УДК 004.023

Прищепа С.В.
Інститут проблем реєстрації інформації
Національної академії наук України

ВИЯВЛЕННЯ НОВИХ ПОДІЙ НА ОСНОВІ РЕЙТИНГУВАННЯ ДЖЕРЕЛ У TWITTER

У статті вивчаються методи виявлення нових подій з інформаційного потоку глобальної мережі Інтернет. Пропонується підхід до виявлення нових подій із використанням методу рейтингування джерел інформації всередині соціальної мережі Twitter. Запропонований метод дає можливість зменшити витрати на моніторинг джерел і необхідні потужності для оброблення і виявлення нових подій. А перерозподіл рейтингу джерел збільшує повнтоту і точність вибірки із плином часу. Для традиційних методів екстрагування нових подій характерні проблеми швидкості виявлення і необхідність у накопиченні й обробленні надвеликих масивів даних. Цю проблему вирішує створення і моніторинг найбільш рейтингових джерел соціальної мережі з найбільшою кількістю нових подій щодо всієї кількості повідомлень.

Ключові слова: *екстрагування подій, рейтингування джерел, обробка тексту, Twitter, інформаційний потік.*

Постановка проблеми. В одній з найбільших соціальних мереж світу – Twitter – щомиті публікуються тисячі матеріалів. Деякі з них цікаві з погляду інформаційності. Але через величезну кількість авторів доводитися працювати з надто великою кількістю даних, що уповільнює й ускладнює виявлення в інформаційному потоці цінних повідомлень (твітів), які потенційно можуть бути новими, невиявленими раніше подіями. Оскільки твіти мають коротку форму, а користувачам не потрібно структурувати текст та стиль повідомлення і чекати модерації, то вони мають перевагу в швидкості поширення й оновлення актуальної інформації. Також варто зазначити, що локальні події, незначні в масштабі країни, але значущі для окремих невеликих локацій та їхніх жителів, теж активно поширюються через дану соціальну мережу, на відміну від засобів масової інформації (далі – ЗМІ). Саме тому для розроблення моделі вилучення нових подій з інформаційних потоків обрана саме ця соціальна мережа. Мінусами можна вважати велику кількість шуму і неправдивої інформації в даному інформаційному потоці.

Аналіз останніх досліджень і публікацій. Приклади якісного виявлення подій описувались нами й іншими вченими в інших статтях. Науковці з усього світу працюють над створенням і вдосконаленням систем виявлення подій, тематики, трендів, емоційності повідомлень з інформаційних потоків. Приклади охоплюють як підходи на основі правил, так і різні статистичні підходи.

Наприклад, система екстрагування подій ATT1, що досягла найкращих результатів на TempEval-3 2013 р., базується як на семантиці текстів, так і на синтаксичному розборі їх. Система виявлення подій ATT1 більше спирається на лексичні, ніж на семантичні особливості в тексті. Вона працює на підході послідовного позначення (sequence labeling) [1, с. 20–24].

В описі свого методу та системи виявлення подій "Niagarino" Андреас Вейлер та інші вчені зазначають, що після попереднього оброблення інформаційного потоку Tweeter призначає методом зворотної частоти документа кожному слову спеціальну оцінку за весь час, а потім найбільш топові терми використовуються для побудови зв'язку, терміни навколо цих термів теж беруться до уваги та мають свою оцінку за IDF-методом. Далі вчені використовують LDA-метод, незважаючи на те, що він зазвичай застосовується для присвоєння категорій, але тут його використовують для визначення – подія / не подія [2, с. 38–40].

Х. Бекер та інші вчені запропонували 2011 р. свій метод виявлення на основі технік кластеризації. Їхній підхід полягає в безперервній кластеризації схожих твітів та розподілі їхнього контенту на «події» та «не події» завдяки методу опорних векторів [3, с. 438–441].

Постановка завдання. У статті ми пропонуємо метод виявлення нових подій зі скороченої кількості джерел, що забезпечує зменшення часу / витрат на виявлення нових подій у потоці повідомлень Twitter.

На рис. 1 зображено графік появи документів у категорії дорожньо-транспортних пригод (далі – ДТП) з наявними в них тригерами подій. На рис. 2 темними кругами виділені нові виявлені події.

Теорія методу полягає в тому, що в мережі Twitter є певний список найбільш рейтингових (подієвих) джерел: акаунти новинних сайтів, ЗМІ, національних лідерів, активістів тощо. Багато з їхніх постів – важливі нові події. Отже, сформувавши список таких джерел (авторів) і оновлюючи його з деякою періодичністю, можна досягти результату, коли 20% локальних джерел охоплюють 80–95% цінних нових подій, що публікуються в мережі. Такий розподіл дає можливість моніторити і виявляти нові події тільки в їхніх твітах, в отже, зменшує скановану базу майже на 80% джерел. Наша попередня експериментальна оцінка показує, що такий підхід працює та підвищує швидкість виявлення і зменшує ресурси, необхідні для сканування.

Виклад основного матеріалу дослідження. Для повноцінної роботи за таким методом спочатку необхідно сформувати список ключових запитів до Twitter за визначеною категорією, наприклад, «Кібербезпека». Сформувавши список основних запитів (понад 100 запитів) за заданою тематикою, ми зробили вивантаження всіх твітів за 1 рік, в яких містилися ключові запити щодо такої тематики та їхніх авторів (джерел). Маючи таку базу даних про твіти, їхній вміст і джерела, ми намагаємося екстрагувати тільки ті твіти, які є подіями. Подія – значна подія певної категорії, явище чи інша діяльність як факт суспільного або особистого життя. Нас цікавлять події, де є хоча б один фігурант, що її вчинив, – суб'єкт, або фігурант, щодо якого її вчинили – об'єкт події. Ми розглядаємо задачу виявлення подій як екстрагування індикаторів (тригерів) подій заданих типів і виявлення їхніх аргументів, зв'язку з фігурантами в реченні. Потім відібрана інформація про ці події має бути розпізнана й об'єднана в єдине уявлення для кожної виявленої події конкретної категорії.

Оскільки більшість подій мають контекстно залежне відношення і більшість із них згадуються лише в декількох документах, написаних протягом якогось часового інтервалу, то використовуємо спеціальні шаблони і словники. Для кожної категорії створюємо свій шаблон поведінки і словники видобутку подій певної категорії. Присвоєння певної категорії документу здійснюється методом опорних векторів. Далі розбиваємо тексти на ключові слова і фрази. Кожному ключовому слову присвоюємо значення для даної предметної області за методом TF-IDF.

Для виявлення подій, їхніх суб'єктів і об'єктів, ми створюємо спеціальні шаблони правил розбору пропозицій із певної тематики і словники індикаторів подій із певної теми, які заповнюються з експертом з урахуванням різних лінгвістичних ознак

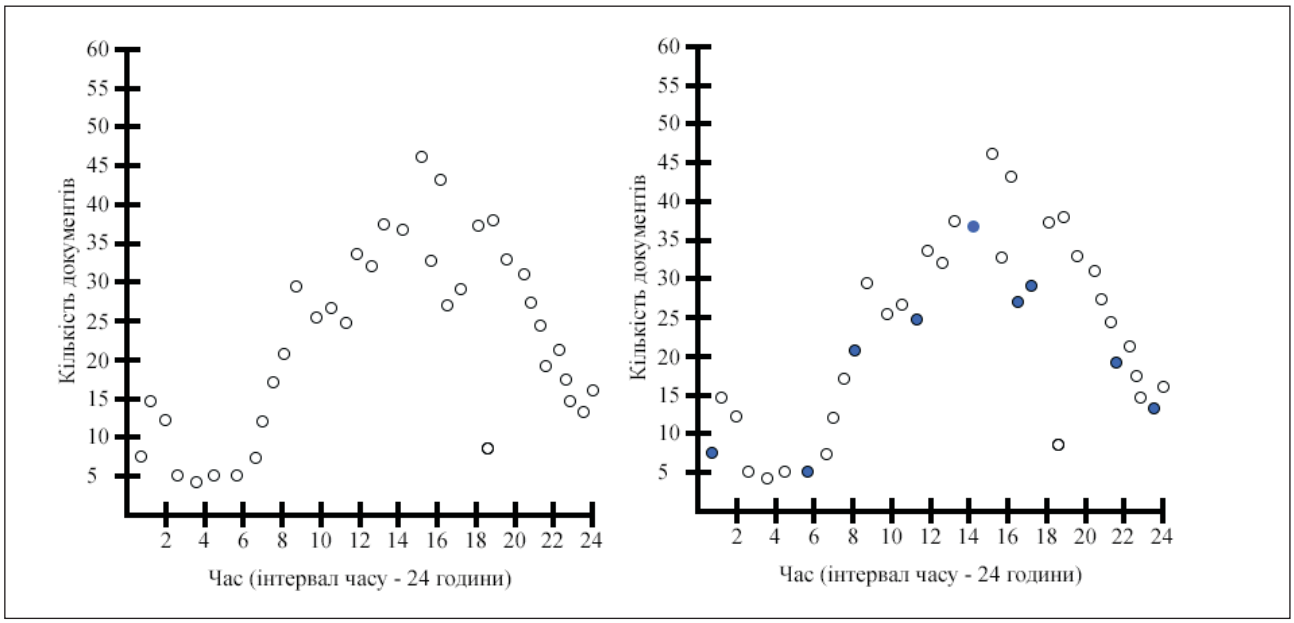

Рис. 1. Поява документів за темою ДТП

Рис. 2. Нові події в категорії ДТП

Таблиця 1

Приклад частини таблиці рейтингу джерел за тематикою

iD	Назва-посилання	Загальна кількість твітів за N часу	Кількість твітів із подією за N часу	Кількість твітів із новою подією за N часу	% ефективності
237	@poroshenko	89	61	49	68,539
238	@interfaxua	9 720	1 276	576	13,127
239	@APUkraine	630	158	114	25,079
240	@CrimeUA	990	489	221	49,393
241	@news112ua	7 863	1 298	648	16,507
242	@kommersant_ua	8 100	1 174	583	14,493
243	@rbc_ukraine	4 500	918	329	20,4

теми (категорії). Індикатори тематичної події представлені словниками. Використовується не один словник індикаторів подій із певної теми, а два. Індикатори події виділяємо пошуком індикаторів у пропозиціях документа з одного зі словників присвоєної документу категорії. Використання подвійних словників дозволяє надати індикатору події спеціальний аргумент приналежності до одного зі словників, що дозволяє будувати більш складні шаблони умов, вивести показники точності вилучення подій на більш високий рівень.

Ми підходимо до питання вилучення подій як до виявлення індикаторів (тригерів) події заданих типів і виявлення їхніх аргументів, зв'язку з фігурантами в реченні. Потім відібрана інформація про ці події розпізнається й об'єднується в єдиному поданні для кожної виявленої події.

Визначимо подію e як кортеж (сутність і дата) + згадування про e, me – може бути твіт, що містить посилання на сутность і написаний у специфічну конкретну дату. Особливості події екстрагуємо зі згадок про подію:

$$x_e = f(\{m_{e'} | e' = e\}),$$

які можуть бути використані для оцінки ймовірності того, що подія належить до категорії E, згідно з деякими заданими параметрами для категорії θE:

$$p_{\theta^E}(y_e = 1 | x_e) = \frac{1}{1 + e^{-\theta_E \cdot x_e}}.$$

Якщо в реченні знайдений індикатор події і хоча б один фігурант, то здійснюється подальший розбір речення – нумерація порядку всіх слів у реченні і присвоєння кожному слову свого порядкового номера. Коли всім сутностям присвоєні типи, аргументи і порядковий номер у реченні, то підбирається тип шаблону, якрму відповідає дане речення з подією. Варіанти шаблонів: індикатор тематичної події одного зі словників і варіанти розташування інших перемінних в реченні документа певної категорії. Шаблони для словників № 1 і № 2 відрізняються.

Виявлення нових подій за таким алгоритмом з величезного масиву документів Twitter – вельми довге і ресурсовитратне завдання, тому ми дійшли висновку, що необхідно поліпшити даний підхід за допомогою рейтингування джерел Twitter для кожної з категорій предметної області.

Для цього, виявивши події і їхніх авторів, ми будуємо рейтинг джерел певної категорії. Рейтинг являє собою ідентифікатор джерела, назву джерела, кількість всіх документів (твітів) від даного джерела за N часу, кількість твітів-подій за N часу, показник ефективності даного джерела. Показник ефективності джерела – це співвідношення твітів-подій до загальної кількості твітів цього джерела * 100. N = 3 місяці від поточної дати.

Із рейтинга для більш ніж 1850 джерел ми побачили, що ~ 20% джерел мають справді високий показник подійності. Отже, залишено лише 270 джерел, за якими йде моніторинг подій і виявлення їхньої новизни.

Це кардинально вирішило проблему швидкості оброблення, виявлення нових подій і масштабів споживаних для цього ресурсів. Водночас повнота майже не страждає. Отже, наш метод можна розділити на такі етапи:

1. Первинне виявлення твітів за заданою категорією і їх попереднє оброблення.
2. Другий етап передбачає кластеризацию і розподіл твітів на групи (категорії).
3. Виявлення подій у базі твітів визначеної категорії.
4. Аналіз найбільш значущих джерел данихт вітів, % їхньої ефективності (подійності) і присвоєння рейтингу джерелам.
5. Подальший моніторинг і екстрагування подій тільки з рейтингових джерел протягом певного часу.
6. Порівняння даних про подію з минулими подіями з бази даних (далі – БП) подій, зважування всіх аргументів і ухвалення рішення про новизну події.

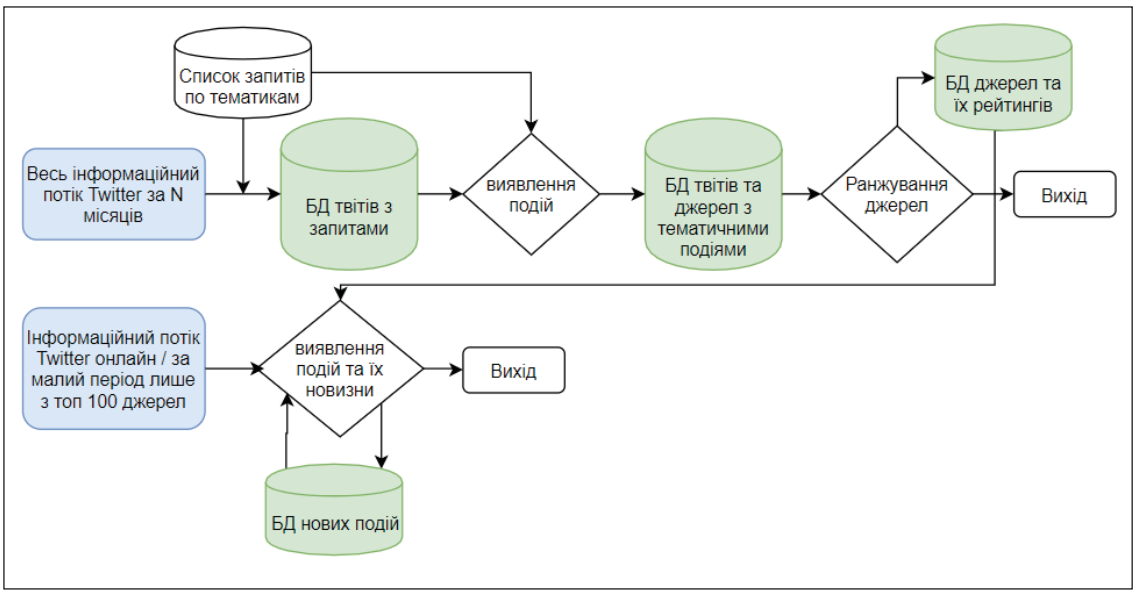

Рис. 3. Схема роботи системи рейтингування джерел та екстрагування нових подій

7. Аналіз ефективності найбільш рейтингових джерел, повторнийаналіз значущості джерел і перерозподіл рейтингу в певний проміжок часу.

На етапі виявлення подій першим кроком в аналізі тексту з використанням статистичних моделей (алгоритмів) є перетворення тексту на числа. Для наших аналітичних методів ми зробили це шляхом підрахунку кількості документів, які містять даний термін (ключове слово) у документах певної категорії. А для кожного документа визначена мітка часу, що дозволяє зробити аналіз тимчасовим. Для забезпечення можливості багатокритеріального порівняння подій і повідомлень у соціальних мережах модель події має містити компоненти, що відображають різні її аспекти. Це дає можливість виявляти саме нові події:

$$\underset{i}{\lambda} = \left\langle M_i^m, W_i^w, T_i^t, S_i^s, B_i^b.Z_i^z, P_i^p, G_i^g, C_i^c \right\rangle .$$

Модель події складається з таких компонентів:

1. Множина повідомлень, що описують подію, – $M_i^m = (m_i^1, m_i^2, m_i^3, ...m_i^{Ni})$.

2. Вектор слів документа, що описує подію, – $W_i^d = (w_i^1, w_i^2, w_i^3, ...w_i^{Ni})$.

3. Вектор заголовків документа, що описує подію, – $T_i^d = (t_i^1, t_i^2, t_i^3, ...t_i^{Ni})$.

4. Вектор слів у реченні – $S_i^d = (s_i^1, s_i^2, s_i^3, ...s_i^{Ni})$.

5. Множина числових значень (дат) із текстів документів – $B_i^b = (b_i^1, b_i^2, b_i^3, ...b_i^{Ni})$, метадані про час та дату – $Z_i^z = (z_i^1, z_i^2, z_i^3, ...z_i^{Ni})$.

6. Множина осіб (фізичних та юридичних), пов'язаних із подією, – $P_i^d = (p_i^1, p_i^2, p_i^3, ...p_i^{Ni})$.

7. Множина локацій – ($G_i^d = (g_i^1, g_i^2, g_i^3, ...g_i^{Ni})$.

8. Приналежність до однієї з тем (категорій) – $C_i^d = (c_i^1, c_i^2, c_i^3, ...c_i^{Ni})$.

Опорні слова для кожної категорії з показником Strange (EDF) та список слів для кожної категорії документів із поліченою за TF-IDF вагою кожного зі слів для певної категорії дають змогу обчислити вагу всього документа. Strange (EDF) визначається як просте відношення кількості знайдених документів із подією з даним тригером події (Ew) до кількості всіх документів у вибірці певної категорії з даним ключовим словом (Dw):

$$EDF_w = \frac{E_w}{D_w}.$$

TF-IDF обчислюємо за загально прийнятою формулою:

$$tf - idf(w, d, D) = tf(w, d) \times idf(w, D);$$

$$tf(w, d) = \frac{n_w}{\sum_k n_k},$$

де nw – кількість входжень слова w у документ, а знаменник – загальна кількість слів у документі.

$$idf(w, D) = \log \frac{|D|}{|\{d_i \in D | w \in d_i|},$$

де $|D|$ – кількість документів у корпусі, а $|\{d_i \in D | w \in d_i|$ – кількість документів із колекції D, в яких трапляється слово w. У разі роботи з текстами великого і малого розміру одночасно проводимо подвійну нормалізацію частоти появи так, що частота появи терміна w:

$$TF_w = 0.5 + 0.5 \times \frac{C_w}{B},$$

де C_w – кількість вживань терміна w у тексті; B – кількість разів, коли трапився найчастотніший термін цього тексту.

Даний вид вимірювання TF нівелює можливу помилку, спричинену впливом розміру тексту.

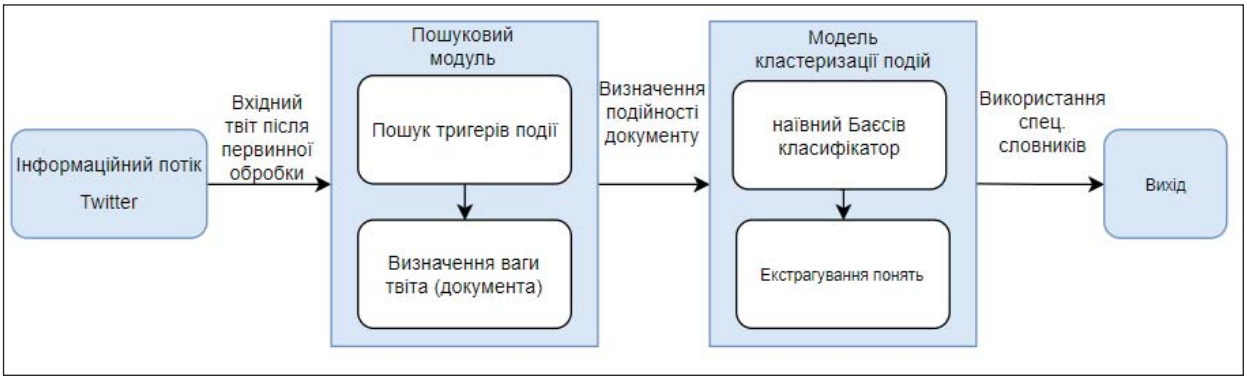

Рис. 4. Схема роботи системи екстрагування нових подій

Наступний крок аналізу документа, в якому знайдений індикатор події (тригер), коли вага документа більше порогового значення – виявлення понять (аргументів) у документі, порівняння цих показників із наявними новими подіями. Алгоритми виявлення понять, сутностей та їх взаємозв'язків детально описані в публікації А. Додонова та Д. Ланде [5, с. 45–51].

Для екстрагування понять використовуємо спеціальні словники для кожного типу понять, а також спеціальні правила (шаблони) їх розбору. Завдяки порівнянню можливої нової події з базою вже накопичених минулих подій з'являється можливість статистичного моделювання розвитку трендів та виявлення пікових показників певної події. А це, у свою чергу, дозволяє виявляти ключові теми для кожної з категорій за певний проміжок часу, кластеризувати документи як доповнення до наявних подій у БД. Оскільки кожен документ із вагою, більшою за порогове значення, розбирається на спеціальні (аргументи) поняття, то можна обчислити відстань між цим документом і минулими подіями з бази даних. Для цього кожен документ має бути представлений аналогічною багатокритеріальною моделлю:

$$d_j = \left\langle M_j^m, W_j^w, T_j^t, S_j^s, B_j^b, Z_j^z, P_j^p, G_j^g, C_j^c \right\rangle .$$

Для порівняння компонентів моделі наявної в БД події та нового документа, що представлені векторами (ключові слова, тригери події, заголовки), використовуємо косинусну міру:

$$\kappa_{i,j}^w = 1 - \frac{\sum_{k=1}^{N^w} \left[w_i^k w_i^k \right]}{\sqrt{\sum_{k=1}^{N^W} (w_i^k)^2} \sqrt{\sum_{k=1}^{N^W} (w_i^k)^2}} .$$

Для порівняння інших компонент, що представлені множинами, використовуємо міру входження:

$$\kappa_{i,j}^w = 1 - \frac{\left| \lambda_i^n \cap d_j^n \right|}{\left| d_j^n \right|} .$$

У нашому разі деякі множини зважені (фізичні й юридичні особи, географічні локації тощо), тому замість кількості елементів ми звертаємо увагу на їхню вагу:

$$\kappa_{i,j}^w = 1 - \frac{\sum\limits_{e \in \lambda_i^e \cap d_j^e} e}{\sum\limits_{e \in d_j^e} e} .$$

Результатом порівняння документа dj та події із БД подій λ_i є вектор, кожен елемент якого ввідображає відстань між d_j та λ_i за деяким критерієм. Задача визначення показників відстані вирішується машиною опорних векторів (SVM), що навчається на базі прикладів пар «подія: документ, що ввідноситься до події» та «подія: документ, що не ввідноситься до події». Усі вхідні документи або прив'язуються до наявних у БД подій, або самі визначаються як нові події. Водночас кожен новий документ d_j порівнюємо з кожним значенням знайдених раніше нових подій λ_i та знаходимо значення відстані між d_j та λ_i. Знаходимо подію із БД подій, що є найбільш близькою до d_j. Якщо відстань між документом і подією менша, ніж порогове значення, то відносимо цей документ до події. Якщо відстань більша за порогове значення, то створюємо в БД нових подій новий запис, що на даному етапі буде представлений лише цим документом (подією).

Висновки. Зменшення бази моніторингу завдяки ранжуванню джерел (авторів) за показником їхньої ефективності використання для виявлення нових подій у соціальній мережі Twitter – це швидкий і ефективний підхід. Дослідження показало, що повнота подій лише в найбільш рейтингових подіях задовольняє наші потреби в повноті та точності виявлення нових подій із даного інформаційного потоку. Завдяки роботі лише з найбільш рейтинговими джерелами можлива робота методу та виявлення нових подій майже в режимі онлайн (з невеликою затримкою). Змен-

шенням або збільшенням показника N (період часу, за який аналізуються джерела Twitter), а також порогового відсотка найбільш рейтингових джерел, з якого вони починають аналізуватися, можна збільшувати або зменшувати показники точності, повноти, та F1-міри.

Список літератури:

1. Jung H., Stent A. ATT1: Temporal annotation using big windows and rich syntactic and semantic features. 2013. P. 20–24.
2. Weiler A., Grossniklaus M., Scholl Marc H. Run-time and Task-based performance of event detection techniques for Twitter. 2015. P. 38–40.
3. Becker H., Naaman M., Gravano L. Beyond trending topics: real-world event identification on twitter. Proc. Intl. Conf on Weblogs and Social Media(ICWSM). 2011. P. 438–441.
4. Прищепа С. Технологія екстрагування нових подій за визначеною тематикою із соціальної мережі Twitter. Реєстрація, зберігання і оброб. даних. 2017. Т. 19. № 3. С. 67–72.
5. Додонов А., Ланде Д. Виявлення понять та їх взаємозв'язків у рамках технології контент-моніторингу. Реєстрація, зберігання і оброб. даних. 2006. Т. 8. № 4. С. 45–51.

ВЫЯВЛЕНИЕ НОВЫХ СОБЫТИЙ НА ОСНОВЕ РЕЙТИНГОВАНИЯ ИСТОЧНИКОВ TWITTER

В данной статье изучаются методы выявления новых событий в информационном потоке глобальной сети Интернет. Предлагается подход к выявлению новых событий с использованием метода рейтингования источников информации внутри социальной сети Twitter. Предложенный метод дает возможность уменьшить затраты на мониторинг источников и необходимые мощности для обработки и выявления новых событий. А перераспределение рейтинга источников увеличивает полноту и точность выборки с течением времени. Для традиционных методов экстрагирования новых событий характерны проблемы скорости выявления событий и необходимость в накоплении и обработке сверхбольших массивов данных. Данную проблему решает создание и мониторинг наиболее рейтинговых источников социальной сети с наибольшим количеством новых событий относительно всего количества сообщений.

Ключевые слова: *экстрагирование событий, рейтингование источников, обработка текста, Twitter, информационный поток.*

DETECTING NEW EVENTS BASED ON THE RATING OF TWITTER SOURCES

In this article, we study methods for identifying new events from the information flow of the global Internet. An approach his proposed for identifying new events using the method of rating information sources with in the sociall network Twitter. The proposed method makes it possible to reduce the cost of monitoring sources and the necessary capacity to hand lean identify new events. A redistribution of the source rating once in the N-time, in creases the fullness and accuracy of the sample overtime. Traditional methods of extracting new events have the problem of the speed of event detection and the need for accumulation and processing of super-large datasets. This problem is solved by the creation and monitoring of themstrated sources of the social network.

Key words: *event extraction, source rating, text processing, Twitter, information flow.*