DOI 10.34185/1562-9945-2-127-2020-10
УДК 004.8

Ю.О. Олійник, О.Є. Афанасьєва, Г.Д. Аршакян

ПІДХІД ДО ВИЯВЛЕННЯ АНОМАЛІЙ В ПОТОКАХ ТЕКТОВИХ ДАНИХ

Анотація. Збільшення потоків інформації веде за собою необхідність розробки спеціалізованих інтелектуальних засобів та методів для обробки величезної кількості даних. Так популярність соціальних мереж, різного роду месенджерів вимагає створення спеціалізованих засобів для обробки потоку текстових повідомлень. Дане дослідження присвячене дослідженню та розробці методу виявлення аномальних елементів в потоках текстових даних. Особлива увага присвячена підтримці україномовних текстів.
Ключові слова: аномалія, isolation forest, text mining, реферація тексту, семантичний аналіз.

Постановка проблеми. Аномалія – певне відхилення від норми [1]. Під виявленням аномалії приймаємо пошук непередбачених значень або певних шаблонів у потоках даних. Відомі методи та підходи виявлення аномалій не розраховані на пряму роботу з текстовими даними, та більше підходять для виявлення аномалій в числових даних або категорійних даних. Тому необхідно представити підхід та реалізувати метод виявлення аномальних елементів в потоках текстових даних. Для цього необхідно вирішити такі завдання:

1. Запропонувати підхід для виявлення аномальних елементів в потоках.

3. Запропонувати метод автоматичної реферації тексту.

3. Реалізувати метод виявлення аномалій в потоках текстових даних.

4. Порівняти результати виявлення аномалій для потоку оригінальних документів та потоку документів після реферації.

5. Забезпечити підтримку україномовних текстів для розробленого підходу та методів.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Аналіз публікацій по темі дослідження. Виявленню аномалій присвячено досить багато досліджень. Дослідження [1] дає рекомендації з виявлення аномалій для різних областей, як то кібербезпека, фінанси, охорона здоров'я, оборона, виробництво та інше. Для виявлення аномалій пропонується використати 3 підходи: оснований на відстані, оснований на щільності, оснований на ранжуванні. В нашому випадку необхідно оброблювати потоки даних, що веде до постійної появи нових даних. В цьому випадку будуть виникати проблеми виявлення аномалій в режимі онлайн. Для таких випадків пропонується використати «віконну» обробку даних.

В роботі [2] проведено аналіз методів та проаналізовано інструменти для роботи з україномовними текстами та запропоновано алгоритм виявлення аномалій методом Isolation Forest. Алгоритм описано в [3] та доведена його ефективність на даних різної природи.

У статті [4] наведено модифікацію методу Isolation Forest для пошуку аномалій в потокових даних з використанням «плаваючого вікна». Метод показує високу ефективність для числових факторів, але в нього відсутня підтримка роботи з текстовими даними. У роботі [5] відбувається аналіз на основі семантичного та статистичного аналізу. Пропонується кожний документ відносити до певної категорії або декількох категорій, що динамічно формуються на основі методу LDA. Семантичний аналіз виконується на основі лексичної мережі англійської мови Wordnet [5]. Такий підхід цікавий, але у ньому відсутня підтримка україномовних текстів.

Мета дослідження. Метою дослідження є підвищення якості аналізу потоків текстової інформації українською мовою.

Основна частина. Потік даних $S = \{(d_0, t_o), (\dots), (d_i, t_i), (d_{i+1}, t_{i+1}), (\dots)\}$ – є нескінченним потоком даних, що надходять з одного або кількох джерел де отримана пара (d_i, t_i) означає, що повідомлення di отримане в час t_i [4].

У такому випадку часовим вікном W_i є інтервал часу фіксованого розміру δ, що починається у точці t_i.

Наведемо схему моделі потоку даних за часовим вікном на рис 1.

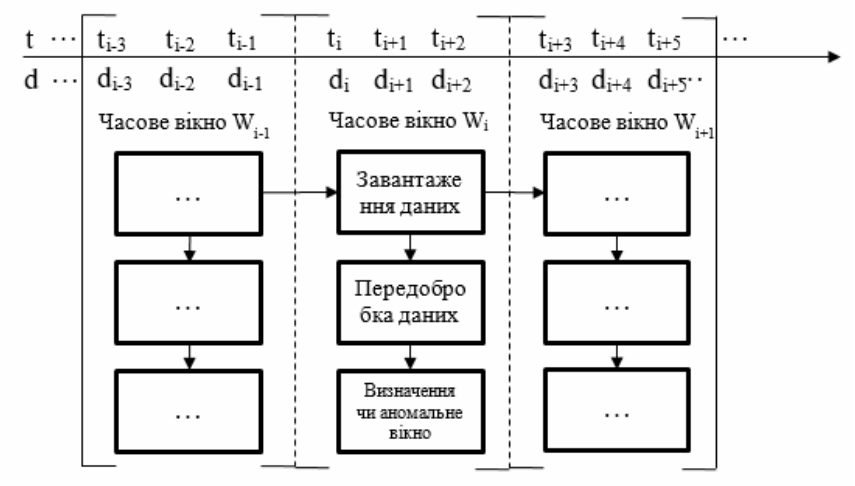

Рисунок 1 – Модель потоку даних

Наведемо алгоритм розбиття потоку даних на часові вікна:

Вхідні дані: S – потік даних, δ – фіксований інтервал часу

Вихідні дані: W_j – часове вікно

1. Встановлення границі часового вікна $g = t_i + \delta$

2. for $t_i \leqslant g$ do

3.$W_j \leftarrow W_j \cup S(^{d_i, t_i})$

4. return W_i

Ще такий потік даних можна представити у вигляді матриці, де колонки представляють партії документів з n елементами.

$$\begin{matrix} d_{1,1}, & d_{1,2}, & \cdots, & d_{1,m_1}; \\ d_{2,1}, & d_{2,2}, & \cdots, & d_{2,m_2}; \\ \cdots; & & & \\ d_{n,1}, & d_{n,2}, & \cdots, & d_{n,m_n}; \\ \cdots: & & & \end{matrix}$$

де $d_{i,j}$ представляє j документ в і-му «вікні», де $d_{i,j} =< X_{i,j}, y_{i,j} >$ де $X_{i,j} \in R_n$ представляє зразок тексту в потоці, $y_{i,j} \in \{+1, -1\}$ визначає чи документ до класу аномальних елементів ($y_{i,j} = +1$) чи не відноситься до класу нормальних елементів ($y_{i,j} = -1$).

1. Попередня обробка тексту

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Для подальшого використання методів визначення аномалій необхідно виконати попередню обробку тексту[6], що включає токенізацію та сегментацію даних, видалення шуму та нормалізацію даних. Видалення шуму використовується для покращення якості даних перед їх обробкою.

При токенізації та сегментації даних відбувається розділення текстового документу на частини – речення, фрази, окремі слова, а для розділення використовуються розділові знаки: «пропуск», «крапка», «кома» і т.д.

При оброці елементів в текстових потоках даних очистка даних потрібна для:

а) усунення нерелевантних символів (наприклад, будь-які символи окрім цифр та букв);

б) видалення нерелевантних слів (таких як згадування в соціальних мережах та посилання на інші ресурси);

в) переведення усіх символів в нижній реєстр.

г) видалення слів з довідника стоп-слів. Не всі слова є необхідними для класифікації. Наприклад, займенники чи прийменники не несуть основний зміст. Тому для покращення ефективності алгоритму такі слова необхідно видаляти. Така процедура відбуваються незалежно від тематики тексту. Для виконання даного кроку використовується спеціальний довідник.

Для нормалізації даних використовується стемінг та лематизація. Стемінг - евристичний спосіб пошуку основи слова. При такому підході основа слова не обов’язково має співпадати з морфологічний коренем заданого слова. В процесі стемінгу відкидаються морфологічні складові слів. В більшості випадків це закінчення, але також можуть відкидатись суфікси чи префікси. Лематизація – це спосіб нормалізації текстових наданих, який використовує словник та морфологічний аналіз, що знайти правильну форму – лему.

Для виділення додаткових ознак з текстових документів використовується модель “Bag of word” та метрика TF-IDF (від англ. TF – term frequency, IDF – inverse document frequency) статистична міра, що використовується для оцінки важливості слова в контексті документа, що є частиною колекції документів або корпусу. Вага деякого слова пропорційна кількості вживання цього слова в документі, і обернено пропорційна частоті вживання слова в інших документах колекції.

$$tf - idf(t,d,D) = \frac{n_t}{\sum_k n_k} \times log \frac{(|D|)}{|\{d_i \in D \mid t \in d_i\}|},$$

де n_t – кількість входів слова t в документ, $\sum_k n_k$- загальна кількість слів у документі, |D| - кількість документів в корпусі, $|\{d_i \in D \mid t \in d_i\}|$ - кількість документів з колекції D, в яких зустрічається t.

Більшу вагу в TF-IDF отримують слова з високою частотою у межах конкретного документа і з низькою частотою вживань в інших документах.

Модель Bag of Words - техніка вилучення ознак, що використовується при роботі з текстовими даними. Вона описує входження кожного слова в текст або корпус текстів.

Векторна модель тексту – представлення тексту у вигляді множини векторів речень (в яких кожному терму проставлена вага - навіть якщо терм не входить до речення.

Програмні засоби та бібліотеки з підтримкою україномовних текстів. Наразі існує обмежена кількість програмних засобів та програмних бібліотек, що мають підтримку україномовних текстів. Pymorphy2[7] - це морфологічний аналізатор та генератор для російської та української мови, що використовує великі ефективно закодовані лексикони, побудовані з даних OpenCorpora[8] та LanguageTool. MITIE [9] – бібліотека

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

та засоби для екстракції інформації побудованій на високопродуктивній бібліотеці «DLIB» [10]. Словник ВЕСУМ (великий електронний словник української мови)[11] містить слова та їхні парадигми з відповідними тегами, а також іншу інформацію, зокрема: додаткові теги, зв'язок між базовими та порівняльними формами прикметників, керування відмінками для прикметників.

2. Автоматична реферація текстів

Реферування тексту (Summarization) - скорочення його обсягу та отримання короткого викладу його змісту Огляд методів автоматичної реферації текстів наведено в роботі [11]. Показано, що наразі практично не існує якісних засобів автоматичної реферації текстів українською мовою. Для оцінки якості реферування використовується косинус подібності $\cos\theta \to 1$.

Косинус подібності розраховується як косинус перетину між векторними поданнями TF-IDF між сукупністю речень, що входять до реферату (input) та оригінального тексту(full).

$$\cos\theta = \frac{TFIDF_{input} \cdot TFIDF_{full}}{|TFIDF_{input}||TFIDF_{full}|}$$

Для автоматичної реферації пропонується наступний алгоритм, наведений на рис.2.

Використано комбінований метод, який вмістить у собі варіативність підходу LSA та TextRank щодо розрахунків вагових коефіцієнтів, містить в собі усі необхідні перетворення для покращення якості та має менше недоліків за LSA та TextRank. За рахунок використання морфологічного аналізатора[7] метод підтримує обробку україномовних та російськомовних текстів. Під час виконання операції реферування, одночасно будуть працювати алгоритми TextRank та LSA, але наприкінці буде видаватись тільки той результат, що є кращим за показником косинуса подібності – як одного з незалежних показників, що використовується для

оцінки якості роботи алгоритму реферування. Для оцінки якості роботи алгоритму автоматичного реферування було сформовано набір україномовних даних з понад 1100 елементів новинного сайту http://korrespondent.net/ за період з жовтня 2019 по січень 2020. Результат роботи алгоритму завантажено на ресурс [13]. Кожний елемент набору даних було оброблено алгоритмом, для формування реферованих документу зі зменшенням об'єму до 40% (Summary 40%) та 20% (Summary 20%). Середнє арифметичне Cos θ для документів Summary 40% склав 0,9033, для Summary 20% - 0,7812, що є досить непоганим результатом.

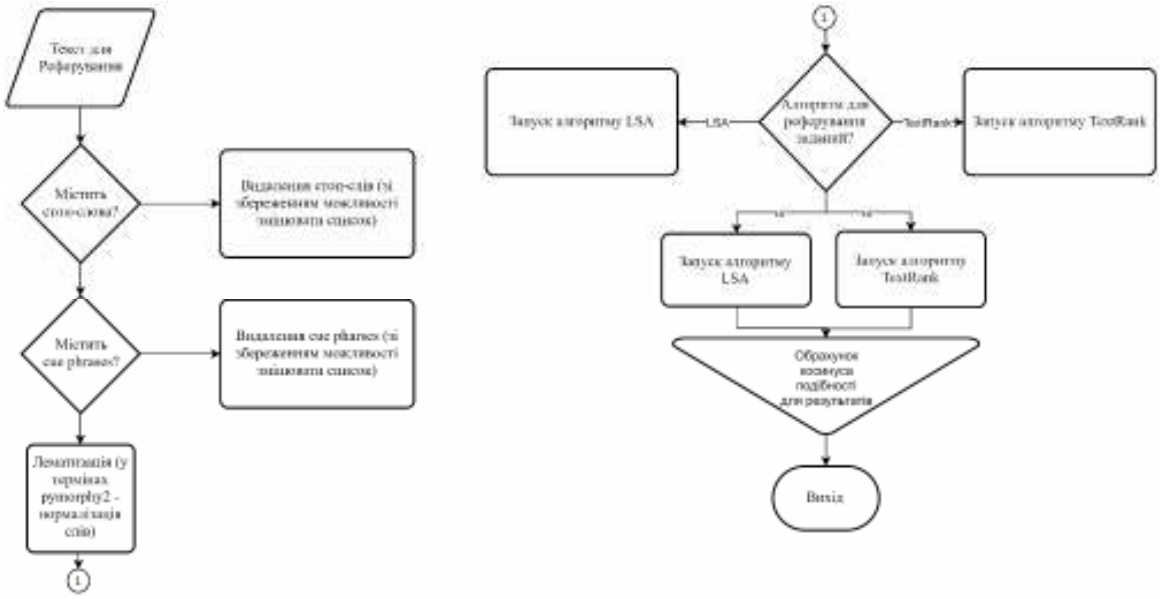

Рисунок 2 – Алгоритм реферації тексту

3. Метод Isolation Forest для виявлення аномалій в потоках текстових даних

Для виявлення аномальних елементів потоках текстових даних виділимо ознаки документів, які можуть бути використані в методах[2,3].

В рамках передобробки даних проводяться операції попередньої обробки, описані вище. Використаємо два фактора для методу Isolation Forest: метрика TF-IDF окремого документа, та довжина L кожного повідомлення.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

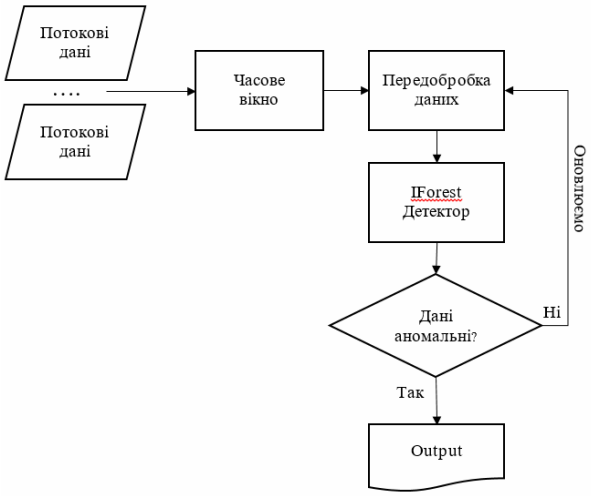

Рисунок 3 – Модернізований метод Isolation Forest

Проведемо експеримент, що визначає метрику TF-IDF для оригінальних документів, та документів після реферування для набору даних[13].

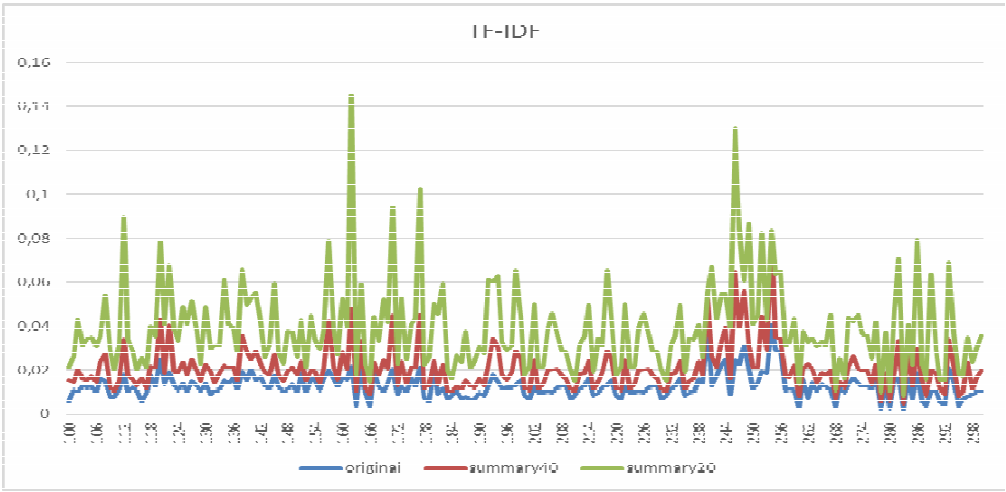

Рисунок 4 – Результат визначення TF-IDF метрики для оригінальних та реферованих текстів

Як видно з графіку (рис.4) метрика TF-IDF оригінальних документів схожу форму з документами «Summary 40%»(„40%”) та «Summary 20%» („20%”). Але для документів «Summary 20%» рівень метрики TF-IDF найвищий. Це пояснюється вилученням неважливих слів та речень з документів, що відповідно збільшує метрику важливих слів.

Проведемо експеримент порівняння швидкості роботи алгоритму виявлення аномалій для потоку оригінальних документів та потоків

реферованих документів. Результат експерименту представлено на рис. 5. Для потоку даних з рефераціею «20%» приріст швидкодії склав 70%, для потоку «40%» відповідно 47%.

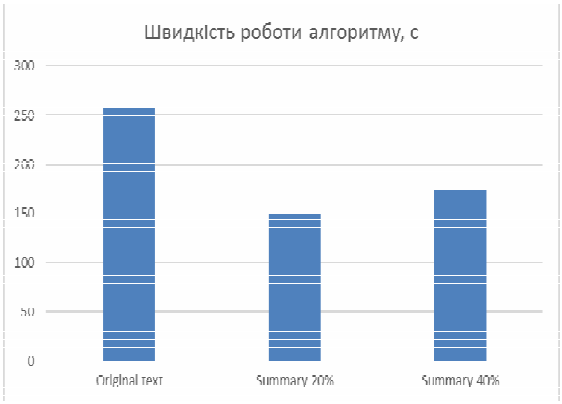

Рисунок 5 – Швидкість роботи алгоритму, с

Результати виявлення аномалій представлено на рис. 6. Параметри для алгоритму Isolation Forest: кількість дерев – 273, максимальна довжина дерева 22, точка розколу – 2, середня довжина дерева 13,867. Відсоток аномальних повідомлень – 7.2%. Аномальні елементи для набору даних оригінальних повідомлень та набору «20%» співпадають на 81%, для «40%» на 87%. Синім кольором на рис.6 виділено аномальні елементи для потоку оригінальних повідомлень.

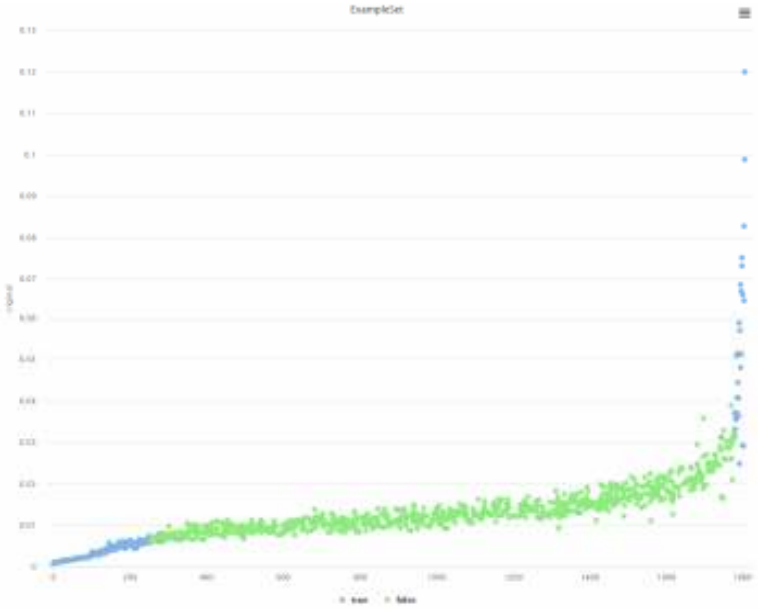

Рисунок 6 – Результат роботи алгоритму Isolation Forest

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Програмна архітектура запропонованого підходу описана в статті [14].

Висновки. Представлено підхід виявлення аномальних елементів в потоках текстових даних з виконанням попередньої обробки текстових даних та проведення реферування тексту. Для проведення автоматичного реферування тексту розроблено комбінований метод на основі LSA та TextRank. За основу методу виявлення аномалій взято метод Isolation Forest та модель потоку даних. Метод підтримує обробку україномовних та російськомовних текстових даних. Визначено, що TF-IDF метрики оригінальних та реферованих документів мають лінійну залежність. Також визначено, що швидкодія обробки потоки реферованих даних збільшується на десятки відсотків в залежності від рівня зменшення об'єму документів. Виявлені аномальні елементи для потоків оригінальних та реферованих співпадають більше ніж на 80%.

ЛИТЕРАТУРА / ЛІТЕРАТУРА

1. Mehrotra K.G., Mohan C.K., & Huang H. (2017). Anomaly detection principles and algorithms (p. 217). New York, NY, USA:: Springer International Publishing.

2. Афанасьєва О.Є. Виявлення аномалій в потоках текстових даних / Афанасьєва О.Є., Олійник Ю.О. // Всеукраїнська науково-практична конференція молодих вчених та студентів «Інформаційні системи та технології управління – ІСТУ-2019». Секція кафедри автоматизованих систем обробки інформації і управління. м. Київ: НТУУ «КПІ ім. Ігоря Сікорського», 26 листопада 2019 р,– С. 88-92

3. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining (pp. 413-422). IEEE.

4. Ding, Z., & Fei, M. (2013). An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. IFAC Proceedings Volumes, 46(20), 12-17.

5. WordNet - A Lexical Database for English [Electronic Resource] – Mode of access: World Wide Web: wordnet.princeton.edu - - Title from the screen

6. Олійник Ю. О., Огляд та аналіз алгоритмів TEXT MINING / Гавриленко О.В., Олійник Ю. О., Г. В. Ханько. // Управління проектами, системний аналіз і логістика. – К.: НТУ, 2017. – Вип., С32-41
7. pymorphy2 - Морфологійчний аналізатор [Электронний ресурс] / Режим доступа: https://pymorphy2.readthedocs.io/ - Назва з екрана
8. Open Corpora [Electronic Resource] – Mode of access: World Wide Web: http://opencorpora.org/ (viewed on September 20, 2019). – Title from the screen.
9. MIT Information Extraction [Electronic Resource] – Mode of access: World Wide Web: https://github.com/mit-nlp/MITIE/ - Title from the screen
10. Dlib toolkit [Electronic Resource] – Mode of access: World Wide Web: http://dlib.net/ - - Title from the screen
11. Великий електронний словник української мови ВЕСУМ - [Электронний ресурс] / Режим доступа: https://github.com/brown-uk/dict_uk - Назва з екрана
12. Аршакян Г.Д. Огляд підходів та методів автоматичного реферування тексту / Аршакян Г.Д. Олійник Ю.О. // Всеукраїнська науково-практична конференція молодих вчених та студентів «Інформаційні системи та технології управління – ІСТУ-2018». Секція кафедри автоматизованих систем обробки інформації і управління. м. Київ: НТУУ «КПІ ім. Ігоря Сікорського», 26 листопада 2019 р,– С. 194-198
13. Набір даних для аналізу [Электронний ресурс] / Режим доступа: https://drive.google.com/open?id=1-aImiiTqKJfIWxmifnI4GZSMbVzfnfvi - Назва з екрана
14. Tomashevskii V.M., Oliynik Y.O., Yaskov V.V., Romanchuk V.M. (2018). Realtime text stream anomalies analysis system. Вісник Херсонського національного технічного університету, (3 (1)), 361-365.

REFERENCES

1. Mehrotra K.G., Mohan C.K., & Huang, H. (2017). Anomaly detection principles and algorithms (p. 217). New York, NY, USA:: Springer International Publishing.
2. Afanasieva O.Ie. Vyiavlennia anomalii v potokakh tekstovykh danykh / Afanasieva O.Ie., Oliinyk Yu.O. // Vseukrainska naukovo-praktychna

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

konferentsiia molodykh vchenykh ta studentiv «Informatsiini systemy ta tekhnolohii upravlinnia – ISTU-2019». Sektsiia kafedry avtomatyzovanykh system obrobky informatsii i upravlinnia. m. Kyiv: NTUU «KPI im. Ihoria Sikorskoho», 26 lystopada 2019 r,– S. 88-92

3. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining (pp. 413-422). IEEE.

4. Ding, Z., & Fei, M. (2013). An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. IFAC Proceedings Volumes, 46(20), 12-17.

5.WordNet - A Lexical Database for English [Electronic Resource] – Mode of access: World Wide Web: wordnet.princeton.edu - - Title from the screen

6. Yu. Oliynik. Review and analysis of algorithms TEXT MINING / O. Gavrilenko, Yu. Oliynik, H. Hanko. // Project management, systems analysis and logistics. – K .: NTU, 2017. - Vol., pp32-41

7. pymorphy2 – Mode of access: World Wide Web: https://pymorphy2.readthedocs.io/ – Title from the screen

8. Open Corpora [Electronic Resource] – Mode of access: World Wide Web: http://opencorpora.org/ (viewed on September 20, 2019). – Title from the screen.

9. MIT Information Extraction [Electronic Resource] – Mode of access: World Wide Web: https://github.com/mit-nlp/MITIE/ - Title from the screen

10. Dlib toolkit [Electronic Resource] – Mode of access: World Wide Web: http://dlib.net/ - Title from the screen

11. BEDUL Dictionary- Mode of access: World Wide Web: https://github.com/brown-uk/dict_uk Title from the screen

12. Arshakian H.D. Ohliad pidkhodiv ta metodiv avtomatychnoho referuvannia tekstu / Arshakian H.D. Oliinyk Yu.O. // Vseukrainska naukovo-praktychna konferentsiia molodykh vchenykh ta studentiv «Informatsiini systemy ta tekhnolohii upravlinnia – ISTU-2018». Sektsiia kafedry avtomatyzovanykh system obrobky informatsii i upravlinnia. m. Kyiv: NTUU «KPI im. Ihoria Sikorskoho», 26 lystopada 2019 r,– S. 194-198

13. Dataset for data analysing Mode of access: World Wide Web: https://drive.google.com/open?id=1-aImiiTqKJfIWxmifnI4GZSMbVzfnfvi - Title from the screen

14. Tomashevskii, V. M., Oliynik, Y. O., Yaskov, V. V., Romanchuk, V. M. (2018). Realtime text stream anomalies analysis system. Вісник Херсонського національного технічного університету, (3 (1)), 361-365.

Received 14.02.2020.
Accepted 18.02.2020.

Подход для обнаружения аномалий в потоках текстовых данных

В статье рассмотрен подход выявления аномальных элементов в потоках текстовых данных с выполнением предварительной обработки данных и проведения реферирования текста. Для проведения автоматического реферирования текста разработан комбинированный метод на основе LSA и TextRank. За основу метода выявления аномалий взято метод Isolation Forest и модель потока данных. Метод поддерживает обработку украиноязычных и русскоязычных текстовых данных. Проведено сравнение быстродействия обработки потоков оригинальных и реферируемых данных.

Text stream data anomalies detection approach

Data stream increasing demands development new intellectual tools and methods for Big-Data processing. Existing anomalies detection approaches based on distance, density and ranking. But these approaches do not take into account data stream features. Unfortunately, in the context of streaming data, those methods more or less have some drawbacks and are not directly applied to streaming data, such as poor adaptability and extensibility, inability to detection novel anomaly, high model updating cost and slow updating speed so on. Besides, existing methods based on numeric or categorical data and do not support Ukrainian text data.

Purpose of research: creating new stream data anomalies detection approach with Ukrainian and Russian language text supporting.

Actually, several software libraries have Ukrainian language text support: Pymorphy2, OpenCorpora, LanguageTool, ВЕСУМ dictionary. Using data preprocessing (normalization, tokenization and noise reduction) and text abstracting for anomalies detection are proposed. Abstracting method developed on base combination of LSA and TextRank methods. For abstracting cosine similarity $\cos\theta$ *is used. For method evaluation was prepared Ukrainian dataset from news portal http://korrespondent.net/. Average cosine similarity between original dataset and dataset with 40% of volume is 0,9033, and for 40% of volume - 0,7812. Anomalies detection speed increase for "20%" dataset in 1.7 time in compare original dataset processing, and in 1.47 time for "40%" dataset.*

Anomalies detection method based on Isolation forest method. Input data: TF-IDF metrics, message lengths, TF-IDF metrics of original messages and "20%" / "40%" messages are similar.

Text stream data anomalies detection approach is presented. Method includes preprocessing and Abstracting stage. Abstracting method developed on base combination of LSA and TextRank methods. Anomalies detection method based on a Isolation Forest method and data

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

stream model. Ukrainian and Russian language text processing is supported. The processing speed of original and abstract data stream is compared.

Олейник Юрий Александрович - старший преподаватель кафедры автоматизированных систем обработки информации и управления, Национальный технический университет Украины «КПИ им. Игоря Сикорского ».
Афанасьева Елена Евгеньевна - магистрант кафедры АСОИУ Национального технического университета Украины «Киевский политехнический институт им. Игоря Сикорского ».
Аршакян Георгий Давыдович - магистрант кафедры АСОИУ Национального технического университета Украины «Киевский политехнический институт им. Игоря Сикорского».

Олійник Юрій Олександрович – старший викладач кафедри автоматизованих систем обробки інформації і управління, Національний технічний університет України «КПІ ім. Ігоря Сікорського».
Афанасьєва Олена Євгенівна - магістрант кафедри АСОІУ Національного технічного університету України «Київський політехнічний інститут ім. Ігоря Сікорського».
Аршакян Георгій Давидович - магістрант кафедри АСОІУ Національного технічного університету України «Київський політехнічний інститут ім. Ігоря Сікорського».

Oliynyk Yuriy - Senior Lecturer, Department of Automated Information Processing and Management Systems, National Technical University of Ukraine «KPI named after Igor Sikorsky».
Afanasyeva Elena - graduate student of the department of ASOIU of the National Technical University of Ukraine «Kyiv Polytechnic Institute named after Igor Sikorsky».
Arshakyan Georgy - undergraduate of the department of ASOIU of the National Technical University of Ukraine «Kiev Polytechnic Institute named after Igor Sikorsky». "