УДК 001.8:004.7

С. В. Прищепа
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Технологія екстрагування нових подій за визначеною тематикою із соціальної мережі Twitter

Розглянуто проблеми і актуальність екстрагування подій. Проаналізовано основні підходи до екстрагування подій з інформаційних потоків. Сформовано свої методи та підходи до виявлення подій і визначення їхньої новизни (дублювання подій). На основі проведеного досліду запропоновано використовувати технологію екстрагування подій на основі спеціальних словників з тригерами подійності, класифікацією «подія/не подія» за методом класифікації наївного Байеса та екстрагування понять та аналізу на дублювання понять з минулими подіями з бази даних подій. Представлено підходи та алгоритм роботи технології на базі цих методів.

Ключові слова: *екстрагування подій, метод екстрагування подій, блок-схема екстрагування нових подій.*

Актуальність

В інформаційних сховищах України та світу зібрані сотні петабайт неструктурованих даних, величезні масиви таких даних виробляються щодня і можуть містити в собі критично важливі данні. Актуальність дослідження методів структурування, класифікації та екстрагування необхідної інформації зумовлена загальною потребою людства своєчасно виявляти нові події з динамічно зростаючих обсягів інформації у глобальних мережах. Разом зі зростанням об'ємів інформації зростає і кількість інформаційних джерел. Наведений нижче графік (рис. 1) яскраво відображає ріст кількості сайтів за минулі роки [1].

Слід відзначити, що ріст популярності соціальних мереж, таких як Facebook, Twitter, YouTube, сповільнив динаміку росту кількості активних сайтів. Це говорить про те, що ці інформаційні площадки стають основними джерелами, де створюється та розповсюджується інформація.

Проте, аби екстрагувати необхідну і актуальну інформацію зі слабо структурованих даних і проводити її аналіз, необхідно мати спеціальний інструментарій, в основі якого має знаходитися певна технологія.

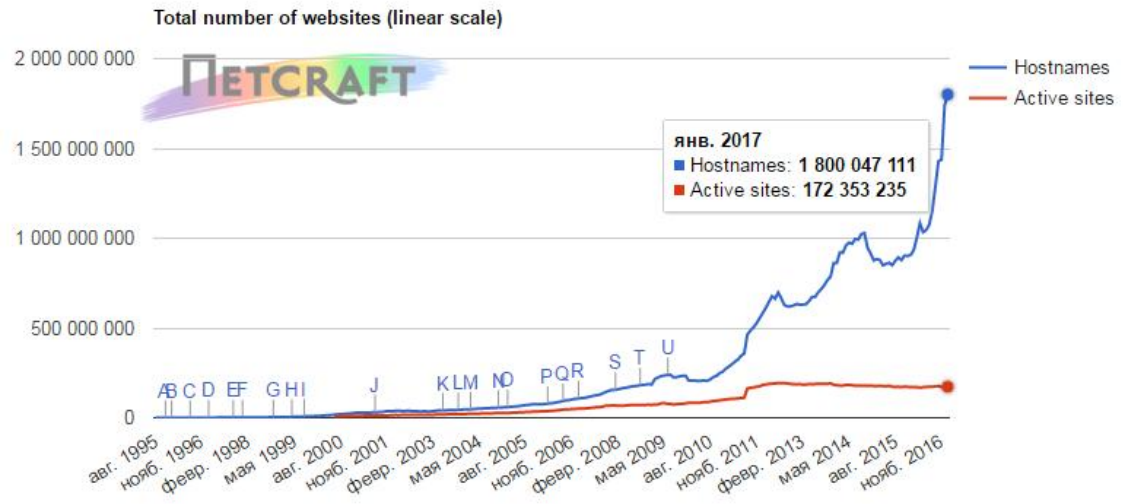

Рис. 1. Динаміка кількості сайтів у всіх доменах світу (за даними Netcraft, лінійна шкала)

Широке та стрімке зростання популярності соціальних мереж призвело до швидкого збільшення їхньої бази користувачів, що охоплюють увесь світ. Колективна інформація, що активно додається користувачами з усього світу, отримана з цих інтернет-платформ, є надзвичайною, з точки зору як обсягу кожного змісту, так і різноманітності обговорюваних тем і подій. Екстрагування виробленої користувачами інформації у режимі реального або близького до реального часу, дає змогу дослідити та швидко проаналізувати поточні події, що в свою чергу дає можливість своєчасно зрозуміти поточний стан справ різноманітних подій.

Світова практика доводить, що вирішення задачі ефективної аналітичної обробки інформації з глобальних мереж, а також оперативне вилучення фактографічних даних (у тому числі подій), виявлення трендів в окремих предметних областях, прогнозування — одні з ключових завдань 21-го сторіччя.

Оскільки твіти мають коротку форму, а користувачам не потрібно структурувати текст і чекати модерації, то вони мають перевагу в розповсюдженні та оновленні актуальної інформації. Також слід зауважити, що локальні події, що є незначними для масштабів країни — теж активно та швидко розповсюджуються через цю мережу, на відмінну від ЗМІ. Тому для розробки технології екстрагування нових подій з інформаційних потоків було обрано саме цю соціальну мережу. Мінусами Twitter можна назвати велику кількість шуму, збільшення кількості неправдивої інформації і специфічних скорочень у даному інформаційному потоці.

Екстрагування нових подій

Існує безліч можливих визначень події. Задачу екстрагування події визначено як виявлення конкретної ситуації в якійсь предметній області (стану, дії, процесу, властивості), що виражена одним або декількома відносинами.

Виявлення подій, як і виявлення сутностей в інформаційних потоках, — одне з головних завдань розбору текстів, яке успішно застосовується в науково-дослідних областях, таких як генерація онтологій, агрегація новин, класифікація текстів, підтримка прийняття рішень і бізнес-аналітика. Це важлива та складна

задача обробки природної мови, так як однакова подія може бути присутня в різних виразах і при цьому може виражати різні події залежно від контексту та часу. Так само, події часто мають вложеність — одна подія може привести до другої, наприклад, подія «злочин», призводить до «розслідування», а це, в свою чергу, може привести до події «затримання» або «арешту» [2].

Сам по собі Twitter має величезну масу короткої і своєчасної інформації, але для того, щоб залишатись у курсі всіх значимих подій, аналітику необхідно постійно сканувати величезну кількість джерел, що призводить до інформаційного перенавантаження людини. Існує велика кількість категорій подій, відслідковування яких може зацікавити аналітика. Саму тому, для екстрагування нових подій повинен бути створений підхід, в якому виявлення подій — досить легко налаштовуваний процес за домогою встановлення деяких правил (шаблонів) для конкретної категорії (тематики) та накопичення шляхом навчання з експертом-учителем спеціальних словників ключових слів тригерів події і мінус-слів для визначених тематик.

Серед напрямків аналізу було обрано такі показники як джерело (автор), дата та час екстрагування документа, приналежність тексту до певної категорії, наявність тригеру події, наявність місця і часу події та наявність можливих об’єктів чи суб’єктів події, відповідність певним шаблонам і приналежність до мережі слів (мови) певного типу подій. Для кращого розуміння подій і з чого складається подія в інформаційному просторі створена карта події (рис. 2).

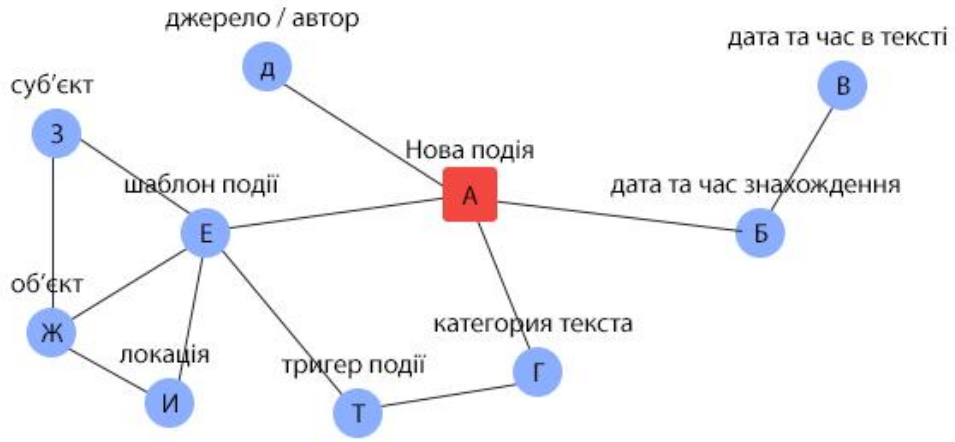

Рис. 2. Карта події

Виявлення подій і їхніх фігурантів у інформаційних потоках — складне завдання, яке має ряд проблем, серед яких можна відзначити:

— велику кількість можливих імен (назв) осіб і варіантів їхнього написання;

— велику ймовірність присутності в тексті осіб, що не мають прямого відношення до події;

— один індикатор події може виражати різні події у різному контексті;

— складні пропозиції з великою кількістю індикаторів подій;

— велику кількість псевдоподій;

— різні лексичні та семантичні складності розбору;

— часто не коректний синтаксис у реченнях.

Запропонована автором технологія екстрагування нових подій базується на використанні теорії математичного аналізу, теорії імовірностей і математичної статистики, теорії складних мереж, методів зважування важливості слів, обробки природних мов, машинного навчання та комп'ютерної лінгвістики.

Для запропонованої методології документ просто визначається як унікальна колекція тексту. Кожен документ — це твіт у соціальній мережі Twitter. Текстовий потік являє собою сукупність документів, в яких кожен документ має пов'язану з ним мітку часу. Кожен документ зазвичай містить метадані, що описують час публікації і дату, а також автора публікації.

Оскільки аналіз минулих наукових робіт у даній темі показав високі оцінки по F1-мірі у підходів, що роблять попередню класифікацію документів, то було прийнято рішення про необхідність розбору тільки тих текстів з інформаційного потоку, що зібрані за спеціальними ключовими словами, які можна відносити до певної категорії. Таким чином, увесь інформаційний потік розбивається на тематичні інформаційні потоки [3] для виявлення подій конкретних тематик. Такий підхід з використанням розробленої технології контент-моніторингу соціальних мереж дає змогу підвищити не тільки якість виявлення нових тематичних подій, але і зменшує кількість інформаційного шуму, що в свою чергу збільшують швидкість роботи та самого виявлення нових подій. Контент-моніторинг збирає весь інформаційний потік соціальної мережі, або його частину певною мовою, а далі проводиться екстрагування та розбір лише з тих масивів документів, у яких присутні специфічні для категорії ключові слова.

Після виділення масиву документів предметної області відбувається розбиття тексту на речення, очищення від розділових знаків (HTML та інші специфічні для соціальної мереж теги). Далі відбувається розбиття речень на окремі ключові слова. На вхід подаються вже оброблені «чисті» документи, що розбиті на окремі ключові слова та очищені від сполучників, розділових знаків тощо.

Перше, що робить запропонована модель з документом, який приходить їй на вхід — виявляє в документі наявність тригера (індикатора) події. Для кожної категорії подій створюються свої словники тригерів події. Словник тригерів події на початковому етапі створюється з експертом-вчителем. Для кожної категорії вони свої, але можуть бути дуже схожими між собою. Найчастіше — це дієслова закінченого часу, що пов'язані з подіями певної тематики. Наприклад: сталося, скоєний, затримано, повідомляється, трапилося тощо. Якщо тригер події не знайдений — документ пропускається. Кожен тригер події має свою вагу в межах від 0 до 1. Вага розраховується як відношення документів-подій з даним тригером до усіх знайдених документів з даним тригером події. Наша технологія дозволяє змінювати пороговий показник ваги тригера події, починаючи від якого тригер вважається дійсно тригером події та використовується в алгоритмі.

У випадку, коли тригер події наявний, переходимо до наступного етапу зважування всіх слів у реченні та визначення ваги документа. Використаємо наївний Баєсів класифікатор для оцінки ймовірності, чи є даний документ подією. При цьому маємо лише два класи — подія та не подія. Перевагою наївного байєсівського класифікатора є мала кількість даних для навчання, які потрібні для оцінки параметрів, що необхідні для класифікації.

Для даної задачі класифікації будемо вважати, що документи вибрані з декількох класів документів, які можуть бути представлені безліччю слів з (незалежною) вірогідністю, що термін w_i даного документа зустрічається в документі класу C:

$$p\left(w_i \middle| C\right).$$

Для розбору даної задачі припустимо, що вірогідність зустрічі слова в документі незалежна від довжини документа, і всі документи мають однакову довжину. Тим більше, що більшість постів у Twitter мають, в основному, довжину 50–140 символів. Тоді вірогідність для даного документа D і класу C вираховується за формулою:

$$p\left(D \middle| C\right) = \prod_i p\left(w_i \middle| C\right).$$

При роботі з великими інформаційними потоками — навчальні дані не настільки великі, щоб частота рідких термінів оцінювалася адекватно. Таким чином, виходить, що, якщо якийсь термін у навчальних даних зустрічається тільки в документах класу «не подія», то оцінка відношення для класу «подія» буде дорівнювати нулю. Саме тому використовується згладжування Лапласа, просто додаючи одиницю до кожної частоти появи терміна w_i з урахуванням багаторазових поява терміна в документі. Таким чином:

$$p\left(w_i \middle| C\right) = \frac{W + 1}{\sum_{w \in V}\left(W + 1\right)} = \frac{W + 1}{\sum_{w \in V} W + \left|V\right|},$$

де $\left|V\right|$ — кількість термінів у словнику; w_i — один із термінів у документі; W — кількість появ терміна w_i у навчальних документах із класу C з урахуванням багаторазових появ терміна в документі.

Цей підхід можна інтерпретувати як апріорний рівномірний розподіл (кожен термін зустрічається в кожному класі по одному разу), що потім уточнюється на основі навчальних даних, які надходять. Відзначимо, що це — апріорна ймовірність появи терміну на рівні документа, а не класу.

Формулу Байеса можна інтерпретувати як визначення вірогідності того, що документ D належить класу C? Чому дорівнює $p\left(C \middle| D\right)$?

Згідно теореми Байеса

$$p\left(C \middle| D\right) = \frac{p\left(C\right)}{p\left(D\right)} p\left(D \middle| C\right),$$

де D — документ, який розглядається; C — один із класів.

У нашому випадку у нас тільки два класи — E (подія) та N (не подія). Тоді задача полягає у вирішенні відношення правдоподібності для конкретно взятого документа:

$$p\left(E \middle| D\right) = \frac{p(E)}{p(D)} \prod_i p\left(w_i \middle| E\right),$$

$$p\left(N\middle|D\right)=\frac{p\left(N\right)}{p\left(D\right)}\prod_i p\left(w_i\middle|N\right).$$

Таким чином, поділивши одне на інше, отримаємо відношення правдоподібності:

$$\frac{p\left(E\middle|D\right)}{p\left(N\middle|D\right)}=\frac{p\left(E\right)}{p\left(N\right)}\prod_i \frac{p\left(w_i\middle|E\right)}{p\left(w_i\middle|N\right)}.$$

Ґрунтуючись на спостереженні, що $p\left(E\middle|D\right)+p\left(N\middle|D\right)=1$, робимо висновок, що ймовірність відношення $p\left(E\middle|D\right)$ може бути порахована за формулою

$$\ln\frac{p\left(E\middle|D\right)}{p(N\middle|D)},\ \text{при цьому}\ p\left(E\middle|D\right)=\frac{e^q}{1+e^q},$$

де $q=\ln\frac{p\left(E\middle|D\right)}{p\left(N\middle|D\right)}$; e — число Ейлера, e = 2,7183.

Отримаємо логарифм правдоподібності:

$$\ln\frac{p\left(E\middle|D\right)}{p(N\middle|D)}=\ln\frac{p\left(E\right)}{p\left(N\right)}+\sum_i \ln\frac{p\left(w_i\middle|E\right)}{p\left(w_i\middle|N\right)}.$$

Для задачі класифікації на подія/не подія, ми вираховуємо та порівнюємо логарифм правдоподібності із заданим пороговим значенням h. Наприклад, якщо порогове значення h = 0,5, то, за умови, що $\ln\frac{p\left(E\middle|D\right)}{p(N\middle|D)}\geq 0,5$, відносимо цей документ до класу E — подія. Якщо менше, то це клас N — не подія. При цьому алгоритм робить перерахунок ваги для даного тригеру, а на вхід подається наступний документ. Якщо ж вага документа більше або дорівнює пороговому значенню, то переходимо до наступного етапу — екстрагування понять.

Під поняттями розуміємо виявлення осіб, місця, дати та часу події. Для виявлення можливих фізичних осіб — створюється загальний словник імен російських/українських і по-батькові з різними закінченнями, потім беруться біграми (триграми) з досліджуваного документа, де присутнє хоча б одне з імен словника (або імен по-батькові) і проводиться розбір слова після нього та до), якщо це слово починається з великої літери, то з великою часткою ймовірності можна стверджувати, що це слово прізвище (або ім'я) — це і є особа в даній події, за умови, що довжина кожного з слів не менше 3 символів.

Для виявлення можливих юридичних осіб у події беруться всі біграми (триграми) з данного документа текстів, в яких хоча би одне зі слів починалося з великої літери, і другим або третім словом була вказана або форма діяльності (ТОВ/ПП/ТОВ/ЗАТ/ООО і т.ін.), або слово типу компанія, компаній, холдинг і т.д.). Даний біграм (триграм) у даному документі — юридична особа в даній події, за умови, що довжина кожного зі слів не менше 3 символів.

Виявлення місця — важка задача через їхню величезну кількість і варіанти написання, а оскільки основною задачею є виявлення нових подій, то ми лише виявляємо наявність назв країн, міст, сіл, річок та областей. Для виявлення місця ми використовуємо словники з назвами можливих місць події.

Задача виявлення дати та часу вирішується таким чином, що йде пошук просто дат за декількома шаблонами запису дат, а також слова, описи дат: «учора, сьогодні, прямо зараз, 20 жовтня і т.д.». У випадках, коли знайдено слова, описи дат, звертаємося до метаданих документа і співвідносимо з датою у метаданих.

Коли проведено аналіз документа та виявлені усі його поняття — проводимо визначення дублювання — виявлення новизни у знайденій події. Чи є подія новою, і її слід занести до бази даних подій, чи це вже минула подія (спогади). Для цієї задачі робимо порівняння з БД наявних подій і досліджуваним документом. Кожне порівняння одного з понять (та тригером події) з кожною із подій у БД має оцінку. Оцінка 1 — поняття не присутнє, 0 — поняття присутнє. При цьому для осіб і місця використовуємо не одиницю, а 0,5, так-як деякі особи, такі як речники або особи, які займають високі державні посади, часто зустрічаються в нових подіях, а виявити всі місця не представляється можливим у рамкам даної технології.

Отже, якщо:

— особа 1 = 0,5;

— особа 2 = 0,5;

— місце = 0,5;

— дата та час у події або у метаданих = 1;

— тригер події = 1,

то сума оцінок = 3,5, і це нова подія. Було обрано порогове значення для прийняття документа як нової події — $L \geq 2,5$

Якщо сума більше або дорівнює пороговому значенню, то подія вноситься до бази даних нових подій, а також робиться перерахунок у словнику тригерів події.

Для кращого розуміння процесу виявлення нових тематичних подій було створено блок-схему роботи запропонованого алгоритму (рис. 3).

При виявленні нових подій важливим є авторитетність автора публікації, на основі якого можна прийняти рішення, що документ від одного з N авторів про одну й ту саму подію — є більш значущим (важливим) при однаковій повноті інформації у кожному з них.

Для вирішення цього завдання пропонується використовувати відомий алгоритм ранжування веб-сторінок, заснованих на зв'язках — HITS (hyperlink induced topic search), запропонований Дж. Клейнбергом [4] та адаптований під спеціальні теги посилань соціальної мережі Twitter. Алгоритм HITS забезпечує вибір з інформаційного масиву кращих «авторів» (вузлів, на які введуть посилання) і «посередників» (вузлів, від яких ідуть посилання цитування). В нашому випадку автор є хорошим посередником, якщо від нього йдуть зв'язки на авторів важливих подій, і навпаки, автор є хорошим автором, якщо на нього ведуть зв'язки від важливих авторів [5]. Відповідно до алгоритму HITS для кожного вузла мережі v_j рекурсивно обчислюється його значимість як автора $a(v_j)$ і посередника $h(v_j)$ за формулами:

$$a\left(v_j\right) = \sum_i h\left(v_i\right), \quad h\left(v_j\right) = \sum_i a\left(v_i\right).$$

У даних формулах підсумовування проводиться по всіх вузлах, які посилаються (або на які посилаються — в другій формулі) на даний вузол. Технологія на даному етапі не включає цей метод оцінки авторитетності авторів документів, але в майбутньому може бути ним доповнена.

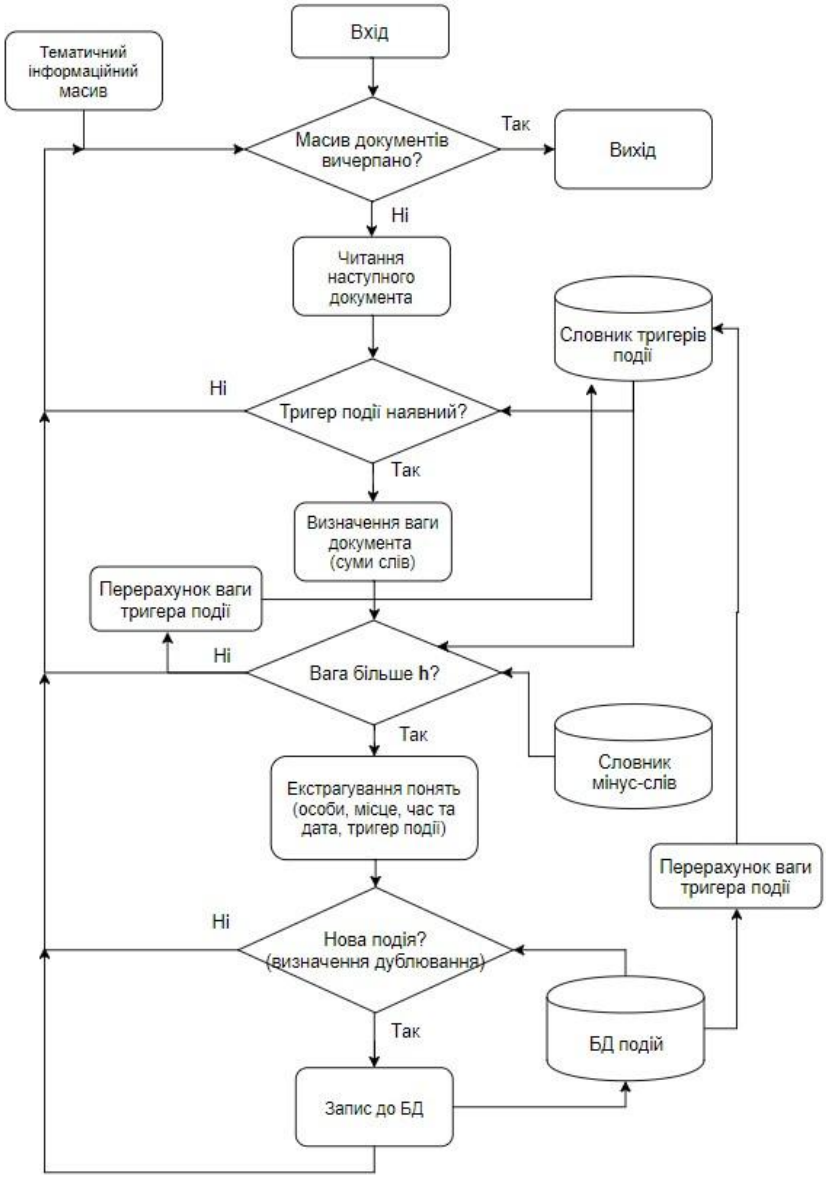

Рис. 3. Блок-схема екстрагування нових тематичних подій

Також, для подальшого поліпшення технології виявлення нових подій із соціальних мереж, можна використовувати крос-текстове виявлення подій та їхніх осіб за допомогою виділення документів за певний проміжок часу з однаковими або пов’язаними подіями і прийняття рішення вже по кластеру речень з певною подією. Також, за допомогою використання WordNet [6] можна розробити модель взаємозв’язку подій з предметами та іншими сутностями в інформаційних пото-

ках. Для забезпечення більшої повноти — використовувати фонові знання [7], здобуті раніше з різних баз даних і сайтів типу Wikipedia, а для цього необхідно будувати мережі слів.

Висновки

Нові методи та поліпшення вже використовуваних методів екстрагування подій — одна з найважливіших задач як у військовій і політичній сферах, так і в бізнесі. Якщо раніше кількість інформації у слабо структурованих джерелах дозволяла обробляти її аналітиком вручну, то зараз зі зростанням кількості доступної інформації і форматів її представлення в мережі, великою кількістю фейкової інформації і спаму, без автоматизації та спеціальних методів екстрагування подій та інших даних — це не можливо. В ході роботи було вивчено предметну область екстрагування нових подій з інформаційних потоків. У результаті аналізу літератури розроблено свою технологію екстрагування нових подій.

Задача обробки природної мови має безліч проблем, частину з яких не вирішується в рамках даної технології, але вирішення яких могло б покращити ефективність виявлення нових подій.

1. Netcraft — Internet Security and Data Mining. URL: https://news.netcraft.com/archives/2017/04/21/april-2017-web-server-survey.html (Last accessed: 30.03.2017). April 2017 Web Server Survey.

2. Nate Chambers and Dan Jurafsky. Unsupervised Learning of Narrative Schemas and their Participants. Proc. of ACL. 2009.

3. Додонов О.Г., Ланде Д.В., Путятін В.Г. Інформаційні потоки в глобальних комп'ютерних мережах. Київ: Наук.думка, 2009. 295 с.

4. Kleinberg J. Authoritative sources in a hyperlinked environment. In Proc. of ACM-SIAM Symposium on Discrete Algorithms. 1998. **46**(5). P. 604–632.

5. Ландэ Д.В., Снарский А.А. Подход к созданию терминологических онтологій. *Онтология проектирования*. 2014. № 2(12). С. 83–91.

6. Josu Goikoetxea, Eneko Agirre, and Aitor Soroa. Single or Multiple? Combining Word Representations Independently Learned from Text and WordNet. 2016.

7. Heng Ji. Relation extraction event extraction. 2014. URL: http://nlp.cs.rpi.edu/course/spring14/lecture9.pdf

Надійшла до редакції 04.09.2017