ДЕЯКІ АСПЕКТИ АНАЛІЗУ ПОТОКІВ ТЕКСТОВИХ ДАНИХ

Олійник Ю.О., Афанасьева О.Є. (студентка), Аршакян Г.Д. (студент)

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Україна

Abstract. Text stream data anomalies detection approach is presented in the article. Using data preprocessing (normalization, tokenization and noise reduction) and text abstracting for anomalies detection are proposed. 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 stream model. Ukrainian and Russian language text processing is supported. The processing speed of original and abstract data stream is compared.

Ключові слова: АНОМАЛІЯ, ISOLATION FOREST, TEXT MINING, РЕФЕРАЦІЯ ТЕКСТУ, СЕМАНТИЧНИЙ АНАЛІЗ.

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.

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

Потік даних $S = \{(d_0, t_0), (...), (d_i, t_i), (d_{i+1}, t_{i+1}), (...)\}$ – є нескінченним потоком даних, що надходять з одного або кількох джерел де отримана пара (d_i, t_i) означає, що повідомлення d_i отримане в час t_i [3]

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

1. Попередня обробка тексту. Для подальшого використання методів визначення аномалій необхідно виконати попередню обробку тексту[4], що включає токенізацію та сегментацію даних, видалення шуму та нормалізацію

даних. Видалення шуму використовується для покращення якості даних перед їх обробкою.

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.

Наразі існує обмежена кількість програмних засобів та програмних бібліотек, що мають підтримку україномовних текстів. Pymorphy2, OpenCorpora, LanguageToo, словник ВЕСУМ.

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).

$$\cos(\theta) = \frac{TFIDF_{input} * TFIDF_{full}}{\left|TFIDF_{input}\right|\left|TFIDF_{full}\right|} \qquad (1)$$

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.

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

Для виявлення аномальних елементів потоках текстових даних виділимо ознаки документів.

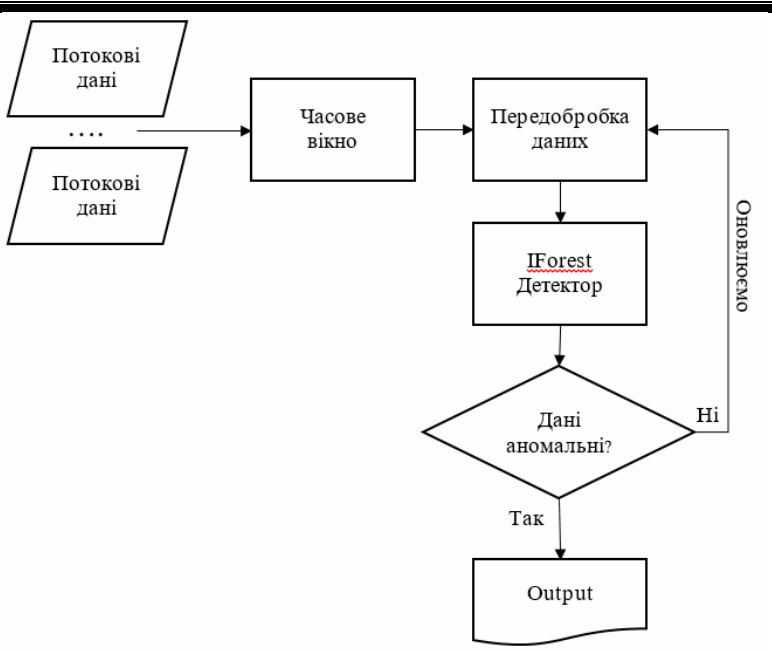

Рисунок 1 – Модернізований метод Isolation Forest

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. Метрика TF-IDF оригінальних документів схожа на форму з документами «Summary 40%» та «Summary 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. Порівняння швидкості роботи алгоритму виявлення аномалій для потоку оригінальних документів та потоків реферованих документів. Результат експерименту представлено на рис. 2. Для потоку даних з рефераційою «20%» приріст швидкодії склав 70%, для потоку «40%» відповідно 47%.

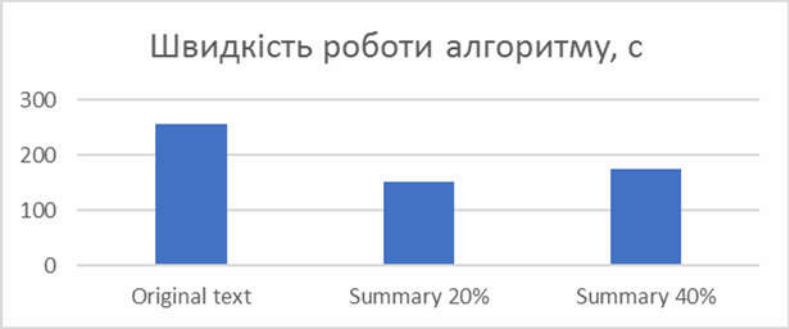

Рисунок 2 – Швидкість роботи алгоритму, с

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. Результати виявлення аномальних елементів для потоків оригінальних та реферованих даних виявили незначне відхилення.

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. 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.

3. 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.

4. 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

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. 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.

3. 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.

4. 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