DOI 10.31891/2307-5732-2019-277-5-203-210
УДК 004.9

О.В. МАЗУРЕЦЬ, О.Ю. ТИМУШ, А.П. ФЕДОРКО
Хмельницький національний університет

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ТЕМАТИЧНОЇ КЛАСИФІКАЦІЇ ТЕКСТОВИХ ПОВІДОМЛЕНЬ

У статті розглянуто інформаційну технологію сортування текстових повідомлень за тематикою. При цьому використовуються розроблений підхід до визначення множин ключових слів для рубрик новин на основі методу оцінки TFIDF та розроблені математико-алгоритмічні моделі для визначення приналежності тестової новини до актуальних рубрик новин. 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. Система визначення множин ключових слів для рубрик новин дозволяє за введеною множиною новин, що мають приналежність до певної конкретної рубрики, визначити множину ключових слів, які розглядаються як еквівалент узагальненого семантичного вмісту для новин цієї рубрики. В результаті використання програмної системи для аналізу вхідних даних у вигляді множин новин для всіх актуальних рубрик, одержуються вихідні дані у вигляді відповідної кількості множин ключових слів, які розглядаються в подальшому як портрети новин цих рубрик. Система тематичного сортування новин дозволяє за вхідними даними у вигляді текстового контенту тестової новини одержати вихідні дані у вигляді цифрових показників, що відображають оцінку приналежності тестової новини до кожної з рубрик. Для цього проводиться автоматизоване порівняння множини слів із контенту новини та множин ключових слів рубрик новин. Розроблені тестові програмні системи були використані для дослідження ефективності інформаційної технології тематичного сортування текстової інформації. Для цього проводилось автоматизоване визначення рубрик для тестових зразків новин за допомогою розроблених програмних продуктів. Одержані результати дослідження ефективності інформаційної технології показали, що в переважній більшості випадків програмна система, виконана відповідно до запропонованої інформаційної технології тематичного сортування текстової інформації, успішно виконала сортування новин за рубриками, й середня успішність сортування за рубриками склала 94,4%.

Ключові слова: текстові повідомлення, класифікація, ключові слова.

O. MAZURETS, O. TYMUSH, A. FEDORKO
Khmelnytskyi National University

INFORMATION TECHNOLOGY FOR THEMATIC CLASSIFICATION OF TEXT MESSAGES

The article considers the information technology for thematic classification of text messages. Developed approach is used to define the sets of keywords for news headings based on the TFIDF evaluation method and developed mathematical and algorithmic models to determine the affiliation of test news to current news headings. 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 news keywords sets definition system allows you to define a set of keywords from the provided data that is considered to be equivalent to generalized semantic content for news items. As a result of using the software system for analysing the input data in the sets of news for all relevant thematic, the output is received in the form of an appropriate number of sets of keywords, which are subsequently considered as news portraits of these sections. The system of news thematic sorting allows the input data in the form of textual content of the test news to get the output data in the form of digital metrics that reflect the assessment of the test news belonging to each heading. For this, an automated comparison of the plurality of news content words and the plurality of news headline keywords is performed. The developed test software systems were used to investigate the effectiveness of information technology themed textual sorting. For this purpose, automated thematic definition for test news samples was carried out using developed software products. 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 of thematic sorting of text information, successfully completed news sorting by headings, and the average success of sorting by headings was 94.4%.

Keywords: text messages, classification, keywords.

Постановка проблеми в загальному вигляді

Поширення інформаційних технологій та розвиток глобальної мережі призвели до надання відкритого доступу пересіченому користувачу до великих обсягів інформації. Інформація, представлена здебільшого в текстовому вигляді, не може бути сприйнята в доступних обсягах. Тому є доречним її фільтрування за певними критеріями відповідно до інтересів та 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 та підвищує якість роботи новинних агрегаторів, що формує актуальний напрямок наукових досліджень.

Аналіз останніх досліджень

Визначення ключових слів для новин та їх приналежності рубрикам може проходити на стороні клієнта (споживача) або на стороні серверу (сайту новин). Забезпечення виконання цих функцій на стороні сайту новин збільшує привабливість такого сайту та підвищує зручність його використання. Тому одними із функцій сайту новин є групування новин за категоріями та визначення для них переліків ключових слів, чим займаються технічні редактори.

Таким чином, при роботі технічного редактора сайту новин до кожної новини потрібно одержати:

- рубрику (тематичну категорію) новини для визначення місця розміщення новини на сайті;
- перелік ключових слів, який використовується в пошукових запитах користувачів та пошукових систем.

Більшість сучасних браузерів та поштових клієнтів працюють з RSS-стрічками [2]. Крім того, існують спеціалізовані програми (RSS-агрегатори), які збирають і опрацьовують інформацію RSS-каналів. Також дуже популярними є веб-агрегатори, які спеціалізуються на збиранні та відображенні RSS-каналів.

Оскільки новини в RSS-стрічках в метаданих в ряді випадків не містять тематичної категорії та переліків ключових слів, але вони потрібні для обробки технічним редактором сайту новин й визначаються вручну, автоматизація процесу визначення цих даних дозволить підвищити ефективність роботи технічних редакторів новинних сайтів. З цією метою необхідно проводити семантичний аналіз цих новин в RSS-стрічках.

Семантичний аналіз тексту є етапом у послідовності дій алгоритму автоматичного розуміння тексту, що полягає у виділенні семантичних відношень і формуванні семантичного представлення. Результати семантичного аналізу можуть бути застосовані для рішення задач у різних областях. В даному випадку, задачею є визначення множин ключових слів для окремих рубрик новин за множинами зразків новин відповідних рубрик.

Застосування різноманітних методів аналізу текстів дозволяє зіставити окремим словам або словосполученням тексту деякі певним чином поставлені у відповідність числові вагові значення, що вказують на міру їх важливості в досліджуваному тексті [3]. Ці методи розрізняються за алгоритмами обрахунку вказаних вагових значень [4]. Для автоматизованого пошуку ключових слів використовуються різноманітні методи аналізу текстів, серед яких найбільш відомими є частотна оцінка TF, оцінка TF-IDF, дисперсійна оцінка DE, оцінка ранжування BM тощо [5].

Постановка задачі

Метою роботи є розробка інформаційної технології тематичного сортування текстової інформації та програмного забезпечення для перевірки його ефективності на прикладі автоматизованого сортування новин по рубриках.

Викладення основних матеріалів дослідження

RSS-стрічки. Зазвичай масиви новин передаються по мережі у вигляді RSS-стрічок. Загалом, RSS є спеціальним форматом, який призначений для опису стрічок новин, анонсів статей, змін у блогах тощо. Інформація з різних джерел, подана у форматі RSS, може бути зібрана, опрацьована і подана користувачеві в зручному для нього вигляді спеціальними програмами (рис. 1). Наприклад, за допомогою RSS подається короткий опис нової інформації, що з'явилася на сайті, і посилання на її повну версію. Інтернет-ресурс у форматі RSS називається RSS-каналом, RSS-стрічкою або RSS-фідом.

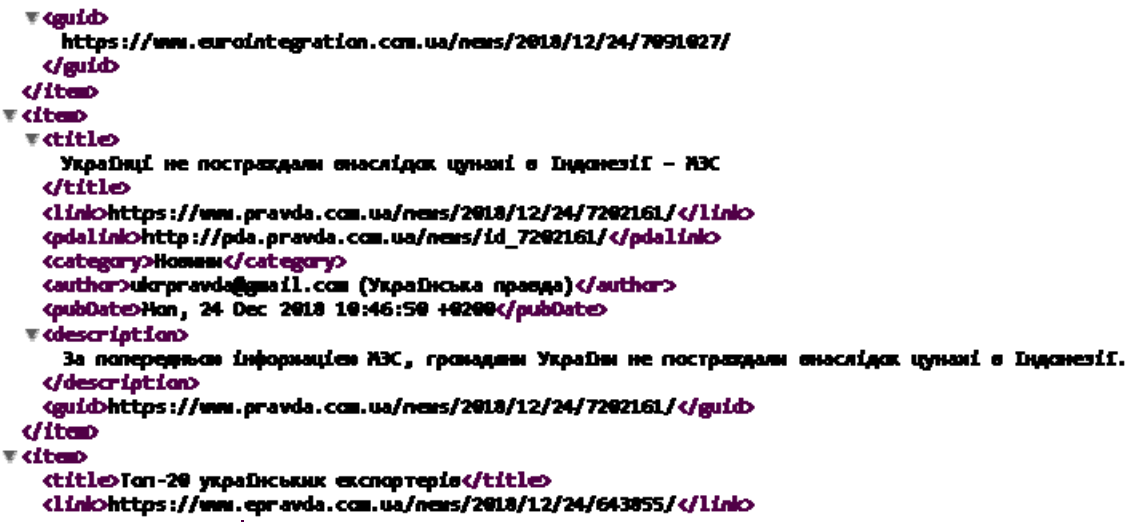

Рис. 1. Фрагмент RSS-стрічки

Програмно RSS є родиною XML-форматів, яка використовується для публікації та постачання інформації, що часто змінюється, наприклад, нових записів в блозі, заголовків новин, анонсів статей, зображень, аудіо і відеоматеріалів (в стандартизованому форматі). Документ в стандарті RSS (який також інколи називають «стрічкою», «веб-стрічкою» або «каналом») складається з повного або часткового тексту і метаданих (дата і авторство) [2].

Множина рубрик новин. Для визначення практично доцільної множини рубрик новин було проведено аналіз наявних множин рубрик новин на 20 існуючих сайтах новин (ukr.net, ukrinform.ua, 24tv.ua, unian.ua, korrespondent.net, gazeta.ua, pravda.com.ua, news.google.com тощо).

Нехай D – це загальна множина рубрик на всіх досліджених сайтах:

$$D = \sum_{i=1}^{k} D_i \,, \tag{1}$$

де D_i – множина рубрик *i*-го сайту новин; k – кількість сайтів новин у вибірці.

Для кожної оригінальної назви рубрики d було обраховано кількість появ у цій множині c_j, що відповідає потужності підмножини з кожною із однакових назв рубрик m:

$$c_j = \left|\left\{m \middle| m \in D \vee m = d\right\}\right| \tag{2}$$

де j – кількість оригінальних назв рубрик новин.

Виходячи з вимоги $c_j \to \max$ при обмеженні загальної кількості категорій в 6 елементів (дане число може бути змінене відповідно до потреб користувача), було визначено результуючу множину $D`$ актуальних рубрик новин:

$$D` = \{Політика, Економіка, Наука, Туризм, Спорт, Здоров'я\} \tag{3}$$

Таким чином, для подальшої роботи необхідно для кожної з цих категорій (Політика, Економіка, Наука, Туризм, Спорт, Здоров'я) визначити відповідні множини ключових слів.

Портрети новин для рубрик. Для кожної з шести визначених рубрик слід сформувати портрети новин цих рубрик у вигляді обмежених переліків ключових слів. З цією метою (рис. 2) для кожної рубрики новин необхідно сформувати вибірки з 100 випадкових новин кожна. Оскільки кінцевою метою є узагальнений пошук ключових слів за рубрикою, то ці вибірки мають бути створені як окремі текстові документи, кожен з яких містить по 100 новин.

Наступним кроком є обрахунок значень оцінки TFIDF для кожного оригінального слова відповідно до (2). При використанні методу оцінки TFIDF в якості альтернативних документів, необхідних для ідентифікації і відділення загальновживаних слів, використовуються відповідні множини зразків новин інших рубрик.

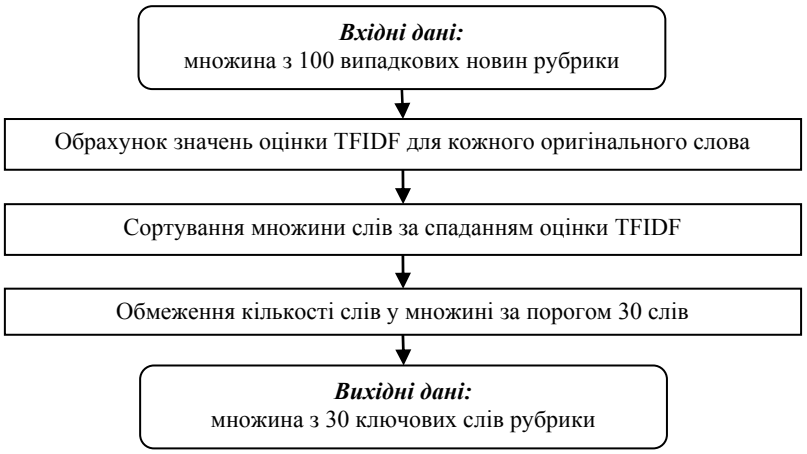

Рис. 2. Послідовність дій при визначення множин ключових слів для рубрик новин

В результаті утворюється множина, яка містить всі оригінальні слова документу й показники оцінки TFIDF для кожного слова. Після чого слова в множині сортуються за спаданням оцінки TFIDF й кількості слів у множині обмежується за порогом 30 слів.

Таким чином, в результаті виконання запропонованого підходу до визначення множин ключових слів для рубрик новин, для кожної з рубрик автоматизовано одержується множина з 30 ключових слів, яка у подальшому може бути використана як портрет новин відповідної рубрики для визначення приналежності тестової новини.

При необхідності збільшення кількості рубрик, наведені дії виконуються для кожної з них для одержання відповідних множин з 30 ключових слів.

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 з 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. 3). На цьому етапі використовуються тільки вхідні дані у вигляді навчальних множин новин для кожної з рубрик. Як було наведено вище, на цьому етапі проводиться обрахунок значень оцінки TFIDF для кожного оригінального слова для кожної з навчальних множин новин.

Після чого для кожної з одержаних множин ключових слів, що містять також значення оцінки

TFIDF, 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.

Рис. 3. Загальна схема інформаційної технології тематичного сортування текстової інформації на прикладі новин

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.

Модель для визначення приналежності новини рубрикам. 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:

$$S_{Per}^{n} = S_{Cont} \bigcap S_{Words}^{n} , \tag{4}$$

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

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 $L = \left| S_{Per}^{n} \right|$.

Процес визначення приналежності тестової новини до кожної з актуальних рубрик новин передбачає почергову перевірку наявності кожного елементу множини ключових слів даної рубрики в множині слів контенту новини; у випадку підтвердження такої наявності відбувається збільшення показника відповідності для поточної рубрики на одиницю. Процес продовжується доти, доки не будуть перевірені всі елементи множин ключових слів для кожної рубрики. Одержані значення показників відповідності і є числовими показниками приналежності тестової новини до актуальних рубрик новин. Для визначення відсоткових значень обраховується відношення показниками приналежності тестової новини до потужності множини ключових слів для кожної рубрики новин.

Отже, запропонована інформаційна технологія тематичного сортування текстової інформації дозволяє за вхідною інформацією у вигляді цифрового текстового контенту одержувати вихідну інформацію у вигляді оцінок приналежності даного контенту до кожної з відомих категорій. При цьому використовується наведений підхід до визначення множин ключових слів для рубрик новин та розроблені математико-алгоритмічні моделі для визначення приналежності тестової новини до актуальних рубрик новин.

Прикладна реалізація інформаційної технології

Для дослідження ефективності інформаційної технології тематичного сортування текстової інформації було розроблено відповідне тестове програмного забезпечення – система визначення множин ключових слів для рубрик новин й система тематичного сортування новин. Для розробки тестових зразків програмного забезпечення було використано платформу .NET Framework, мову програмування C# та формат JSON для роботи з проміжними даними. Оскільки розробка була виконана в межах одного простору імен, програмні продукти можна розглядати як складові єдиної системи.

Система визначення множин ключових слів для рубрик новин, наведена на рис. 4, дозволяє за введеною множиною новин, що мають приналежність до певної конкретної рубрики, визначити множину ключових слів, які розглядаються як еквівалент узагальненого семантичного вмісту для новин цієї рубрики. Одержана множина ключових слів сортується за значенням показника TFIDF та обмежується кількісно до 30 слів. В результаті використання програмної системи для аналізу вхідних даних у вигляді множин новин для всіх актуальних рубрик, одержуються вихідні дані у вигляді відповідної кількості множин ключових слів.

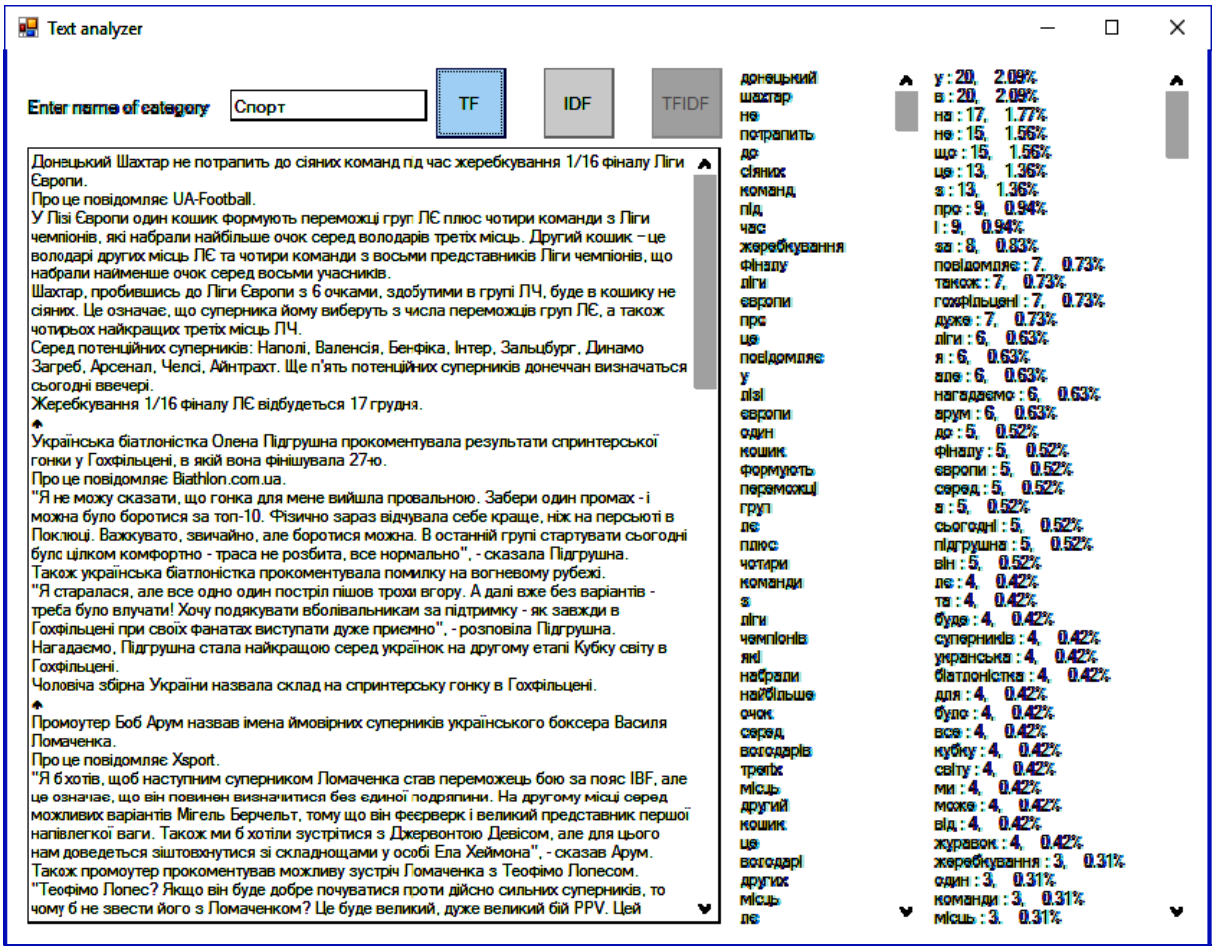

Рис. 4. Інтерфейс системи визначення множин ключових слів для рубрик новин

Система тематичного сортування новин, наведена на рис. 5, дозволяє за вхідними даними у вигляді текстового контенту тестової новини одержати вихідні дані у вигляді цифрових показників, що відображають оцінку приналежності тестової новини до кожної з рубрик. Для цього проводиться автоматизоване порівняння множини слів із контенту новини та множин ключових слів рубрик новин.

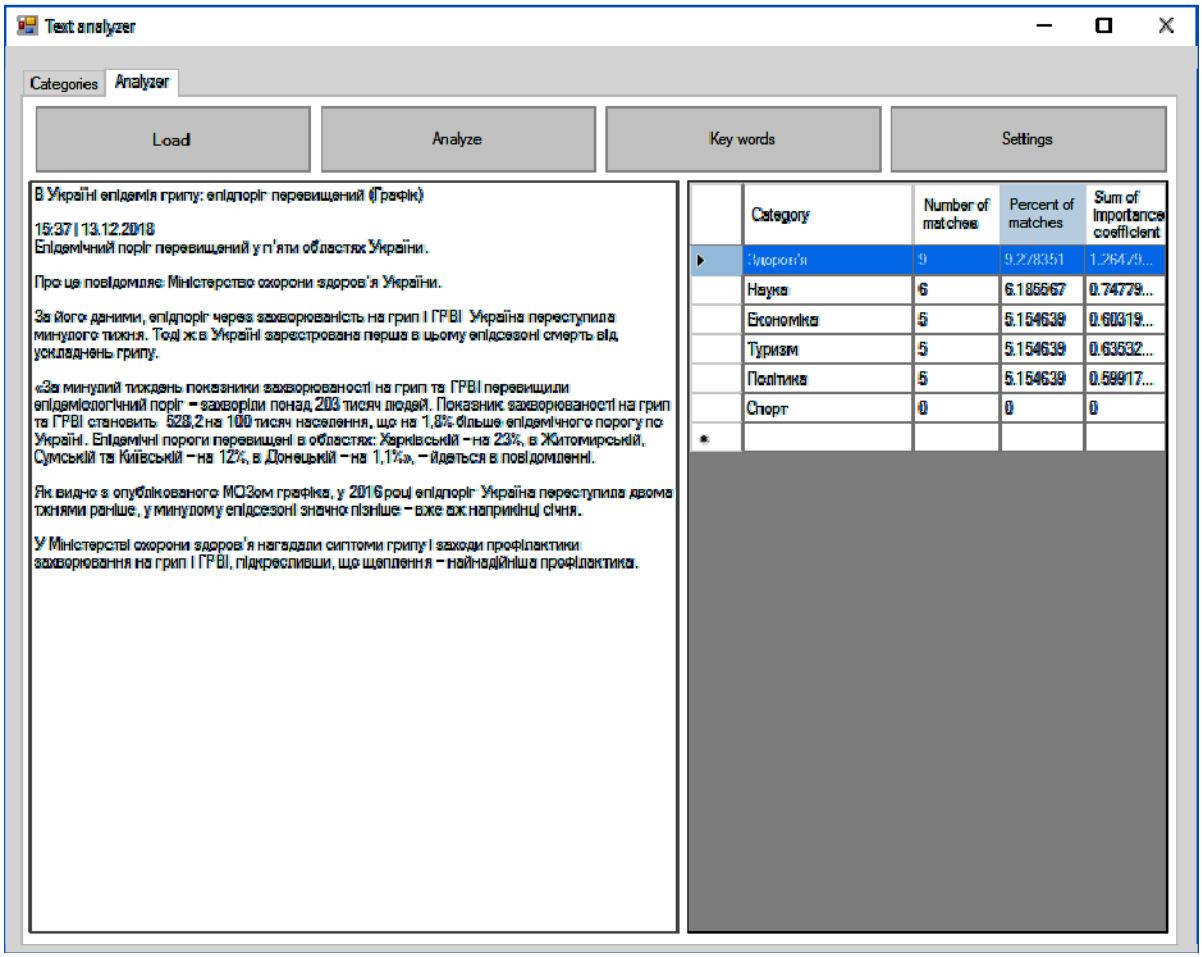

Рис. 5. Інтерфейс системи тематичного сортування новин

Розроблені тестові програмні системи дозволили провести прикладне дослідження інформаційної технології тематичного сортування текстової інформації.

Дослідження ефективності інформаційної технології

Дослідження ефективності інформаційної технології тематичного сортування текстової інформації виконано за результатами застосування розроблених тестових програмних систем, що виконані на засадах розробленої інформаційної технології. 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.

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.

В відсотковому вигляді одержано дані, подані у таблиці 2. Для визначення успішності сортування новин за рубриками U було використано відношення кількості вірних результатів до загальної кількості одержаних результатів:

$$U = \frac{T_{Ok}}{T_{All}}, \quad (5)$$

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

Таблиця 1

Кількісні результати ефективності тематичного сортування новин

Тип результату	Результати за рубриками						
	Політика	**Економіка**	**Наука**	**Туризм**	**Спорт**	**Здоров'я**	**Всього**
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

Таблиця 2

Відсоткові результати успішності тематичного сортування новин за рубриками

Політика	**Економіка**	**Наука**	**Туризм**	**Спорт**	**Здоров'я**	**Всього**
100%	86,7%	93,3%	86,7%	100%	100%	94,4%

У вигляді діаграми відсоткові результати успішності тематичного сортування новин за рубриками наведені на рис. 6.

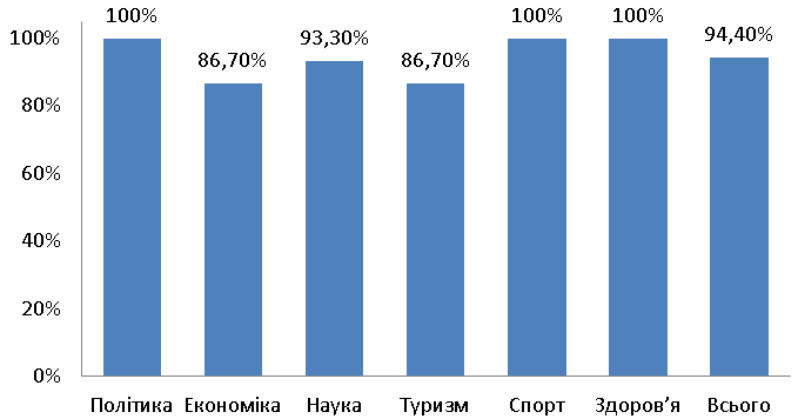

Рис. 6. Діаграма відсоткових результатів успішності тестового тематичного сортування новин за рубриками

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.

Висновки

В статті було розглянуто інформаційну технологію сортування текстових повідомлень за тематикою. При цьому використано розроблений підхід до визначення множин ключових слів для рубрик новин на основі методу оцінки TFIDF та розроблені моделі для визначення приналежності тестової новини до актуальних рубрик новин.

Одержані результати дослідження ефективності інформаційної технології показали, що в переважній більшості випадків програмна система, виконана відповідно до запропонованої інформаційної технології тематичного сортування текстової інформації, успішно виконала сортування новин за рубриками, й середня успішність сортування за рубриками склала 94,4%.

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.

Література

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. – P. 691–702.

2. RSS 2.0 Specifications. URL: http://www.rssboard.org/rss-specification

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

4. Бармак 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.

5. Krak Y. The practice implementation of the information technology for automated definition of semantic terms sets in the content of educational materials / Y. Krak, O. Barmak, O. Mazurets // CEUR Workshop Proceedings, 2139. – 2018. – P. 245–254.

References

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. – P. 691–702.

2. RSS 2.0 Specifications. URL: http://www.rssboard.org/rss-specification

3. Lande D. V. Kompaktificirovannyj gorizontalnyj graf vidimosti dlya seti slov / D. V. Lande, A. A. Snarskij // Trudy Mezhdunarodnoj nauchnoj konferencii «Intellektualnyj analiz informacii IAI-2013. Znaniya i rassuzhdeniya». – Kiev : KPI, 2013. – C. 158–164.

4. Barmak O. V. Metody avtomatyzatsii vyznachennia semantychnykh terminiv u navchalnykh materialakh / O. V. Barmak, O.V. Mazurets // Herald of Khmelnytskyi National University. Ser.: Tekhnichni nauky. – 2015. – № 2(223). – S. 209–213.

5. Krak Y. The practice implementation of the information technology for automated definition of semantic terms sets in the content of educational materials / Y. Krak, O. Barmak, O. Mazurets // CEUR Workshop Proceedings, 2139. – 2018. – P. 245–254.

Рецензія/Peer review : 19.05.2019 р. Надрукована/Printed : 23.07.2019 р.
Рецензент: д.т.н., проф. Сорокатий Р. В.