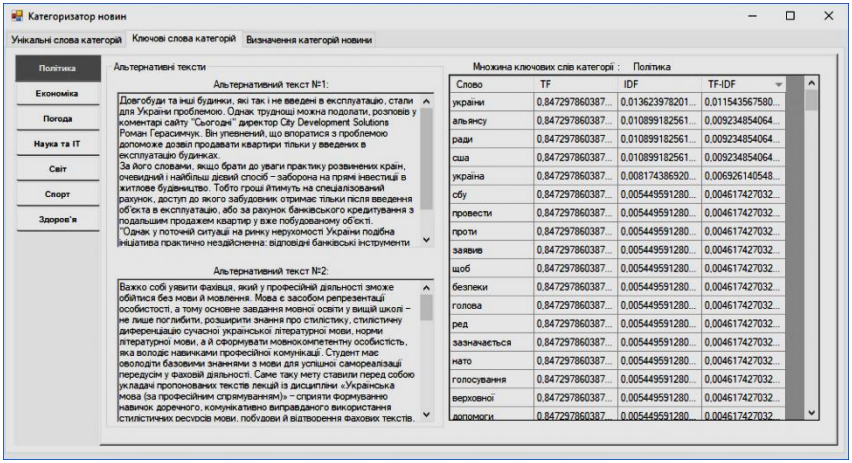

Рисунок 2 – Інтерфейс програмної системи для визначення множин ключових слів рубрик новин

Наступним кроком було використання системи тематичного сортування новин для автоматизованого визначення приналежності тестових зразків новин до актуальних рубрик новин. Для цього було сформовано 6 множин по 15 тестових новин для кожної з 6 рубрик, причому новини з тестових вибірок не були використані для навчання системи.

Обраховані показниками приналежності тестових новин до рубрик були використані для остаточного визначення приналежності кожної новини до однієї рубрики, для цього обиралися рубрики з найбільшими показниками приналежності.

Перевірка коректності сортування новин за категоріями полягала у визначенні відповідності сортування системою тематичного сортування новин із сортуванням, що було реалізовано на сайті новин – джерелі експериментальних зразків.

Результати проведеного експерименту з дослідження ефективності інформаційної технології тематичного сортування текстової інформації відповідно до наведених вище умов наведено у таблиці 1.

Наведені результати свідчать, що в більшості випадків сортування новин за рубриками системою тематичного сортування новин було виконано вірно.

Таблиця 1 – Результати ефективності тематичного сортування новин

Тип результату	**Результати за рубриками**						**Всього**
	Політика	**Економіка**	**Наука**	**Туризм**	**Спорт**	**Здоров'я**	
Коректно	15	13	14	13	15	15	85
Не коректно	0	2	1	2	0	0	5
Загалом	15	15	15	15	15	15	90

Для визначення успішності сортування новин за рубриками U було використано відношення кількості вірних результатів до загальної кількості одержаних результатів:

$$U = \frac{T_{Ok}}{T_{All}}, \tag{2}$$

де T_{Ok} – кількість коректних результатів сортування новин; T_{All} – загальна кількість одержаних результатів сортування новин.

У вигляді діаграми відсоткові результати успішності тематичного сортування новин за рубриками наведені на Рис. 3.

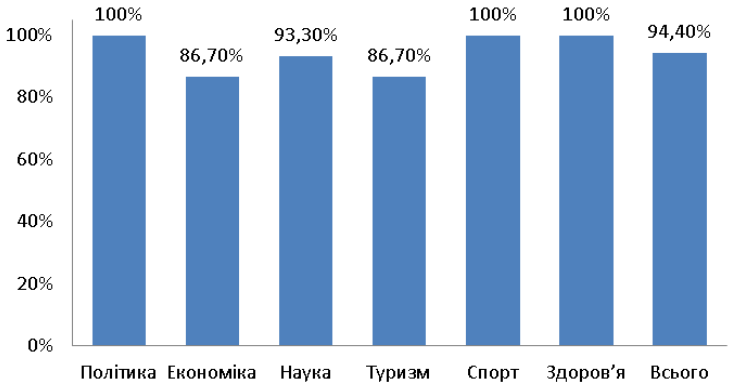

Рисунок 3 – Діаграма відсоткових результатів успішності тестового тематичного сортування новин за рубриками

Одержані результати свідчать, що для категорій «Політика», «Спорт» та «Здоров'я» успішність сортування за рубриками сягала 100%, проте для категорій новин «Економіка», «Наука» й «Туризм» були відзначені випадки невірної класифікації тестових зразків новин, що знизило успішність сортування новин за деякими рубриками до 86,7%. З наведеного можна зробити висновок, що в переважній більшості випадків програмна система, виконана відповідно до запропонованої інформаційної технології тематичного сортування текстової інформації, успішно виконала сортування новин за рубриками.

Для підвищення ефективності роботи системи можна збільшити навчальні вибірки новин для рубрик, що дозволить більш точно визначати відповідні множини ключових слів рубрик. Проте частина похибок може випливати із особливостей предметної області, наприклад, коректності підбору рубрик новин до загальної множини, оскільки деякі рубрики можуть семантично перетинатись. З другого боку, деякі новини, що на сайтах новин належать певним рубрикам, цілком коректно можуть бути автоматизовано віднесені до інших рубрик за їх контентом. Цьому явищу є характерні аналогії в предметній області, коли говорять, наприклад, про комерціалізацію спорту чи політизацію економіки.

Перелік посилань

1. Ventura J. New Techniques for Relevant Word Ranking and Extraction / J. Ventura, J. Silva // Proceedings of the artificial intelligence 13th Portuguese conference on Progress in artificial intelligence, EPIA'07. – Berlin: Springer-Verlag, Berlin, Heidelberg, 2007. – С.691-702.
2. RSS 2.0 Specifications [Електронний ресурс]. – Режим доступу: http://www.rssboard.org/rss-specification
3. Шпичко А. В. Методи автоматизованого визначення семантичних термінів у цифрових текстах / А. В. Шпичко, О. В. Мазурець // Матеріали VIII Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ICST-ODESSA-2019». Одеса – 2019. – С.166-168.
4. Мазурець О. В. Моделі оцінки ефективності методів пошуку ключових термінів у контенті навчальних матеріалів / О. В. Мазурець, О. М. Якимюк // Матеріали Всеукраїнської науково-практичної конференції з міжнародною участю «Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи». Тернопіль – 2017. – С.258-261.