УДК 004.432:004.934 DOI: https://doi.org/10.30837/IYF.CVSAMM.2024.031

ДОСЛІДЖЕННЯ МЕТОДІВ СЕМАНТИЧНОЇ КЛАСТЕРИЗАЦІЇ ДЛЯ АНАЛІЗУ НОВИН

Гончарова О.В.
Науковий керівник – к.н.т., доц. Кобилін О.А.
Харківський національний університет радіоелектроніки, каф. ІНФ,
м. Харків, Україна,
тел.: (057) 702-14-19, e-mail: oksana.honcharova@nure.ua

This work is dedicated to clustering news based on their topics and exploring methods of semantic clustering. Semantic clustering can be used to increase the number of matches and differences between text elements. For advanced semantic clustering, machine learning methods, such as cluster analysis, clustering algorithms, or neural measures that capture semantic interactions between objects, can be used. An analysis of the most popular methods of semantic clustering is conducted, along with testing their effectiveness and speed of operation. Nevertheless, remaining trends in advanced semantic clustering of text documents include the emergence of current models of deep innovation and hybrid approaches.

У сучасному світі новини займають величезну частину нашого життя, отже, семантична кластеризація новин є актуальною проблемою. Під кластеризацією розуміють розбиття даних за певними групами, які називають кластерами, в яких дані згруповано за спільними характеристиками. Завданням кластеризації є відображення множини вхідних даних у множині кластерів. Семантична кластеризація може використовуватися для покращення розуміння подібностей та відмінностей між текстовими елементами, полегшуючи подальший аналіз та обробку [1, 2]. Для виконання семантичної кластеризації можуть використовуватися методи машинного навчання, такі як кластерний аналіз, алгоритми кластеризації, або нейронні мережі, які враховують семантичні взаємодії між об’єктами [3].

Існує декілька методів семантичної кластеризації:

– кластерний аналіз (Hierarchical Clustering): використовує ієрархічний підхід для групування об’єктів. Об’єкти поступово об’єднуються в кластери відповідно до їхньої схожості. Можна використовувати різні метрики схожості, такі як евклідова відстань або косинусна схожість;

– K-Means: визначається кількість кластерів (K), і об’єкти розподіляються між ними на основі мінімізації середньої квадратичної відстані між об’єктами та центрами кластерів;

– DBSCAN (Density-Based Spatial Clustering of Applications with Noise): визначає кластери на основі щільності об’єктів у просторі. Кластери формуються там, де є висока щільність, і об’єкти, які знаходяться в менш щільних областях, розглядаються як шум чи викиди;

– Spectral Clustering: використовує властивості собствених значень та собствених векторів графа схожості об’єктів для групування. Ефективний для виявлення нелінійних залежностей та кластерів складної форми;

– Agglomerative Hierarchical Clustering: об’єднує найбільш схожі об’єкти на кожному етапі. Результатом є дерево, яке можна представити у вигляді дендрограми;

– Latent Semantic Analysis (LSA): використовує сингулярний розклад матриці для зменшення розмірності текстових даних та виявлення семантичних взаємодій;

– Word Embeddings-Based Clustering: використання векторних представлень слів (word embeddings), таких як Word2Vec або Doc2Vec, для вимірювання семантичної схожості та кластеризації текстових даних.

Для задачі кластеризації новин можна використати комбінацію кількох методів. Doc2Vec та інші методи, які враховують контекст документів, дозволяють враховувати семантику всього тексту новини при кластеризації, що робить результати більш точними та репрезентативними. Останні тенденції у дослідженні семантичної кластеризації новин включають використання моделей глибокого навчання та гібридних підходів. Завдяки цьому дослідженню відкриваються перспективи для покращення інформаційного пошуку, рекомендацій та організації новинного контенту, що відповідає вимогам сучасного інформаційного суспільства [4].

У ході експериментального дослідження для кластеризації новин було виявлено, що найшвидшим методом семантичної кластеризації є K-Means, а найбільш ефективним є Word Embeddings-Based Clustering, а саме Doc2Vec, оскільки він враховує контекст вхідних даних. Отже для швидкої та ефективної роботи застосунку семантичної кластеризації новин найліпшим буде використання комбінації цих двох методів семантичної кластеризації.

Список використаних джерел:

1. Оченашко М. О. Використання вагових коефіцієнтів для дескрипторів зображення у задачі класифікації. Proceedings of the XIV International Scientific and Practical Conference. Oslo, Norway. 2023. pp. 544–548.

2. Gorokhovatskyi V., Tvoroshenko I., Yakovleva O. Transforming image descriptions as a set of descriptors to construct classification features, Indonesian Journal of Electrical Engineering and Computer Science, 2024. 33 (1), 113–125.

3. Gorokhovatskyi, V., Vlasenko, N.. Редукція опису зображення у складі множини дескрипторів на основі метричного критерію інформативності. Advanced Information Systems, 2021. 5(4), pp. 10–16.

4. Gorokhovatskyi V., Tvoroshenko I., Kobylin O., Vlasenko N. Search for visual objects by request in the form of a cluster representation for the structural image description, Advances in Electrical and Electronic Engineering, 2023. 21 (1), pp. 19–27.