Современные задачи анализа научных публикаций требуют эффективного выявления тематической структуры больших текстовых коллекций. При использовании современных методов векторизации текста документы представляются в виде точек в высокоразмерном embedding-пространстве. Однако даже работы, относящиеся к различным подтемам одной области знания, часто оказываются близкими друг к другу. В результате пространство документов принимает вид плотного облака точек, в котором стандартные методы кластеризации затруднительно применить для выявления тонких тематических различий между исследованиями.

В задачах кластеризации текстовых embeddings используются различные классы алгоритмов. Наиболее распространены методы разбиения на кластеры, такие как K-means и его модификации, включая mini-batch K-means [4,6]. Также применяются иерархические алгоритмы кластеризации, например агломеративные методы и алгоритм BIRCH [1,4]. Помимо этого рассматриваются альтернативные подходы, включая самоорганизующиеся карты Кохонена, вероятностные модели на основе гауссовских смесей и методы графовой кластеризации [3,5,8].

Современные исследования показывают, что использование embeddings, полученных из трансформерных моделей (например, BERT или RoBERTa), позволяет существенно повысить качество кластеризации текстов по сравнению с традиционными представлениями, такими как TF-IDF [2]. Тем не менее многие существующие алгоритмы не учитывают локальную геометрическую структуру embedding-пространства.

В настоящей работе предлагается использовать топологические характеристики локальной структуры embedding-пространства для выделения внутренних и граничных точек кластеров. Для каждой точки embedding-пространства рассматривается её локальная окрестность, на которой строится Vietoris–Rips комплекс. На основе диаграмм персистентности вычисляются топологические признаки, включая числа Бетти, суммарную персистентность и персистентную энтропию Шеннона для гомологий H_0 и H_1.

На основе вычисленных топологических признаков обучается классификатор, который определяет, является ли точка внутренней точкой кластера или принадлежит его границе. После этого кластеры восстанавливаются путем распространения меток от внутренних точек к граничным точкам на основе соседства в embedding-пространстве. Для обучения классификатора использовался алго-

ритм Random Forest.

Эксперименты проводились на наборах научных публикаций, сформированных по тематическим кластерам из базы OpenAlex. Полученные результаты показывают, что предложенный подход позволяет эффективно восстанавливать структуру кластеров в embedding-пространстве. Значения метрик качества кластеризации достигали $ARI = 0.84$ для сложных наборов данных и $ARI = 0.99$ для более явно разделимых тематических кластеров. При этом алгоритм DBSCAN на тех же данных показывал значительно более низкие результаты.

Литература

1. Abdalgader K. Hierarchical clustering approaches for document analysis. 2024.
2. Alagöz F. Transformer-based embeddings for text clustering. 2024.
3. Boutalbi R. Graph-based clustering using tensor representations. 2022.
4. Costa A. et al. Text clustering methods comparison. 2023.
5. Miller J. Gaussian mixture models for high-dimensional text embeddings. 2024.
6. Simanjuntak H. et al. Mini-batch K-means for text clustering. 2023.
7. Vedmiediev M. et al. Evaluation of clustering algorithms for text embeddings. 2023.
8. Wang X. Self-organizing maps for document clustering. 2023.