УДК 004.912

ХАТЯН Олег Анатолійович

АЛГОРИТМ ПОБУДОВИ «ТЕМ ДНЯ» ІНТЕРНЕТ ЗМІ

Постановка проблеми. Протягом останніх років має місце значне зростання обсягу новинного контенту Інтернет ЗМІ. При цьому в ритмі сьогодення можливості сприйняття інформації кінцевим споживачем досить обмежені. Тому поряд із такими характеристиками узагальненого контенту інформаційного ресурсу, як швидкість реагування на реальні події, повнота, стислість та достовірність подання не останню роль має і форма його (контенту) представлення. Іншими словами, споживач бажає оглянути перелік новин, вибрати важливе, ознайомитись із його змістом. Однак без додаткових засобів впорядкування контенту перегляд стрічки новин, що оновлюється декілька разів на хвилину, викликає значне інформаційне навантаження, навіть для фахово підготовленої особи.

Тому для підвищення швидкості та якості інформування споживача інформаційного продукту електронні ресурси пропонують структуроване подання контенту. Це відокремлені фрагменти електронних сторінок (таблоїди, переліки найбільш популярних новин) або організація сторінки (розподіл на обмежені стрічки тематичних рубрик), які містять узагальнення контенту щодо найбільш важливих та/або останніх подій. На відміну від зазначених способів подання останніх новин деякі інформаційні портали надають стислий перелік подій усього дня у вигляді переліку сюжетів новин.

З метою якісного формування/наповнення своїх інформаційних ресурсів потужні інформаційні та інформаційно-аналітичні агенції поряд із комплексами технічних засобів, зазвичай, залучають кваліфікований персонал. Інші ресурси застосовують автоматизовані системи інтеграції новин, що базуються на алгоритмах класифікації та рубрикації масивів текстових повідомлень. У будь-якому разі головною метою застосування таких алгоритмів є отримання стислого переліку груп повідомлень, кожна з яких обмежена певним *тематичним змістом*. Для позначення таких груп, отриманих протягом доби, ми використовуємо термін «теми дня».

Поряд з підвищенням якості інформування громадян щодо подій у житті держави подання контенту в стислому вигляді «тем дня» є особливо важливим для інформаційного та аналітичного забезпеченням усіх сфер дія-

льності суспільства. Саме цим зумовлена *актуальність проблеми* пошуку відповідного ефективного алгоритму побудови «тем дня».

Аналіз останніх досліджень і публікацій. Проблематиці автоматизованої побудови тематично пов'язаних інформаційних повідомлень присвячені наукові роботи багатьох дослідників. Так, російські дослідники М. С. Агеєв, Б. В. Добров, Н. В. Лукашевич висвітлюють основні проблеми та систематизують методи автоматизованої рубрикації текстів [1], описують технологію автоматичної рубрикації документів суспільно-політичної тематики (правових актів та матеріалів ЗМІ) за допомогою складних ієрархічних рубрикаторів. Зазначену технологію Б. В. Добров та Н. В. Лукашевич пропонують будувати на основі використання спеціалізованого тезаурусу (поняття та співвідношення між ними), який вживається експертами для формування правил відношення до окремої з множини наперед визначеного переліку рубрики [2]. М. С. Агеєв доповнює алгоритм методом автоматичної класифікації текстів, оснований на машинному навчанні, який використовує опис рубрик через булеві функції [3]. Подібний алгоритм рубрикації текстових документів, що містять науково-технічну інформацію українською та російською мовами, надають вітчизняні науковці О. Є. Архипов, М. В. Михайлова, В. М. Панченко, але основою відношення до певної рубрики для цього алгоритму є виділення ключових груп слів з урахуванням змінюваності словоформ текстів навчальної вибірки [4]. При побудові тематичних сюжетів та «тем дня» така апріорна інформація відсутня. Тому практично затребуваним для нас є алгоритми класифікації множин текстових повідомлень, які є основою методів пошуку схожих документів довільної вибірки. Серед них на особливу увагу, на наш погляд, заслуговує модифікація алгоритму розрахунку матриці найближчих сусідів [5] та метод, використаний Д. В. Ланде, В. М. Фурашевим, С. М. Брайчевським для встановлення дублів і пошуку подібних документів [6; 7]. Високу обчислювальну продуктивність показав підхід до кластеризації колекції документів із невідомою наперед кількістю кластерів, запропонований вітчизняними науковцями О. А. Амонсом, Ю. О. Яновим та І. О. Безпалим, базою якого є метод, оснований на статистиці появи ключових термів, модифікація методу знаходження матриці подібності на основі схожості косинуса та використання FRiS-функції [8]. Втім результатом застосування зазначених підходів є побудова груп повідомлень за *подібністю лексики*.

Отже, **метою** роботи є отримання ефективного алгоритму стислого подання контенту шляхом побудови *змістовно-смислових* груп документів – «тем дня» з масиву інформаційних повідомлень Інтернет ЗМІ. Для її досягнення визначимо термін «теми дня» та проаналізуємо рішення в межах класу

задач автоматизованої класифікації масиву (колекції документів) текстових повідомлень, що і буде нашим головним **завданням**.

Виклад основного матеріалу. Серед інших технологій інтеграції та узагальнення контенту електронних інформаційних ресурсів мережі Інтернет найефективнішими з економічної точки зору є автоматизовані системи. Саме так найбільш технологічні та потужні інформаційні портали (інтегратори новин) поряд із досить обмеженою стрічкою новин надають стрічку основних сюжетів подій. Сюжетом називають ланцюжок (або кластер, визначений на множині актуальних повідомлень) подібних документів [6; 7].

Дослідження існуючих систем інтеграції новинних потоків, отриманих із ресурсів Інтернет ЗМІ (*ukr.net*, *news.meta.ua*, *uaport.net*), показало наявність серед окремих сюжетів семантично близьких груп новин щодо однієї події, які стосуються певних її аспектів. Наприклад, стрічка сюжетів ресурсу uaport.net за 02.08.2015 р. на одній сторінці містили сюжети *«Порошенко поздравил военных Воздушных сил Украины и десанта с профессиональным праздником»* (35 повідомлень) та *«Порошенко поздравил военнослужащих Воздушных сил ВСУ и отметил их важность в боевом потенциале армии»* (28 повідомлень) (рис. 1).

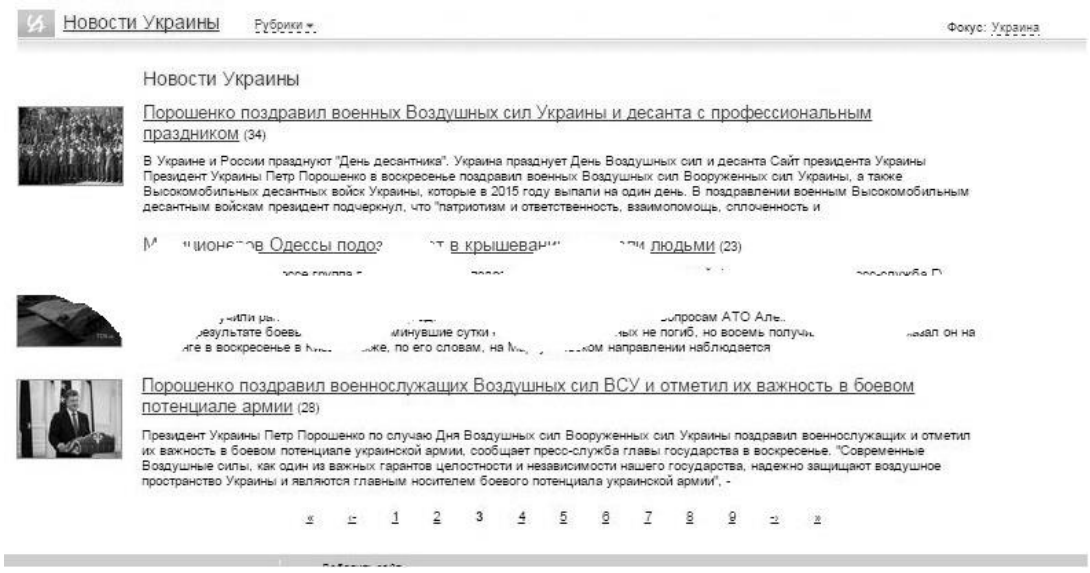

Рис. 1. Сюжети семантично близьких груп новин на uaport.net за 02.08.2015 р.

Роз'яснення ситуації ми отримали шляхом дослідження відповідних баз – методичної та алгоритмічної [6; 7], застосованих засобів автоматизації інтеграції новин. З'ясувалось, що основою побудови сюжетів потоку новин є алгоритми класифікації вихідної колекції інформаційних повідомлень за

векторами значимих термів. Отже, такі алгоритми, по-перше, обмежуються аналізом лексики, а по-друге, з метою прискорення роботи алгоритму класифікації, що доцільно з огляду потужності множини повідомлень, які підлягають обробці, обмеженню підлягає кількість лексичних одиниць (значимих термів), які складають характеристичний вектор документа. Таким чином, сюжетний ланцюг – є статистичною похідною лексичного вмісту інформаційних повідомлень. При цьому гарантованість відображення вектором значимих термів змісту та смислу повідомлення залишається відкритою темою для досліджень.

З метою концептуального вирішення зазначеної проблеми пропонується визначити поняття *«тема дня»* як таке, що поєднує протягом певного періоду (доби) у тематично пов'язану групу повідомлення, які подають за змістом: фактичний опис події; розгортка події стосовно різних контекстів; розвиток події; оцінка події; аналіз події та її наслідків.

За побудовою така група повідомлень повинна враховувати як лексичні, так і змістовно-смислові ознаки документів. Формування таких тем, на нашу думку, доцільно здійснювати на метарівні стосовно аналізу лексики окремого повідомлення в межах певної колекції документів, але з урахуванням результатів цього аналізу.

Отже, розглянемо більш детально певні рішення щодо автоматизованої класифікації масиву (колекції документів) текстових повідомлень.

Насамперед зазначимо, що з точки зору загального підходу функціонально усі відомі рішення складаються з чотирьох етапів і полягають у наступному:

1. Перехід від множини документів до множини векторів значимих термів (множина таких термів вихідної колекції документів задають простір ознак).

2. Розрахунок ваги кожного терму у кожному документі.

3. Побудова матриці близькості між документами.

4. Виконання процедури кластеризації.

Деякі автори перші два пункти об'єднують, хоча, на нашу думку, це обумовлено певними технологічними рішеннями.

Відповідно існують сталі методи розв'язання окремих кроків, що вважаються стандартними.

Так, оцінку ваги певного терму l в документі стандартно [5; 6; 7] розраховують за формулою TF*IDF:

$$TFIDF_D(l) = \beta + (1-\beta) \cdot tf_d(l) \cdot idf_d(l), \qquad (1)$$

$$tf_d(l) = \frac{freq_d(l)}{freq_d(l) + 0{,}5 + 1{,}5 \cdot \dfrac{dl_D}{avg\_dl}},$$

$$idf_d(l) = \frac{\log\left(\dfrac{|c| + 0{,}5}{df_l}\right)}{\log(|c| + 1)},$$

де $tf_D(l)$ – частотний коефіцієнт терму в документі D, $freq_D(l)$ – частота терму l в документі D, dl_D – міра довжини документа D, avg_dl – середня довжина документа, $\beta = 0{,}4$, $idf(l)$ – зворотній частотний коефіцієнт терму в документі D, $|c|$ – кількість документів у колекції, $df(l)$ – кількість документів, де присутній терм l.

Також стандартною вважається оцінка міри наближення документів $x \in D_1$ та $y \in D_2$ і розраховується як косинус між векторами документів у просторі ознак [5; 8]:

$$\cos(x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}, \qquad (2)$$

де n – вимірність простору ознак (кількість термів у колекції), x_i, y_i – TF*IDF-вага i-тої ознаки документа x та y відповідно.

Для проведення процедури кластеризації, зазвичай, використовують алгоритм K найближчих сусідів [5; 6; 7; 8] на основі матриці близькості між документами, суть якого полягає у тому, що певний об'єкт l відносять до того класу, елементів (ознак) якого виявиться більше серед k найближчих сусідів $x_l^{(i)}$, $i=1, \ldots, k$ [10]:

$$w(i,l) = [i \le k]; \qquad a(l, X^m, k) = \arg\max_{y \in Y} \sum_{i=1}^{k} \left[y_l^i = y\right], \qquad (3)$$

де X^m – метрична множина документів, для якої визначені ваги $w(I,k)$ l-го документа; $y \in Y$ – шукане розбиття множини X^m на класи за допомогою алгоритму $a : X \to Y$.

При цьому оптимальне значення параметра k визначають за критерієм «з виключенням об'єктів по-одному» (*leave-one-out*). Для кожного об'єкта

$x_i \in X^m$ перевіряється, чи вірно він класифікується за своїми k найближчими сусідами.

$$LOO(k, X^m) = \sum_{i=1}^{m} \left[a\left(x_i; X^m \setminus \{x_i\}, k\right) \neq y \right] \to \min_k . \qquad (4)$$

Не зупиняючись на інших поширених методах класифікації множини документів, описаних в науковій літературі, розглянемо особливості окремих рішень класичного підходу.

Так, в [5] автори пропонують підвищення продуктивності алгоритму розрахунку *K* найближчих сусідів завдяки організації обчислювальної системи, а саме застосування інвертованих індексів таблиці бази даних, що містить записи «документ-терм-вага», впорядкованих за парою (індекс-документа, індекс-терму) і з оцінкою ваги терму як *cos(x,y)* за виразом (2).

У монографіях [6; 7] для побудови векторів значимих термів пропонується використовувати фіксовано 12 найбільш вагових з лінгвістичної точки зору термів, що пройшли процедуру морфологічної обробки; термами автори називають ключові слова, відібрані за законом Зіпфа, тобто такі, що не є словами зі «стоп-словника» та не є рідкими.

Підхід, запропонований авторами у [8], відзначається цілою низкою удосконалень класичного рішення. Так, на етапі зважування термів у просторі ознак, замість обчислення TF*IDF-ваги (1), кожному терму t_i документа T_l ставиться у відповідність статистична міра w_{li}:

$$w_{li} = -\log(p(t_i) \cdot f_t(T_l, t_i)), \qquad (5)$$

$$f_t(T_l, t_i) = \frac{f_r(T_l, t_i)}{f_r(T_l, t_i) + 1 + \frac{d(T_l)}{350}},$$

$$p(t_i) = 1 - e^{-1.5\frac{f_c(t_i)}{n}},$$

де $f_r(T_l, t_i)$ – число входжень терма t_i в документ T_l; $d(T_l)$ – кількість термів у документі T_l; $f_c(t_i)$ – число входжень терма t_i в колекцію.

На етапі побудови матриці близькості науковці пропонують удосконалення функції схожості косинуса (2) з метою підвищення точності оцінки міри зв'язку текстів у випадку невеликої кількості термів, що входять в обидва оцінювані документи T_1 та T_2 за рахунок ваги термів ($w1_{1i}$, $w2_{2i}$), присутніх у двох текстах:

$$sim(T_1, T_2) = \frac{\sum_i w_{1i} w_{2i}}{\sqrt{\sum_i w_{1i}^2 \sum_i w_{2i}^2}} \cdot \sum_i \frac{w1_{1i} + w2_{2i}}{w1_{1i} + w2_{2i}} . \quad (6)$$

Кластеризація, на відміну від застосування алгоритму *K* найближчих сусідів (3), виконується на основі модифікованої FRiS-функції [9], яка є мірою подібності об'єкта *T* зі стовпом b_i у конкуренції з іншими стовпами:

$$F^*(T, B) = \left(\frac{r1(T, b) - r2^*(T, b)}{r1(T, b) + r2^*(T, b)} \right). \quad (7)$$

Середнє значення FRiS-функції для вибірки дозволяє розрахувати характеристику повноти набору стовпців, характерних для колекції повідомлень:

$$F(B) = (1/m) \sum_{T_i \in T} F^*(T_i, b) . \quad (8)$$

Втім цей метод, як і інші розглянуті, дозволяє будувати з практичної точки зору лише сюжетні ланцюжки повідомлень, які характеризуються подібним набором ключових термів. Врахування ж змістовно-смислових ознак документів в зазначеній роботі, як і в розглянутих вище, не вирішується.

Наш підхід до формування «тем дня» також базується на методі класифікації колекції документів, отриманих протягом доби, з певними відмінностями.

Так, з метою підвищення сукупної продуктивності методу ми пропонуємо технологічно розділити етапи класичної схеми на дві групи:

1. «Перехід від множини документів до множини векторів значимих термів» і «Розрахунок ваги кожного терму в кожному документі».

2. «Побудова матриці близькості між документами» і «Виконання процедури класифікації».

Такий розподіл, на нашу думку, дозволить виконувати обчислення для першої групи етапів у процесі поточного збору інформації, а для другої групи – безпосередньо при виконанні запиту на класифікацію певної колекції документів.

Для побудови *вектора ключових термів* автор вважає за доцільне модифікувати алгоритм, описаний в [5], де ключовими словами називають такі терми l_i, для яких частота входжень у текст більша за певне граничне значення:

$$N(l_i) > \alpha , \quad (9)$$

де α залежить від розміру вихідного тексту, типу документа і визначається експериментально як стала (для науково-технічних документів $\alpha = 10$) [5].

Ми пропонуємо як граничне значення використовувати визначену нами оцінку часткової інформативної потужності *K* повідомлення стосовно кількості унікальних слів повідомлення без слів із словника «стоп-лист» – S_{nS} [11]:

$$K = \frac{S_{nS} \times 2}{100} + \sigma ,$$

де σ – статистичне відхилення щільності групового розподілу термів l_i повідомлення. Тоді вираз (9) набуває вигляду:

$$N(l_i) > K . \qquad (10)$$

Таким чином, зазначена модифікація алгоритму переходу від документа до вектора ключових термів, на нашу думку, є певним еквівалентом оцінки ваги терму *l* в документі за формулою TF*IDF (1), що водночас дозволяє зменшити обчислювальну складність етапу класифікації колекції повідомлень щодо відповідного запиту.

Реалізація запропонованої нами процедури побудови «тем дня» виконується шляхом застосування дворівневої кластеризації колекції повідомлень, відібраних протягом доби. Результатом першого рівня є побудова множини *сюжетів повідомлень* на основі ключових термів повідомлень (різновид методу кластеризації *K* найближчих сусідів). Полягає він у наступному.

Для кожного повідомлення m_i відібраної для аналізу колекції $M = \{m_1, m_2, \text{K}, m_i, \text{K}, m_n\}$ формуються, відповідно до критерію (10), вектори ключових термів, які інтегрально утворюють простір ознак (*загальний вектор ключових термів*) $K = \{k_1, k_2, \text{K}, k_j, \text{K}, k_m\}$ вихідної колекції повідомлень, і будується матриця зв'язків «повідомлення-ключові терми»:

$$D_{MK} = \begin{pmatrix} w_{11} & w_{12} & \Lambda & w_{1n} \\ w_{21} & w_{22} & \Lambda & w_{2n} \\ \Lambda & \Lambda & \Lambda & \Lambda \\ w_{m1} & w_{m2} & \Lambda & w_{mn} \end{pmatrix}, \qquad (11)$$

де $w_{ij}, \left(i = \overline{1,n};\ j = \overline{1,m}\right)$ – частота ключового терму *i* у повідомленні *j*. Після цього розраховується коефіцієнт кореляції повідомлень за ключовими термами (матриця векторного добутку $D_{M\,M}$), що є еквівалентом матриці близькості між документами:

$$D_{M\,M} = \begin{pmatrix} \overline{w}_{11} & \overline{w}_{12} & \Lambda & \overline{w}_{1n} \\ \overline{w}_{21} & \overline{w}_{22} & \Lambda & \overline{w}_{2n} \\ \Lambda & \Lambda & \Lambda & \Lambda \\ \overline{w}_{n1} & \overline{w}_{n2} & \Lambda & \overline{w}_{nn} \end{pmatrix}, \tag{12}$$

по рядкам матриці D_{MK}, за формулою, яка фактично забезпечує порівняння кожному терму t_i документа T_l його статистичної міри $\overline{w}_{ij}$:

$$\overline{w}_{ij} = \left. \frac{\left(\sum_{j=1}^{m} w_{ij} \right)}{\lambda_i} \right|_{i = \overline{1,n}}, \tag{13}$$

де $\lambda_i = \max(w_{ij}), \left(j = \overline{1,m}\right)$ – максимальне значення входження *i*-го ключового слова серед усіх *j*-тих повідомлень.

Тоді остаточне розбиття множини *M* повідомлень за класами $\Omega = \{\omega_1, \omega_2, \text{K}, \omega_l\}$ (де $\omega_i = m_{\omega i}$ – є повідомлення-центроїд, на роль якого послідовно тестуються повідомлення із множини *M*) відбувається за правилом: $m_i \in \omega_k, \left(k = \overline{1,l}\right) |_{\overline{w}_{ik} - \overline{w}_{\omega k} < \alpha}$, де $\overline{w}_{\omega k}$ – коефіцієнт кореляції центроїда, а α – граничне значення відстані повідомлення від центроїду класу або коефіцієнт подібності повідомлення до певного класу, тобто *сюжетного ланцюжка*.

Подальші міркування стосовно формування «*теми дня*» виходять з ідеї, що отримані з колекції документів сюжети фактично дають перелік описів подій та оцінок щодо окремих фактів та об'єктів реального світу. З іншого боку, окреме повідомлення змістовно стосується ширшого семантичного кола аспектів у певних контекстах ніж ті, з яких лексично складається сюжет. Цим, зокрема, пояснюється наведений нами на початку роботи приклад про два сюжети стосовно теми поздоровлення Президентом України ВПС України. А тому, на нашу думку, класифікація сюжетів дозволить отримати шукане змістовно-смислове групування повідомлень.

Отже, на другому рівні *сюжетні ланцюжки* об'єднуються в «***теми дня»*** на основі кластеризації повідомлень, які увійшли до різних сюжетних ланцюжків (процедура побудови аналогічна наведеній вище, але простором ознак виступають ідентифікатори повідомлень). Назва теми визначається назвою повідомлення-центроїду, яке об'єднало інші повідомлення кластера.

Практична реалізація

Описаний нами алгоритм формування «тем дня» розраховано для стислого подання потоку новин комплексу програмних засобів Системи аналізу інформаційного простору (САІП). Цей комплекс створювався у рамках проведення окремих науково-дослідних робіт та будувався нами за модульним принципом із використанням наступних freeware засобів та пакетів: БД – MySQL; веб-сервер – Apache; засоби програмування – веб-сервер: PHP; веб-клієнт: JavaScript, AJAX; додаткові модулі збору повідомлень Інтернет ЗМІ та засоби побудови лінгво-статистичних індексів: MS Visual C/C++; додаткові алгоритми та засоби: mysqlclient.lib – API бібліотека доступу до БД MySQL; алгоритм стемменгу Портера, адаптований для української і російської мов та реалізований мовою програмування PHP. Словники напрацьовані групою експертів: «стоп-слів», фактографічної (подієвої) лексики, позитивної/негативної лексики, маніпулятивної лексики, емоційної лексики.

Реалізація та дослідне тестування запропонованого нами алгоритму проводилось на обчислювальній системі на базі Intel(R) Core(R) i5-3450S CPU @ 2,8GHz 2,8GHz 3Gb RAM.

Результати тестування

Тестування алгоритму формування «тем дня» проводили на масиві повідомлень, зібраних САІП за добу 03.08.2015 року за допомогою веб-клієнт – АРМ аналітика для встановлення обчислювальної продуктивності та ефективності стиснення контенту. Інформаційні повідомлення були відібрані із 115 найбільш популярних джерел українського сегмента Інтернет ЗМІ. Максимальна для обчислення колекція документів становила 3600 повідомлень, при чому алгоритм працював 98,9 сек. і сформував 192 «теми дня» (801 сюжет), що, за нашими оцінками, дуже добре для швидкого інформування стосовно подій, які висвітлювались Інтернет ЗМІ протягом доби (рис. 2).

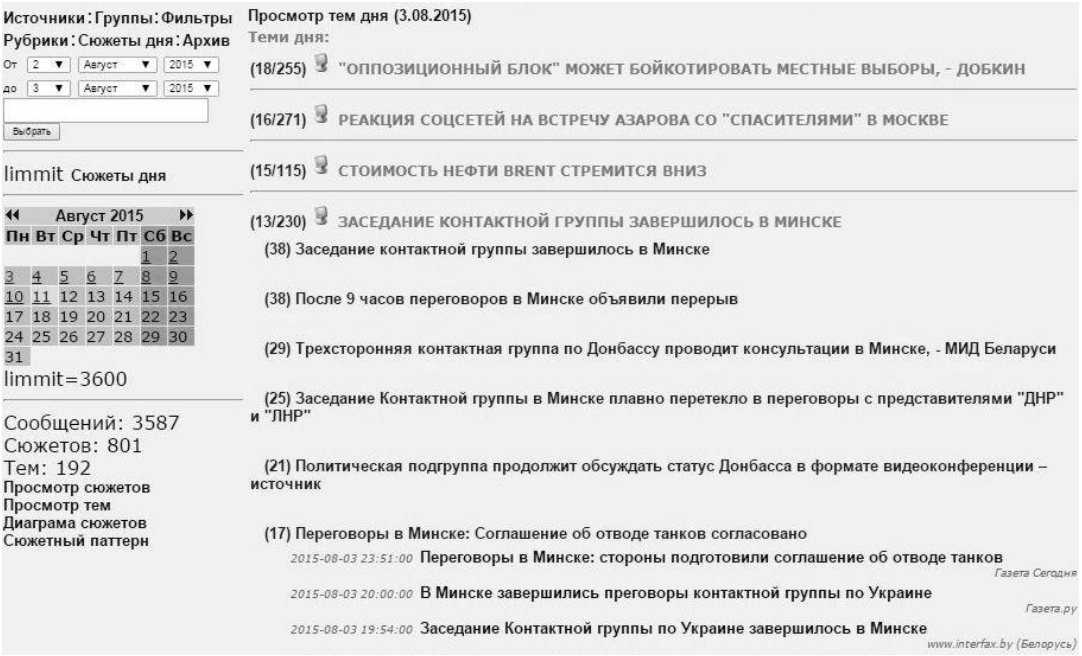

Рис. 2. Фрагмент списку «тем дня», ранжируваних за кількістю групованих сюжетів, сформованих за 03.08.2015 р.

Результати тестування продуктивності алгоритму для різних колекцій документів представлені на діаграмі (рис. 3).

Рис. 3. Діаграма залежності швидкості роботи алгоритму формування «тем дня» від потужності вхідної колекції (крок – 100 документів)

Колекція документів із 1600 повідомлень була опрацьована нашим алгоритмом за 19,9 сек., що у порівнянні з розглянутими у [8] алгоритмами Claster (6 сек.) та FRiSClaster (44 сек.) є порівнянним з відомими алгоритмами.

Для дослідження ефективності стиснення контенту ми розраховували коефіцієнт стиснення інформації:

$$k = \frac{S_0}{S_c},$$

де S_0 – потужність вихідної колекції документів; S_c – кількості отриманих при роботі алгоритму сюжетів або «тем дня». Більше значення коефіцієнта характеризує більший ступінь стиснення. Розрахунки виконували окремо для результатів обробки алгоритмами побудови сюжетів та «тем дня» щоденної колекції повідомлень протягом липня 2015 року (рис. 4).

Рис. 4. Діаграма тестування ефективності стиснення контенту алгоритму формування «тем дня»

Отримали середнє значення коефіцієнта стиснення для сюжетів 4,5, в той час як для «тем дня» 18,4, що свідчить про більше ніж чотирикратну ефективність стиснення вихідної множини повідомлень. Крім того, отримані нами «теми дня» є категорією більш інформативною, яка до того ж характеризує смислове навантаження декількох подібних сюжетів, що у інших розглянутих підходах не вирішується.

Зауважимо, що отриманий за нашою процедурою перелік «тем дня» змінюється протягом доби і стає сталим тільки після її закінчення. Така особливість дає змогу у прикладних дослідженнях проводити ретроспективний аналіз контенту за цією ознакою.

Висновки та перспективи дослідження. Таким чином, детальне ознайомлення з існуючими системами інтеграції новинних потоків, отриманих з ресурсів Інтернет ЗМІ з метою пошуку найбільш ефективного та стислого

способу подання значного обсягу інформації зі збереженням тематичного змісту, дозволило нам визначити «теми дня» як конструкцію, що враховує лексичні та змістовно-смислові ознаки групи пов'язаних документів.

Проведений нами аналіз стандартної схеми, сучасних підходів, рішень та окремих алгоритмів автоматизованої класифікації колекції текстових повідомлень дозволив завдяки певним модифікаціям запропонувати алгоритм формування «тем дня».

Тестування практичної реалізації запропонованого алгоритму не тільки показало достатність обчислювальної потужності для впровадження онлайнового веб-додатку, але й засвідчило його ефективність та інформативність для стислого відображення змісту потужної колекції вихідних документів.

У подальшому спрямуємо наші дослідження на вивчення можливостей прикладних аспектів застосування «тем дня», зокрема з метою виявлення інформаційних впливів через Інтернет ЗМІ.

Список використаних джерел

1. Агеев М. С. Автоматическая рубрикация текстов: методы и проблемы [Электронный ресурс] / М. С. Агеев, Б. В. Добров, Н. В. Лукашевич // Ученые записки Казанского государственного университета. Серия Физико-математические науки. – 2008. – Т. 150, кн. 4. – С. 25–40. – Режим доступа : http://www.cir.ru/docs/ips/publications/2008_kgu_classif.pdf.

2. Добров Б. В. Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры [Электронный ресурс] / Б. В. Добров, Н.В. Лукашевич // VIII Нац. конф. по искусственному интеллекту КИИ-2002. – М. : Физматлит, 2002. – Т. 1. – С. 178–186. – Режим доступа : http://www. cir. ru/docs/ips/publications/2002_cai_rubr.pdf.

3. Агеев М. С. Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов : дис. канд. физ-мат. наук : 05.13.11 / М. С. Агеев ; Московский гос. унив. – М., 2005.

4. Архипов О. Є. Алгоритм автоматизованої рубрикації текстових документів / О. Є. Архипов, М. В. Михайлова, В. М. Панченко // Інформаційна безпека людини, суспільства, держави. – 2011. – № 1(5). – С. 41–51.

5. Агеев М. С. Метод эффективного расчета матрицы ближайших соседей для полнотекстовых документов [Электронный ресурс] / Б. В. Агеев, Б. В. Добров // Вестник Санкт-Петербургского университета. – 2011. – Сер. 10. – Вып. 3. – С. 72–84. – Режим доступа : http://www.cir.ru/docs/ips/publications/2011_vspb_knn.pdf.

6. Ландэ Д. В. Основы интеграции информационных потоков : моногр. / Д. В. Ландэ. – К. : Инжиниринг, 2006. – 240 с.

7. Ландэ Д. В. Основы моделирования и оценки электронных информационных потоков : моногр. / Д. В. Ландэ, В. Н. Фурашев, С. М. Брайчевский, А. Н. Григорьев. – К. : Инжиниринг, 2006. – 176 с.

8. Амонс О. А. Кластеризація документів на основі статистичної близькості термів / О. А. Амонс, Ю. О. Янов, І. О. Безпалий // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка : зб. наук. праць. – 2008. – № 49. – С. 55–62. – Бібліогр.: 11 назв.

9. Борисова И. А. Использование FRiS-функции для построения решающего правила и выбора признаков / И. А. Борисова, В. В. Дюбанов, Н. Г. Загоруйко, О. А. Кутненко // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–07). – Новосибирск, 2007. – Т. 2. – С. 67–76.

10. Воронцов К. В. Математические методы обучения по прецедентам. Курс лекций по машинному обучению [Электронный ресурс] / К. В. Воронцов. – 2011. – 141 с. – Режим доступа : http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf.

11. Хатян О. А. Потужність множини ключових слів як критерій інформативності новинного повідомлення / О. А. Хатян, М. О. Рябий // Актуальні питання забезпечення кібернетичної безпеки та захисту інформації : зб. наук. праць наук.-практ. конф. ; м. Київ, 25–28 лютого 2015 р., Європейський університет / редкол. : О. І. Тимошенко (голова) та ін. – К. : Вид-во Європейського університету, 2015. – С. 117–119.

Рецензенти:
кандидат технічних наук, доцент
В. Рябцев,
кандидат технічних наук, старший
науковий співробітник В. Панченко

Аннотация: В статье дается определение «тем дня» как конструкции, которая учитывает лексические и содержательно-смысловые признаки группы связанных документов. Предложены некоторые модификации общего алгоритма, а также отдельных составляющих классификации множества сообщений, полученных из ресурсов Интернет СМИ, и алгоритм построения «тем дня». Приведенные результаты тестирования предложенного алгоритма показывают его информативность и эффективность для сжатого отображения содержания многочисленной коллекции исходных документов.

Ключевые слова: тема дня, алгоритм классификации, сюжеты новостей, сжатое содержание, Интернет СМИ.

Abstract: The article defines «pick-of-the-day» as construction that takes into account the lexical and meaningful-significative signs of the related documents group. Some modifications of general algorithm, separate constituents of set of messages classification received from the Internet MASS-MEDIA as well as the algorithm of the construction of the «pick-of-the-day» are suggested. Testing results of the proposed algorithm show its informational content and efficiency for the compressed reflection of powerful collection of initial documents.

Key words: pick-of-the-day, classification algorithm, news plots, compressed content, Internet MASS-MEDIA.