

**Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«Национальный исследовательский университет ИТМО»
(Университет ИТМО)**

Факультет Программной инженерии и компьютерной техники

**Образовательная программа Проектирование и разработка систем
больших данных**

**Направление подготовки (специальность) 09.04.04 Программная
инженерия**

О Т Ч Е Т

о Научно-исследовательской работе

**Тема задания: Обзор предметной области тематики выпускной
квалификационной работы**

Обучающийся: Постнов Степан Андреевич, Р4135

**Руководитель практики от университета: Маркина Татьяна Анатольевна,
старший преподаватель факультета программной инженерии и
компьютерной техники**

Дата 29.12.2025

**Санкт-Петербург
2025**

СОДЕРЖАНИЕ

| | |
|--|----|
| ВВЕДЕНИЕ | 3 |
| МЕТОД ПОИСКА ИСТОЧНИКОВ | 7 |
| НАПИСАНИЕ АНАЛИТИЧЕСКОГО ОБЗОРА | 10 |
| СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ | 12 |
| ПРИЛОЖЕНИЕ 1: СПИСОК ИСТОЧНИКОВ ДЛЯ АНАЛИТИЧЕСКО- ГО ОБЗОРА | 16 |
| ПРИЛОЖЕНИЕ 2. АНАЛИТИЧЕСКИЙ ОБЗОР | 19 |

ВВЕДЕНИЕ

Целью проведения практики является выполнение систематического обзора предметной области, связанной с методами кластеризации для определения семантической близости и совместимости слов в поисковых запросах. В ходе практики предполагается провести анализ научных работ за период с 2010 по 2025 год, выявить ключевые подходы, выделить актуальные проблемы и перспективные направления исследований. Результаты обзора будут использованы для обоснования выбора методологии и определения фокуса дальнейшей выпускной квалификационной работы.

Данный отчёт описывает выполнение всех этапов практики в соответствии с индивидуальным заданием, представленным в таблице 1.

Таблица 1 – Индивидуальное задание на практику

| № этапа | Название этапа | Задание |
|---------|---|--|
| 1 | Инструктаж обучающегося | Инструктаж обучающегося по ознакомлению с требованиями охраны труда, техники безопасности, пожарной безопасности, а также правилами внутреннего трудового распорядка. |
| 2 | Заполнение и утверждение задания | Заполнить в разделе Практика задание на НИР и отправить на согласование руководителю практики. |
| 3 | Составить список источников для аналитического обзора | Источники должны быть не старше 5 лет. Допускаются источники старше 5 лет, но они должны быть в меньшем количестве. В качестве источников могут быть статьи, тезисы, книги, учебные пособия и др. Минимальное количество источников: 7. Должен быть оформлен в соответствии с ГОСТ 7.32-2017. |

Продолжение таблицы 1

| № этапа | Название этапа | Задание |
|----------------|--|--|
| 4 | Написать аналитический обзор по тематике ВКР | <p>Провести анализ научных работ по методам кластеризации за период с 2010 по 2025 год с целью систематизации существующих знаний, выявления актуальных проблем и перспективных направлений исследований. Систематизировать информацию из не менее 20 научных статей. Выделить не менее 3 ключевых актуальных проблем в области применения алгоритмов кластеризации с целью определения семантической близости и совстречаемости слов в поисковых запросах.</p> <p>Определить не менее 3 перспективных направлений исследований, подкрепленных обоснованием и ссылками на релевантные работы. Оформить результаты анализа в виде обзора литературы объемом 15 страниц. Анализ будет проведен с использованием доступных научных баз данных (PubMed, Scopus, Web of Science, Google Scholar, Wikipedia, Arxiv, eLibrary, РИНЦ и документаций open-source проектов) и ресурсов библиотеки университета. Результаты анализа позволят нашей исследовательской группе определить наиболее перспективные направления для дальнейших исследований в области кластеризации с целью определения семантической близости и совстречаемости слов в поисковых запросах, что приведет к разработке новых алгоритмов для распределенных систем, которые позволят ускорить обработку поисковых запросов.</p> |

Продолжение таблицы 1

| № этапа | Название этапа | Задание |
|--------------------|---------------------------|--|
| | | <p>Анализ должен быть завершен к 11 января 2026 года.</p> <p>Требования к аналитическому обзору: - Обзор не менее 20–30 научных источников (Scopus, WoS, РИНЦ, патенты, arxiv.org). - Классификация подходов (таблица или схема). - Сравнительная таблица (методы, преимущества, недостатки). - Выявление тенденций и "белых пятен"(что уже изучено, а что требует исследований). - Критический анализ (противоречия в литературе, применимость к тематике). - Описание актуальности, включающая в себя обоснование выбора темы ВКР (почему эта проблема важна) или почему не стоит этим заниматься. - Выводы (на чем следует сосредоточиться в работе). Структура аналитического обзора (научно-исследовательский вариант): 1) Введение в тему (2-4 стр.) - Определение области исследования. - Исторический контекст (учесть необходимую глубину исследования, текущее состояние решения проблемы не обязательно state of the art). - Актуальные вызовы. 2) Таксономия подходов (4-8 стр.) - Систематизированная классификация существующих методов, технологий и теорий по заданным критериям, позволяющая наглядно представить структуру исследуемой области. - Хронологическая таблица развития методов. - Сравнительная таблица (методы, преимущества, недостатки). - Классификационная схема.</p> |

Продолжение таблицы 1

| № этапа | Название этапа | Задание |
|----------------|---|--|
| | | <p>3) Критический анализ (2-4 стр.) - Противоречия в литературе. - Выявление тенденций и "белых пятен". - Методологические ограничения. - Неисследованные аспекты. - Применимость к тематике.</p> <p>4) Описание актуальности (2-4 стр.) - Обоснование выбора темы ВКР (почему эта проблема важна) или почему не стоит этим заниматься.</p> <p>5) Перспективные направления (1-2 стр.) - Тренды последних 2 лет. - Рекомендации для будущих исследований.</p> <p>6) Выводы (на чем следует сосредоточиться в работе).</p> <p>(1-2 стр.)</p> <p>7) Список использованных источников</p> |
| 5 | Оформление отчёта, получение отзыва с оценкой от руководителя | <p>Написание отчёта, оформление в соответствии с требованиями, получение отзыва от руководителей.</p> <p>Оформление отчётных документов в соответствии с требованиями:</p> <p>1. Оформление отчёта должно быть выполнено в соответствии с методическим пособием (https://books.ifmo.ru/file/pdf/2622.pdf, разделы "Набор текста" "Нумерация страниц и разделов" "Заголовки разделов" "Оформление рисунков" "Оформление таблиц" "Оформление формул" "Оформление ссылок на источники" "Оформление списка использованных источников" "Оформление приложений").</p> <p>2. Структура документа: титульный лист, введение, основная часть, заключение, приложения.</p> <p>3. В основной части подробно описывается выполнение задач всех этапов. Получить отзыв с оценкой у руководителя практики в модуле Практика.</p> |

МЕТОД ПОИСКА ИСТОЧНИКОВ

В ходе выполнения работы я сформировал список источников на основе систематического поиска в научных базах данных и репозиториях. Поиск проводился итеративно, так как в первую очередь я собрал базовые обзорные работы по алгоритмам кластеризации и их применению, затем расширил выборку специализированными публикациями по кластеризации текстов/запросов и графовым подходам, а в конце проверил, что итоговый набор покрывает период 2010–2025 гг. и содержит достаточно число современных источников. В силу того, что методы кластеризации впервые были описаны достаточно давно, некоторые ранние источники пришлось оставить в силу их полноты и фундаментальности.

Основными поисковыми ресурсами и базами данных выступали:

- Google Scholar — как основной инструмент для подбора статей и использования умного поиска (сайт поддерживает поиск с использованием нейронных сетей);
- arXiv — для поиска свежих обзоров и препринтов по кластеризации и масштабируемым алгоритмам;
- eLibrary/РИНЦ — для русскоязычных публикаций по кластеризации и применению в информационном поиске;
- IEEE Xplore — для инженерно-ориентированных обзоров и прикладных работ;
- Deepseek — для поиска релевантных научных статей с использованием ИИ.

Я использовал как русские, так и английские запросы, чтобы не потерять релевантные источники и терминологические варианты. Используемые запросы и их результаты:

- eLibrary: «алгоритмы кластеризации поисковых запросов (последние 10 лет)», «обзор методов кластеризации», «индекс поисковых запросов на основе кластеризации» (найденные источники: [1; 2; 3; 4]);

- Google Scholar: «*community detection*», «*overview of clustering algorithms*», «*k-means and other clustering algorithms*» (найденные источники: [5; 6; 7; 8; 9; 10; 11; 12; 13; 14; 23; 24]);
- Deepseek: «*semantic search clustering articles*», «*clustering scientific articles*» (найденные источники: [18; 19; 20; 22]).

Также от научного руководителя были получены следующие источники, которые были добавлены в работу — [15; 16; 17; 21].

На этапе первичного отбора я задавал временной фильтр 2010–2025 гг., а также отдельно просматривал последние 2–5 лет, чтобы обеспечить актуальность обзора. В итоговый список я включал источники, которые удовлетворяют следующим условиям:

- соответствие теме (кластеризация, детектирование сообществ, масштабирование алгоритмов, кластеризация текстов/запросов, графовые методы, семантический индекс);
- наличие чётко описанного метода/таксономии/ограничений (для обзорных работ) либо экспериментальной части и метрик оценки (для прикладных работ);
- авторитетность и возможность корректно идентифицировать библиографические данные.

В процессе поиска и написания отчета мной были исключены следующие источники:

1. Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review. // ACM Computing Surveys. – 1999. — не отражает современные протоколы оценки и масштабируемые сценарии коротких текстов (до 2010 г.) [30].
2. Shi J., Malik J. Normalized cuts and image segmentation. // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2000. — ориентирована на сегментацию изображений; методологически значима, но не содержит валидации на коротких текстах/поисковых запросах (до 2010 г.) [31].
3. Ng A. Y., Jordan M. I., Weiss Y. On spectral clustering: Analysis and an algorithm. // Advances in Neural Information Processing Systems (NIPS).

- 2002. — отсутствует прикладная проверка на задачах кластеризации поисковых запросов (до 2010 г.) [32].
4. Girvan M., Newman M. E. J. Community structure in social and biological networks. // Proceedings of the National Academy of Sciences (PNAS). – 2002. — не рассматривает текстовые представления и совстречаемость терминов в запросах [33].
 5. Newman M. E. J. Fast algorithm for community detection in networks. // Physical Review E. – 2004. — отсутствуют эксперименты для коротких текстов (до 2010 г.) [34].
 6. Lloyd S. Least squares quantization in PCM. // IEEE Transactions on Information Theory. – 1982. — историческая основа k -means; устарела для целей обзора и не отражает современные ограничения для разреженных текстов (до 2010 г.) [35].
 7. MacQueen J. Some methods for classification and analysis of multivariate observations. // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. – 1967. — исторический ввод k -means; значительно старше целевого периода, отсутствуют релевантные метрики для коротких текстов (до 2010 г.) [36].
 8. Shi H., Liang Y., Wang J. Spectral clustering for community detection: A review. // IEEE Access. – 2018. — обзор сосредоточен на графах социальной сети; не содержит протоколов для семантической кластеризации поисковых запросов [37].
 9. Ding C., He X. K-means clustering via principal component analysis. // Proceedings of ICML. – 2004. — отсутствует проверка на коротких запросах/разреженных признаках (до 2010 г.) [38].
 10. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. // Information Processing & Management. – 1988. — классическая работа по взвешиванию термов; существенно вне целевого периода, не фокусируется на кластеризации запросов (до 2010 г.) [39].

НАПИСАНИЕ АНАЛИТИЧЕСКОГО ОБЗОРА

На основе рекомендаций научного руководителя и предварительного анализа темы работы, в качестве типа обзора был выбран научно-исследовательский обзор. Данный формат ориентирован на систематизацию существующих методов, выявление ограничений и перспективных направлений исследований, что наиболее полно соответствует задачам выпускной квалификационной работы.

Структура обзора формировалась на основе анализа требований индивидуального задания и рекомендаций по оформлению научно-исследовательских обзоров. Данная структура позволяет последовательно перейти от общего описания проблематики к частным вопросам применения методов кластеризации в распределённых системах поиска.

В процессе подготовки обзора возникали следующие методические и содержательные сложности: в литературе по кластеризации и по информационному поиску используются различные термины для близких понятий (например, «similarity», «proximity», «co-occurrence»), часть базовых алгоритмов (например, k -means, иерархическая кластеризация) описана в классических работах вне целевого периода, первоначально отбор литературы дал более 40 потенциально релевантных публикаций, что превышало требуемый объём. На завершающем этапе подготовки обзора был проведён дополнительный поиск специализированных публикаций, которые позволили усилить обоснование выбора графовых методов и уточнить метрики оценки. В список были включены следующие источники:

- Обзорная статья по детектированию сообществ в графах, которая систематизирует семейство алгоритмов на основе оптимизации модульности и близких критериев. Работа детально описывает концепцию сообществ как плотных подграфов с минимальным числом внешних связей, что напрямую применимо к задаче выделения семантических групп терминов [25].
- Публикация, описывающая алгоритм Leiden, который устраняет известные недостатки метода Louvain (в частности, проблему образования несвязных сообществ). Leiden обеспечивает гарантии связности выделяемых кластеров и демонстрирует высокую масштабируемость на разре-

женных графах с миллионами вершин [26].

- Методологическая работа, посвящённая нормализованной попарной взаимной информации (NPMI) в задаче извлечения коллокаций [27].
- Научная работа, вводящая метрику силуэта (silhouette) для оценки качества кластеризации [28].
- Публикация, фокусирующаяся на эмбеддингах, а не на кластеризации, которая необходима для определения векторных подходов [29].

Включение данных источников позволило усилить методологическую базу обзора, обосновать центральную роль графовых методов и заложить основу для дальнейшего исследования гибридных подходов.

В результате написания обзора была сформирована целостная картина методов кластеризации (применительно к поисковым запросам), выделены актуальные проблемы (разреженность, шум, масштабируемость), проведён критический анализ ограничений и предложены конкретные направления для дальнейшей работы. Итоговый отчет включает 29 основных источников (в том числе обзорные статьи последних 2–5 лет) и 10 исключённых источников. Структура обзора соответствует требованиям научно-исследовательского формата и позволяет обосновать выбор алгоритма Leiden для реализации в рамках выпускной квалификационной работы.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 Бова В. В., Кравченко Ю. А., Родзин С. И. Методы и алгоритмы кластеризации текстовых данных (обзор). // Известия ЮФУ. Технические науки. – 2022. – № 4 (228). – С. 122–143.
- 2 Черникова Д. А. Алгоритм кластеризации поисковых запросов. // Евразийский научный журнал. – 2017. – № 12.
- 3 Миронов А. И., Мунерман В. И. Создание частичного индексирования таблицы для оптимизации поисковых запросов. // Современные информационные технологии и ИТ-образование. – 2022. – Т. 18, № 3. – С. 558–565.
- 4 Люнченко С. Применение методов кластеризации для управления запасами товарно-материальных ценностей. // Евразийский союз ученых. – 2020. – № 4-4 (73). – С. 29–37.
- 5 Курейчик В. В., Герасименко П. С. Основные подходы к извлечению текстовой информации (обзор). // Известия ЮФУ. Технические науки. – 2024. – № 4 (240). – С. 6–14.
- 6 Pitafi S., Anwar T., Sharif Z. A taxonomy of machine learning clustering algorithms, challenges, and future realms. // Applied Sciences. – 2023. – Т. 13, № 6. – С. 3529.
- 7 Wani A. A. Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions. // PeerJ Computer Science. – 2024. – Т. 10. – e2286.
- 8 Mahnoor, Shafi I., Chaudhry M., Caro Montero E., Silva Alvarado E., de la Torre Diez E., Abdus Samad M., Ashraf I. A Review of Approaches for Rapid Data Clustering: Challenges, Opportunities, and Future Directions. // IEEE Access. – 2024. – Т. 12. – С. 138086–138120.
- 9 Miraftabzadeh S. M., Colombo C. G., Longo M., Foiadelli F. K-Means and Alternative Clustering Methods in Modern Power Systems. // IEEE Access. – 2023. – Т. 11. – С. 119596–119633.
- 10 Alasali T., Ortakcı Y. Clustering Techniques in Data Mining: A Survey of Methods, Challenges, and Applications. // Computer Science. – 2024. – Т. 9, № 1. – С. 32–50.
- 11 Oyelade J., Isewon I., Oladipupo O., Emebo O., Omogbadegun Z., Aromolaran O., Uwoghiren E., Olaniyan D., Olawole O. Data clustering: Algorithms and its applications. // Proceedings of the 2019 19th International Conference on

Computational Science and Its Applications (ICCSA). – 2019. – C. 71–81.

12 Xu D., Tian Y. A comprehensive survey of clustering algorithms. // Annals of Data Science. – 2015. – T. 2, № 2. – C. 165–193.

13 Aggarwal C. C., Reddy C. K. Data clustering. // Algorithms and applications. Chapman&Hall/CRC Data Mining and Knowledge Discovery Series. – 2014.

14 Ezugwu A. E., Shukla A. K., Agbaje M. B., Oyelade O. N., José-García A., Agushaka J. O. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. // Neural Computing and Applications. – 2021. – T. 33, № 11. – C. 6247–6306.

15 Ezugwu A. E. Nature-inspired metaheuristic techniques for automatic clustering: a survey and performance study. // SN Applied Sciences. – 2020. – T. 2, № 2. – C. 273.

16 Shahid N. Comparison of hierarchical clustering and neural network clustering: an analysis on precision dominance. // Scientific Reports. – 2023. – T. 13, № 1. – C. 5661.

17 Bushra A. A., Yi G. Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms. // IEEE Access. – 2021. – T. 9. – C. 87918–87935.

18 Nagpal A., Jatain A., Gaur D. Review based on data clustering algorithms. // In: 2013 IEEE Conference on Information & Communication Technologies. – 2013. – C. 298–303.

19 Guyeux C., Chrétien S., Bou Tayeh G., Demerjian J., Bahi J. Introducing and comparing recent clustering methods for massive data management in the Internet of Things. // Journal of Sensor and Actuator Networks. – 2019. – T. 8, № 4. – C. 56.

20 Ahmad A., Khan S. S. Survey of state-of-the-art mixed data clustering algorithms. // IEEE Access. – 2019. – T. 7. – C. 31883–31902.

21 Fahad A., Alshatri N., Tari Z., Alamri A., Khalil I., Zomaya A. Y., Foufou S., Bouras A. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. // IEEE Transactions on Emerging Topics in Computing. – 2014. – T. 2, № 3. – C. 267–279.

22 Wegmann M., Zipperling D., Hillenbrand J., Fleischer J. A review of systematic selection of clustering algorithms and their evaluation. // arXiv preprint arXiv:2106.12792. – 2021.

- 23 Nasraoui O., N'Cir C-E Ben. Clustering methods for big data analytics. // Techniques, Toolboxes and Applications. – 2019. – T. 1. – C. 91–113.
- 24 Reddy C. K., Vinzamuri B. A survey of partitional and hierarchical clustering algorithms. // In: Data clustering. – 2018. – C. 87–110.
- 25 Fortunato S. Community detection in graphs. // Physics Reports. – 2010. – T. 486, № 3–5. – P. 75–174.
- 26 Traag V. A., Waltman L., van Eck N. J. From Louvain to Leiden: guaranteeing well-connected communities. // Scientific Reports. – 2019. – T. 9. – Article 5233.
- 27 Bouma G. Normalized (pointwise) mutual information in collocation extraction. // Proceedings of GSCL. – 2009.
- 28 Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. // Journal of Computational and Applied Mathematics. – 1987. – T. 20. – P. 53–65.
- 29 Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. // arXiv preprint arXiv:1301.3781. – 2013.
- 30 Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review. // ACM Computing Surveys. – 1999.
- 31 Shi J., Malik J. Normalized cuts and image segmentation. // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2000.
- 32 Ng A. Y., Jordan M. I., Weiss Y. On spectral clustering: Analysis and an algorithm. // Advances in Neural Information Processing Systems. – 2002.
- 33 Girvan M., Newman M. E. J. Community structure in social and biological networks. // Proceedings of the National Academy of Sciences. – 2002.
- 34 Newman M. E. J. Fast algorithm for community detection in networks. // Physical Review E. – 2004.
- 35 Lloyd S. Least squares quantization in PCM. // IEEE Transactions on Information Theory. – 1982.
- 36 MacQueen J. Some methods for classification and analysis of multivariate observations. // Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability. – 1967.
- 37 Shi H., Liang Y., Wang J. Spectral clustering for community detection: A review. // IEEE Access. – 2018.
- 38 Ding C., He X. K-means clustering via principal component analysis. // Proceedings of the 21st International Conference on Machine Learning (ICML). –

2004.

39 Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. // Information Processing & Management. – 1988.

ПРИЛОЖЕНИЕ 1: СПИСОК ИСТОЧНИКОВ ДЛЯ АНАЛИТИЧЕСКОГО ОБЗОРА

- 1 Бова В. В., Кравченко Ю. А., Родзин С. И. Методы и алгоритмы кластеризации текстовых данных (обзор). // Известия ЮФУ. Технические науки. – 2022. – № 4 (228). – С. 122–143.
- 2 Черникова Д. А. Алгоритм кластеризации поисковых запросов. // Евразийский научный журнал. – 2017. – № 12.
- 3 Миронов А. И., Мунерман В. И. Создание частичного индексирования таблицы для оптимизации поисковых запросов. // Современные информационные технологии и ИТ-образование. – 2022. – Т. 18, № 3. – С. 558–565.
- 4 Люнченко С. Применение методов кластеризации для управления запасами товарно-материальных ценностей. // Евразийский союз ученых. – 2020. – № 4-4 (73). – С. 29–37.
- 5 Курейчик В. В., Герасименко П. С. Основные подходы к извлечению текстовой информации (обзор). // Известия ЮФУ. Технические науки. – 2024. – № 4 (240). – С. 6–14.
- 6 Pitafi S., Anwar T., Sharif Z. A taxonomy of machine learning clustering algorithms, challenges, and future realms. // Applied Sciences. – 2023. – Т. 13, № 6. – С. 3529.
- 7 Wani A. A. Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions. // PeerJ Computer Science. – 2024. – Т. 10. – e2286.
- 8 Mahnoor, Shafi I., Chaudhry M., Caro Montero E., Silva Alvarado E., de la Torre Diez E., Abdus Samad M., Ashraf I. A Review of Approaches for Rapid Data Clustering: Challenges, Opportunities, and Future Directions. // IEEE Access. – 2024. – Т. 12. – С. 138086–138120.
- 9 Miraftabzadeh S. M., Colombo C. G., Longo M., Foiadelli F. K-Means and Alternative Clustering Methods in Modern Power Systems. // IEEE Access. – 2023. – Т. 11. – С. 119596–119633.
- 10 Alasali T., Ortakcı Y. Clustering Techniques in Data Mining: A Survey of Methods, Challenges, and Applications. // Computer Science. – 2024. – Т. 9, № 1. – С. 32–50.
- 11 Oyelade J., Isewon I., Oladipupo O., Emebo O., Omogbadegun Z., Aromolaran O., Uwoghiren E., Olaniyan D., Olawole O. Data clustering: Algorithms

and its applications. // Proceedings of the 2019 19th International Conference on Computational Science and Its Applications (ICCSA). – 2019. – C. 71–81.

12 Xu D., Tian Y. A comprehensive survey of clustering algorithms. // Annals of Data Science. – 2015. – T. 2, № 2. – C. 165–193.

13 Aggarwal C. C., Reddy C. K. Data clustering. // Algorithms and applications. Chapman&Hall/CRC Data Mining and Knowledge Discovery Series. – 2014.

14 Ezugwu A. E., Shukla A. K., Agbaje M. B., Oyelade O. N., José-García A., Agushaka J. O. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. // Neural Computing and Applications. – 2021. – T. 33, № 11. – C. 6247–6306.

15 Ezugwu A. E. Nature-inspired metaheuristic techniques for automatic clustering: a survey and performance study. // SN Applied Sciences. – 2020. – T. 2, № 2. – C. 273.

16 Shahid N. Comparison of hierarchical clustering and neural network clustering: an analysis on precision dominance. // Scientific Reports. – 2023. – T. 13, № 1. – C. 5661.

17 Bushra A. A., Yi G. Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms. // IEEE Access. – 2021. – T. 9. – C. 87918–87935.

18 Nagpal A., Jatain A., Gaur D. Review based on data clustering algorithms. // In: 2013 IEEE Conference on Information & Communication Technologies. – 2013. – C. 298–303.

19 Guyeux C., Chrétien S., Bou Tayeh G., Demerjian J., Bahi J. Introducing and comparing recent clustering methods for massive data management in the Internet of Things. // Journal of Sensor and Actuator Networks. – 2019. – T. 8, № 4. – C. 56.

20 Ahmad A., Khan S. S. Survey of state-of-the-art mixed data clustering algorithms. // IEEE Access. – 2019. – T. 7. – C. 31883–31902.

21 Fahad A., Alshatri N., Tari Z., Alamri A., Khalil I., Zomaya A. Y., Foufou S., Bouras A. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. // IEEE Transactions on Emerging Topics in Computing. – 2014. – T. 2, № 3. – C. 267–279.

22 Wegmann M., Zipperling D., Hillenbrand J., Fleischer J. A review of systematic selection of clustering algorithms and their evaluation. // arXiv preprint

arXiv:2106.12792. – 2021.

23 Nasraoui O., N'Cir C-E Ben. Clustering methods for big data analytics. // Techniques, Toolboxes and Applications. – 2019. – T. 1. – C. 91–113.

24 Reddy C. K., Vinzamuri B. A survey of partitional and hierarchical clustering algorithms. // In: Data clustering. – 2018. – C. 87–110.

ПРИЛОЖЕНИЕ 2. АНАЛИТИЧЕСКИЙ ОБЗОР

СОДЕРЖАНИЕ

| | | |
|-------|---|----|
| 1 | Введение в тему | 20 |
| 1.1 | Исторический контекст развития | 21 |
| 1.2 | Объект исследования и постановка задачи | 21 |
| 1.3 | Рабочая гипотеза исследования | 21 |
| 2 | Таксономия подходов к кластеризации | 22 |
| 2.1 | Описание существующих подходов | 22 |
| 2.2 | Хронологическая таблица развития исследований | 23 |
| 2.3 | Классификация подходов | 24 |
| 2.3.1 | Векторный подход | 26 |
| 2.3.2 | Иерархический подход | 27 |
| 2.3.3 | Плотностный подход | 27 |
| 2.3.4 | Графовый подход | 27 |
| 2.3.5 | Смешанные данные | 28 |
| 2.4 | Сравнение подходов и методов | 28 |
| 3 | Критический анализ | 30 |
| 3.1 | Противоречия в литературе и методологические ограничения | 30 |
| 3.2 | Выявленные тенденции | 31 |
| 3.3 | «Белые пятна» и неисследованные аспекты для поисковых запросов | 31 |
| 3.4 | Актуальные проблемы применения кластеризации в поисковых запросах | 32 |
| 4 | Описание актуальности исследования и обоснование выбора темы | 33 |
| 4.1 | Актуальность с точки зрения распределённых систем | 33 |
| 4.2 | Актуальность с точки зрения разработки и инфраструктуры | 33 |
| 4.3 | Актуальность с точки зрения пользователя | 34 |
| 4.4 | Актуальность с точки зрения аналитики | 34 |
| 4.5 | Риски и ограничения темы | 34 |
| 5 | Перспективные направления исследований | 35 |
| 5.1 | Тренды последних 2 лет | 35 |
| 5.2 | Рекомендации для будущих исследований | 36 |
| | ВЫВОДЫ | 37 |
| | СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ | 39 |

1 Введение в тему

Кластеризация относится к базовым методам интеллектуального анализа данных и применяется для выявления устойчивых групп объектов без использования обучающей разметки. В области информационного поиска методы кластеризации находят применение для нескольких целей.

1. Группировка поисковых запросов по намерению пользователя (*intent*) и тематике [2].
2. Построение тематических и семантических групп для улучшения качества подсказок и расширения запросов.
3. Анализ совстречаемости слов и фраз для выявления скрытых ассоциаций.
4. Оптимизация структуры индекса в распределённых системах путём семантической локализации данных [3].

Поисковые запросы, в отличие от полнотекстовых документов, характеризуются рядом специфических особенностей:

- малая длина (типичная длина 2–4 слова, редко больше 10 слов);
- высокая шумность (опечатки, сокращения, нестандартные написания);
- морфологическая вариативность (особенно для русского языка);
- смешение языков в одном запросе;
- неоднозначность и контекстная зависимость формулировок.

Указанные особенности существенно усложняют построение моделей сходства и повышают роль предварительной обработки (очистка, нормализация, токенизация) в качестве определяющего фактора успеха кластеризации. В то же время кластеризация широко применяется и в других прикладных сценариях, что подчёркивает универсальность подходов [4; 11].

1.1 Исторический контекст развития

Историческое развитие исследований по кластеризации в рассматриваемый период целесообразно трактовать как последовательное расширение требований к задаче и уточнение методологии. Начиная с 2010-х годов, фокус смещается к практической процедуре применения и решению вопросов выбора параметров, устойчивой работы с разреженными и смешанными данными, переносу подходов в потоковые и индустриальные сценарии. Отсюда растёт интерес к автоматическому подбору параметров и к методам, ориентированным на массовые объёмы и ограниченные ресурсы [14; 15; 19; 20; 23].

В 2021–2025 гг. заметно усиливается методологическая составляющая, поэтому возрастает роль протоколов сравнения, анализа ограничений и требований к воспроизводимости. Для кластеризации поисковых запросов эта линия важна особенно, поскольку качество результата существенно зависит от предобработки и протокола оценки, а также от масштаба данных. Поэтому вместо категоричных утверждений «метод X лучше» чаще предлагаются рамки систематического выбора и проверки применимости, а также обзоры быстрых прикладных решений [6; 7; 8; 22].

1.2 Объект исследования и постановка задачи

Объектом исследования в рамках данного обзора является кластеризация поисковых запросов для целей определения семантической близости и совстречаемости. Под семантической близостью в данном контексте понимается практическая близость слов/запросов, проявляющаяся через их совместное употребление в одном и том же пользовательском контексте (в пределах запроса, сессии или локального временного окна). Под совстречаемостью понимается статистически значимая совместная встречаемость слов, которая не может быть объяснена только высокой частотностью каждого слова отдельно (для чего используются меры типа NPMI, PPMI, PPMI²). Мера NPMI рассматривается как нормированная модификация PMI, удобная для сравнения связей разной частотности [27].

1.3 Рабочая гипотеза исследования

Рабочая гипотеза, лежащая в основе данного исследования, состоит в том, что представление данных в виде взвешенного графа совстречаемости

терминов, где вес ребра отражает статистическую значимость связи между словами, позволяет применять графовые алгоритмы детектирования сообществ (community detection algorithms) для получения семантически интерпретируемых и вычислительно масштабируемых кластеров. Ожидаемые преимущества:

- интерпретируемость кластеров как сообществ связных терминов;
- использование результатов для построения распределённого полнотекстового индекса.

2 Таксономия подходов к кластеризации

2.1 Описание существующих подходов

В литературе подходы к кластеризации обычно классифицируют по нескольким независимым осям. На практике наиболее распространены три критерия. Обобщающие обзоры методов и их ограничений приводятся, например, в работах [10].

1. Представление данных (вектор, матрица расстояний/сходства, граф, смешанные признаки).
2. Семейство алгоритмов (разбиение на части, иерархические подходы, плотностные методы, методы детектирования сообществ и др.).
3. Ограничения применения (масштабируемость, требования к памяти/времени, воспроизводимость, интерпретируемость и устойчивость к шуму).

Для задач семантики поисковых запросов ключевым оказывается именно представление данных. Причина состоит в том, что запросы коротки и разрежены, а сходство между ними (или между терминами) формируется не напрямую, а через набор допущений о признаках и метрике. Поэтому далее используется практико-ориентированная группировка, где тип представления данных задаёт основной контекст, а семейство алгоритмов рассматривается как уровень детализации. В рамках обзора для дальнейшего изложения выделяются следующие базовые подходы.

1. Векторный подход: разреженные признаки (TF-IDF, n -граммы) и плотные представления (эмбеддинги).

2. Иерархический подход: матрица расстояний/сходства и построение дендрограммы с выбором уровня разбиения.
3. Плотностный подход: выделение кластеров как областей высокой плотности и отделение шума.
4. Графовый подход: граф совстречаемости и граф сходства, где вес ребра является явной мерой связи.
5. Подход для смешанных данных: совместное использование текстовых, категориальных и поведенческих сигналов.

2.2 Хронологическая таблица развития исследований

В таблице 2 представлена хронологическая динамика за последние 15 лет.

Таблица 2 – Хронологическая динамика исследований по кластеризации за последние 15 лет

| Период | Доминирующие подходы/темы | Ключевые источники (примеры) |
|---------------|---|-------------------------------------|
| 2010–2015 | Фундаментальные обзоры и таксономии; классические алгоритмы; первые систематизации для big data | [12; 13; 18; 21] |
| 2016–2020 | Усиление фокуса на big data и mixed-data; автоматический подбор параметров; перенос методов в прикладные домены и IoT | [15; 19; 20; 23] |
| 2021–2025 | Систематический выбор и протоколы оценки; анализ ограничений; обзоры быстрых/прикладных решений | [6; 7; 8; 14; 22] |

2.3 Классификация подходов

Классификация подходов к кластеризации в задачах семантической близости и совстречаемости представлена в таблице 3.

Таблица 3 – Классификация подходов к кластеризации в задачах семантической близости и совстречаемости

| Подход | Представление | Замечания по применимости к запросам |
|-----------------------|-----------------------------|--|
| Векторный (частотный) | TF-IDF, n-граммы | Методы простые и быстрые, но чувствительны к разреженности и коротким текстам; требовательны к выбору метрики [7; 12] |
| Иерархический | Матрица сходства/расстояния | Методы формируют иерархию тематик, однако плохо масштабируются на больших выборках и чувствительны к метрике и критерию связи [16; 22; 24] |
| Плотностный | Плотность в метрике | Методы способны выделять шум и кластеры произвольной формы, но параметро-чувствительны и ухудшаются в высокой размерности [17] |

Продолжение таблицы 3

| Подход | Представление | Замечания по применимости к запросам |
|------------------|---|--|
| Графовый | Граф совстречаемости/сходства | Методы естественны для моделирования совстречаемости; критичны определение веса связи и фильтрация шумовых рёбер [6; 21] |
| Смешанные данные | Текст + категориальные / поведенческие признаки | Методы перспективны для поисковых сценариев, но требуют корректной интеграции и балансировки сигналов [20] |

2.3.1 Векторный подход

Векторные методы представляют объект (запрос, документ или термин) как вектор признаков. Наиболее распространены разреженные признаки на основе частот (TF-IDF, n -граммы) и плотные векторные представления (эмбеддинги). Современные подходы к обучению эмбеддингов для слов и коротких контекстов описаны, например, в [29]. TF-IDF интерпретируется как произведение частоты терма в объекте на обратную частоту документов в корпусе, что снижает вклад общеупотребительных слов. n -граммные признаки позволяют частично учитывать локальный контекст, но при малой длине запроса часто приводят к росту разреженности.

Сильная сторона векторного подхода — простота и высокая скорость базовых операций (вычисление сходства, работа с линейными моделями), что делает его удобным базовым уровнем [12; 13]. Ключевое ограничение для поисковых запросов связано с разреженностью и шумом. При длине 2–4 слова многие объекты оказываются почти неразличимы при стандартных метриках, а небольшой сдвиг предобработки (лемматизация, фильтрация стоп-слов, обра-

ботка опечаток) меняет геометрию пространства. Это повышает чувствительность методов, работающих в векторном пространстве [7; 21].

2.3.2 Иерархический подход

Иерархические методы формируют дерево вложенных кластеров (дендограмму), что позволяет выбирать уровень детализации постфактум. Как правило, на вход подаётся матрица расстояний или сходства, рассчитанная по выбранному представлению (векторному или иному). Достоинство подхода — диагностическая ценность и возможность анализа структуры на разных масштабах [24]. Ключевые ограничения — высокая вычислительная стоимость при прямой реализации ($O(n^2)$ по памяти/времени) и сильная зависимость результата от выбранного критерия связи (linkage) и метрики. Поэтому в задачах с большими логами запросов иерархические методы обычно применимы либо на выборках, либо в сочетании с предварительным сжатием/агрегацией данных.

2.3.3 Плотностный подход

Плотностные методы (например, DBSCAN) выделяют кластеры как связанные области повышенной плотности в выбранной метрике и отделяют шум как «редкие» точки. Их достоинства — отсутствие необходимости заранее задавать число кластеров и способность находить кластеры произвольной формы [17]. Однако на высоких размерностях и разреженностих плотность становится менее информативной (эффект «проклятия размерности»), что ухудшает устойчивость метода.

2.3.4 Графовый подход

Графовые подходы моделируют объекты как вершины графа, а их связи — как взвешенные рёбра. В отличие от векторных методов, мера близости задаётся явно через вес ребра и может быть построена как по совстречаемости, так и по сходству эмбеддингов. Для исследуемой постановки (совстречаемость терминов в запросах) графовая модель является естественной, так как статистически значимые совместные появления образуют структуру связей. На таких графах применяются алгоритмы детектирования сообществ (семейство Louvain/Leiden, label propagation и др.). Классические алгоритмы детектирования сообществ Louvain и Leiden подробно описаны в [25; 26]. Сильные

стороны — масштабируемость на разреженных графах и интерпретируемость сообществ как групп взаимосвязанных терминов [21]. Основные риски — зависимость результата от построения графа, нормирования весов и фильтрации слабых рёбер [6; 22].

2.3.5 Смешанные данные

Методы для смешанных данных объединяют текстовые признаки с категориальными и поведенческими сигналами (клики, переходы, время, географический контекст). Их потенциал связан с тем, что поведение пользователя и контекст дополняют чисто текстовую близость и помогают различать омонимию и близкие формулировки. В то же время интеграция модальностей требует аккуратной нормировки вкладов и выбора функции качества. Практический риск состоит в доминировании одного сильного сигнала (например, популярности) [20].

2.4 Сравнение подходов и методов

Сравнительная таблица 4 включает репрезентативные методы, которые часто рассматриваются как «базовые» в обзорах и применяются как ориентиры в прикладных работах. Выбор методов обусловлен следующими критериями.

1. Распространённость и роль базовых ориентиров в обзорах [12; 13; 24].
2. Наличие выраженных преимуществ и ограничений для коротких текстов и разреженных признаков [7; 17].
3. Практическая релевантность в графовой постановке «совстречаемость как сеть связей» [21].

Таблица 4 – Сравнение репрезентативных методов для описанных подходов в контексте поисковых запросов

| Подход | Метод | Преимущества | Недостатки |
|---------------|-------------------------------|---|--|
| Векторный | k -means | Быстрый базовый уровень; эффективен на компактных признаках/эмбедингах [9; 12] | Требует задания числа кластеров; чувствителен к инициализации; ограниченно применим к разреженным коротким текстам [7] |
| Иерархический | Агломеративная (Ward linkage) | Формирует иерархию тематик; допускает выбор уровня детализации [24] | Плохо масштабируется на больших выборках; высокая зависимость от метрики и критерия связи [16; 22] |
| Плотностной | DBSCAN | Не требует задания числа кластеров; выделяет шум и кластеры произвольной формы [17] | Чувствителен к параметрам; ухудшается в высокой размерности; сложен для разреженных текстовых признаков [17] |

Продолжение таблицы 4

| Подход | Метод | Преимущества | Недостатки |
|------------------|-----------------------------------|--|---|
| Графовый | Leiden (детектирование сообществ) | Естественны для совместимости; интерпретируемые; масштабируемые на разреженных графах [21] | Зависимость от построения графа и весов; эффект разрешения; необходимость фильтрации ребер и настройки протокола оценки [6; 22] |
| Смешанные данные | Gower distance + агломеративная | Позволяет интегрировать текстовые и поведенческие сигналы; перспективен для поисковых сценариев [20] | Требует аккуратной нормировки вкладов; риск доминирования одного сигнала; сложность выбора функции качества [20] |

3 Критический анализ

3.1 Противоречия в литературе и методологические ограничения

Систематические обзоры фиксируют, что результаты сравнения методов кластеризации зависят от состава данных, выбранного протокола оценивания и подбора параметров [22]. Это означает, что выводы типа «метод А лучше метода В» часто не переносятся между разными приложениями и не совпадают при использовании разных метрик качества. В задачах кластеризации поисковых запросов это противоречие усиливается, так как вариативность предобработки достаточно велика, от простого приведения к нижнему регистру до морфологического анализа, удаления стоп-слов, стемминга или лемматизации. Изменение предобработки может кардинально изменить результаты кластеризации.

Обзоры ограничений и вызовов подчёркивают риск некорректных обобщений при переносе результатов между доменами и культурами [6; 7]. Например, метод, хорошо работающий на английском корпусе, может работать плохо на русском или китайском из-за различий в морфологии, синтаксисе и семантике.

3.2 Выявленные тенденции

Среди ключевых тенденций в современной литературе можно выделить следующие:

1. Растёт требование предоставлять не только результаты, но и полный код, параметры, данные для их переиспользования и верификации. Это отражает общую тенденцию в науке о данных к воспроизводимости исследований (*reproducibility crisis* в других областях).
2. Новые обзоры фокусируются на алгоритмах, которые работают на миллиардах объектов за разумное время, даже если это требует некоторых компромиссов в качестве [8].
3. Наблюдается движение в сторону гибридизации, при которой используются как явные векторные признаки, так и структурная информация (графы совместимости, социальные графы).

3.3 «Белые пятна» и неисследованные аспекты для поисковых запросов

Несмотря на обширную литературу, для задачи кластеризации поисковых запросов остаются недостаточно исследованными следующие направления:

- *Устойчивое обновление кластеров при потоковом поступлении данных.* Хотя есть работы по инкрементальной кластеризации, практические вопросы остаются открытыми: как переопределять границы кластеров при дрейфе распределения, как избежать взрыва числа кластеров, как сохранять стабильность идентификаторов кластеров.
- *Связь качества кластеров с производственными метриками.* Отсутствуют или малочисленны работы, связывающие качество кластеризации (например, силуэт или модульность) с реальными метриками системы по-

иска, такими как скорость обработки запроса, качество подсказок (CTR), удовлетворённость пользователя. Это затрудняет выбор между алгоритмами из производственной перспективы.

3.4 Актуальные проблемы применения кластеризации в поисковых запросах

На основе анализа литературы и практических соображений выделены три критические проблемы:

1. *Короткие тексты и разреженность сигналов.* Поисковые запросы формируют крайне разреженные признаки. Например, в типичном словаре из 100 тыс. слов каждый запрос содержит 2–4 слова, то есть в разреженном векторе ненулевыми оказываются лишь 2–4 компоненты из 100 тыс. (доли процента). Такая разреженность влечёт за собой несколько последствий [7; 21]:

- статистика совместной встречаемости нестабильна (две редкие пары слов могут случайно встретиться);
- большинство стандартных метрик расстояния (евклидово, косинусное) дают похожие значения для случайных пар объектов, снижая дискриминативность;
- методы становятся очень чувствительны к параметрам, так как малые изменения в построении матрицы приводят к большим изменениям результата.

2. *Шум, морфология и полисемия.* Для русскоязычных и смешанных запросов предварительная обработка (очистка, нормализация, морфологический анализ) является критическим шагом, влияние которого нередко превосходит влияние выбора алгоритма кластеризации. Опечатки требуют нечёткого сопоставления, морфологические варианты слова требуют правильной лемматизации или стемминга, стоп-слова требуют фильтрации, но их удаление может потерять контекстную информацию. Полисемия (многозначность слов) усложняет определение сходства. Например, слово «интернет» в контексте может означать технологию, компанию, услугу или товар [1; 5].

3. Масштабируемость и воспроизводимость при распределённой обработке. При работе с логами поисковых запросов объёмы данных исчисляются миллиардами. Требуется разработка алгоритмов, которые могут работать в ограничениях по памяти и времени, допускают разбиение на шарды и эффективно параллелизуются. Одновременно, воспроизводимость кластеров при повторных запусках и обновлении данных является бизнес-критичным требованием, так как изменение идентификаторов кластеров может нарушить работу множества зависимых систем [6; 22].

4 Описание актуальности исследования и обоснование выбора темы

4.1 Актуальность с точки зрения распределённых систем

В первую очередь, кластеризация актуальна для оптимизации обработки запросов в распределённых системах. Если термины или короткие запросы, которые часто встречаются вместе (совстречаются), размещены на одном сервере или машине (в одном шарде), то запросы, содержащие эти термины, обрабатываются локально без дорогостоящих межсетевых обращений. Такая локализация данных уменьшает сетевой трафик, снижает задержку (*latency*) обработки запроса и повышает пропускную способность системы. На масштабе компаний, обрабатывающей миллиарды запросов в день, такие оптимизации могут экономить миллионы долларов на инфраструктуре.

4.2 Актуальность с точки зрения разработки и инфраструктуры

В прикладных сценариях информационного поиска модели машинного обучения не всегда являются оптимальным выбором по соотношению затрат и эффекта. Во многих задачах кластеризации запросов критичны воспроизводимость, прозрачность правил построения групп и предсказуемость поведения при обновлении данных, а сложные ML-модели могут давать менее стабильные результаты без строгого контроля данных, параметров и протокола сравнения [6; 22]. Кроме того, обучение и сопровождение моделей часто требует разметки, регулярного переобучения и инфраструктуры мониторинга качества, что увеличивает время внедрения и операционные риски. На этом фоне классические алгоритмы кластеризации выступают как более изученный и инженерно надёжный базовый уровень, который проще в реализации и отладке, легче мас-

штабируются в пакетной обработке и допускает чёткую интерпретацию причин объединения объектов [7; 12]. В результате на ранних этапах построения системы и в условиях ограничений по данным и ресурсам кластеризация часто оказывается практическое сложных обучаемых моделей как по скорости внедрения, так и по стабильности результата.

4.3 Актуальность с точки зрения пользователя

Кластеризация запросов и терминов также актуальна для повышения качества пользовательского опыта в поисковых системах. На основе кластеров можно строить множество полезных сервисов:

- подсказки (suggestions) основываются на выборе близких запросов или слов;
- расширение запроса (query expansion) использует кластеры для поиска релевантных синонимов или связанных терминов;
- определение интента запроса (интент-классификация) помогает понять, что пользователь на самом деле ищет.

Качество подсказок измеряется через CTR (click-through rate) и конверсию. Если пользователь видит релевантную подсказку, он кликает на неё, что улучшает пользовательский опыт.

4.4 Актуальность с точки зрения аналитики

Кластеры могут быть использованы для углубленной аналитики поведения пользователей. Такая информация помогает компании принимать решения о развитии нового функционала, выявлении и предотвращении проблем (например, возросшего спроса на определённую услугу).

4.5 Риски и ограничения темы

К основным рискам и ограничениям можно отнести следующие пункты:

- без внешней валидации легко получить кластеры, которые выглядят правдоподобно по внутренним метрикам (силуэт, модульность графа), но не улучшают реальные бизнес-метрики (CTR подсказок, скорость обработки запроса). Это явление известно как «хорошие внутренние метрики,

плохие внешние метрики», и требует осторожности при интерпретации результатов [28];

- кластеризация может усилить нежелательные перекосы в данных. Например, если в логи поисковых запросов попадают определённые темы, которые вызывают асимметрию, то кластеризация может стабилизировать и закрепить такие перекосы;
- без правильной постановки эксперимента трудно отличить улучшение, вызванное самими кластерами, от улучшения, вызванного другими факторами (например, временем года, маркетинговой кампанией);

Эти ограничения не означают, что темой «не стоит заниматься», но требуют строгой методологии оценки, прозрачной интерпретации результатов и проверки на реальных данных [7; 22].

В совокупности изложенные аргументы показывают, что тема кластеризации поисковых запросов и терминов является обоснованной и практико-значимой. Она напрямую связана с ускорением обработки запросов в распределённых системах, улучшением пользовательского опыта, а также аналитикой поведения пользователей. Выбранная постановка допускает реализацию, которая дает прозрачные правила группировки и воспроизводимые результаты, что важно при регулярных обновлениях данных и ограничениях инфраструктуры.

Таким образом, исследование и разработка методов кластеризации в данной предметной области оправданы как с научной точки зрения, так и с прикладной.

5 Перспективные направления исследований

5.1 Тренды последних 2 лет

Современные обзоры подчёркивают, что в 2024–2025 гг. развитие кластеризации всё чаще определяется не столько появлением «ещё одного алгоритма», сколько совершенствованием инженерной и методологической части. Выполняется стандартизация протоколов сравнения, анализ ограничений и требований к воспроизводимости результатов [7; 22]. Параллельно усиливается спрос на решения, способные работать в условиях массовых данных и ограниченных ресурсов (быстрая кластеризация, приближённые вычисления, практико-ориентированные компромиссы качества и затрат) [8]. Для задач

поиска и коротких текстов растёт роль гибридизации представлений, то есть совместного использования графовых связей и векторных признаков, а также подходов для смешанных данных [20].

В прикладных постановках заметен сдвиг к интеграции кластеризации с архитектурными решениями распределённых систем. Результаты группировки используются не только для аналитики, но и как элемент оптимизации хранения и обработки (локализация, шардирование, снижение межсетевых обменов) [21; 23].

5.2 Рекомендации для будущих исследований

С учётом выявленных проблем и описанных тенденций целесообразно сосредоточиться на следующих направлениях.

1. *Инкрементальная (стриминговая) кластеризация графов.* Разработать и эмпирически сравнить методы обновления графа совместимости и частичного пересчёта сообществ без полного пересчёта «с нуля», с акцентом на стабильность кластеров при дрейфе данных и ограничениях по времени [7; 8].
2. *Гибридные графо-векторные модели.* Исследовать способы объединения статистических графовых весов (NPMI/PPMI) с эмбеддингами и/или поведенческими сигналами, чтобы компенсировать разреженность и повысить устойчивость семантики на коротких запросах, сохраняя интерпретируемость результата [20].
3. *Связь кластеризации с распределённой архитектурой (семантическое шардирование).* Проверить, как результаты детектирования сообществ могут использоваться для логического разбиения индекса и уменьшения межсетевых обращений, и сформулировать практические критерии балансировки нагрузки и обработки «пограничных» терминов [21; 23].
4. *Методология оценки и воспроизводимость.* Для каждого варианта протокола фиксировать параметры предобработки, метрики качества и downstream-метрики, чтобы результаты сравнения были переносимыми и проверяемыми; это особенно важно при регулярных перезапусках и обновлениях данных [7; 22].

ВЫВОДЫ

По результатам систематического анализа научных источников за 2010–2025 гг. (PubMed, Scopus, Web of Science, Google Scholar, Wikipedia, arXiv, eLibrary, РИНЦ), а также документации open-source проектов, сформулированы следующие основные выводы.

1. Кластеризация поисковых запросов рассмотрена как инструмент выявления семантической близости и совстречаемости терминов; показано, что результат определяется не только выбранным алгоритмом, но и способом представления данных, предобработкой и протоколом оценивания.
2. Построена классификация подходов по типу представления данных (векторные, графовые, гибридные/смешанные) с фокусом на применимость к коротким, шумным и морфологически вариативным текстам.
3. Выполнено аналитическое сравнение базовых методов (k -means, иерархическая кластеризация, DBSCAN, детектирование сообществ) в контексте кластеризации запросов; сделан вывод об отсутствии универсально лучшего решения и необходимости выбора по компромиссу качества, устойчивости и вычислительных ограничений.
4. Определены актуальные проблемы кластеризации поисковых запросов:
 - разреженность коротких текстов и нестабильность сигналов;
 - высокая чувствительность к предобработке (морфология, шум, опечатки);
 - требования к масштабируемости и воспроизводимости при распределённой обработке больших объёмов данных.
5. Выявлены тенденции на основе научных работ и отмечен сдвиг от «простых сравнений» к систематическому выбору алгоритмов, усиление требований к воспроизводимости и прозрачности, фокус на быстрые масштабируемые решения и гибридизацию представлений.
6. Сформулированы рекомендации, отражающие выявленные «белые пятна» и практические требования:

- инкрементальная (стриминговая) кластеризация графов с сохранением стабильности кластеров при обновлении данных;
- гибридные графо-векторные модели для компенсации разреженности коротких запросов;
- использование результатов детектирования сообществ для семантического шардирования распределённых систем;
- развитие методологии оценки (фиксация предобработки, метрик и downstream-метрик) для переносимости и воспроизводимости результатов.

С учётом рассмотренной постановки и требований к масштабу итоговый выбор алгоритма кластеризации в работе предлагается сделать в пользу **Leiden** как метода детектирования сообществ на взвешенных графах [26]. Решение обосновано следующими основными преимуществами:

1. Представление совстречаемости как графа и поиск сообществ напрямую соответствуют задаче выделения семантических групп.
2. Алгоритм ориентирован на большие разреженные графы и подходит для обработки массивных логов запросов.
3. Параметр разрешения позволяет контролировать гранулярность кластеров без смены семейства алгоритмов.
4. Наличие активно поддерживаемых open-source реализаций (*igraph*, *leidenalg*) упрощает внедрение и повышает воспроизводимость.
5. Результаты сообществ естественно использовать для локализации данных и семантического шардирования в распределённых системах.

Таким образом, выполненный обзор задаёт обоснованный выбор методологии и формирует чёткий фокус дальнейшей работы. В последующей работе стоит сосредоточиться на построении взвешенного графа совстречаемости, выборе протокола оценки, обеспечении воспроизводимости и проверке прикладного эффекта в сценариях распределённой обработки.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 Бова В. В., Кравченко Ю. А., Родзин С. И. Методы и алгоритмы кластеризации текстовых данных (обзор). // Известия ЮФУ. Технические науки. – 2022. – № 4 (228). – С. 122–143.
- 2 Черникова Д. А. Алгоритм кластеризации поисковых запросов. // Евразийский научный журнал. – 2017. – № 12.
- 3 Миронов А. И., Мунерман В. И. Создание частичного индексирования таблицы для оптимизации поисковых запросов. // Современные информационные технологии и ИТ-образование. – 2022. – Т. 18, № 3. – С. 558–565.
- 4 Люнченко С. Применение методов кластеризации для управления запасами товарно-материальных ценностей. // Евразийский союз ученых. – 2020. – № 4-4 (73). – С. 29–37.
- 5 Курейчик В. В., Герасименко П. С. Основные подходы к извлечению текстовой информации (обзор). // Известия ЮФУ. Технические науки. – 2024. – № 4 (240). – С. 6–14.
- 6 Pitafi S., Anwar T., Sharif Z. A taxonomy of machine learning clustering algorithms, challenges, and future realms. // Applied Sciences. – 2023. – Т. 13, № 6. – С. 3529.
- 7 Wani A. A. Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions. // PeerJ Computer Science. – 2024. – Т. 10. – e2286.
- 8 Mahnoor, Shafi I., Chaudhry M., Caro Montero E., Silva Alvarado E., de la Torre Diez E., Abdus Samad M., Ashraf I. A Review of Approaches for Rapid Data Clustering: Challenges, Opportunities, and Future Directions. // IEEE Access. – 2024. – Т. 12. – С. 138086–138120.
- 9 Miraftabzadeh S. M., Colombo C. G., Longo M., Foiadelli F. K-Means and Alternative Clustering Methods in Modern Power Systems. // IEEE Access. – 2023. – Т. 11. – С. 119596–119633.
- 10 Alasali T., Ortakcı Y. Clustering Techniques in Data Mining: A Survey of Methods, Challenges, and Applications. // Computer Science. – 2024. – Т. 9, № 1. – С. 32–50.
- 11 Oyelade J., Isewon I., Oladipupo O., Emebo O., Omogbadegun Z., Aromolaran O., Uwoghiren E., Olaniyan D., Olawole O. Data clustering: Algorithms and its applications. // Proceedings of the 2019 19th International Conference on

Computational Science and Its Applications (ICCSA). – 2019. – C. 71–81.

12 Xu D., Tian Y. A comprehensive survey of clustering algorithms. // Annals of Data Science. – 2015. – T. 2, № 2. – C. 165–193.

13 Aggarwal C. C., Reddy C. K. Data clustering. // Algorithms and applications. Chapman&Hall/CRC Data Mining and Knowledge Discovery Series. – 2014.

14 Ezugwu A. E., Shukla A. K., Agbaje M. B., Oyelade O. N., José-García A., Agushaka J. O. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. // Neural Computing and Applications. – 2021. – T. 33, № 11. – C. 6247–6306.

15 Ezugwu A. E. Nature-inspired metaheuristic techniques for automatic clustering: a survey and performance study. // SN Applied Sciences. – 2020. – T. 2, № 2. – C. 273.

16 Shahid N. Comparison of hierarchical clustering and neural network clustering: an analysis on precision dominance. // Scientific Reports. – 2023. – T. 13, № 1. – C. 5661.

17 Bushra A. A., Yi G. Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms. // IEEE Access. – 2021. – T. 9. – C. 87918–87935.

18 Nagpal A., Jatain A., Gaur D. Review based on data clustering algorithms. // In: 2013 IEEE Conference on Information & Communication Technologies. – 2013. – C. 298–303.

19 Guyeux C., Chrétien S., Bou Tayeh G., Demerjian J., Bahi J. Introducing and comparing recent clustering methods for massive data management in the Internet of Things. // Journal of Sensor and Actuator Networks. – 2019. – T. 8, № 4. – C. 56.

20 Ahmad A., Khan S. S. Survey of state-of-the-art mixed data clustering algorithms. // IEEE Access. – 2019. – T. 7. – C. 31883–31902.

21 Fahad A., Alshatri N., Tari Z., Alamri A., Khalil I., Zomaya A. Y., Foufou S., Bouras A. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. // IEEE Transactions on Emerging Topics in Computing. – 2014. – T. 2, № 3. – C. 267–279.

22 Wegmann M., Zipperling D., Hillenbrand J., Fleischer J. A review of systematic selection of clustering algorithms and their evaluation. // arXiv preprint arXiv:2106.12792. – 2021.

- 23 Nasraoui O., N'Cir C-E Ben. Clustering methods for big data analytics. // Techniques, Toolboxes and Applications. – 2019. – T. 1. – C. 91–113.
- 24 Reddy C. K., Vinzamuri B. A survey of partitional and hierarchical clustering algorithms. // In: Data clustering. – 2018. – C. 87–110.
- 25 Fortunato S. Community detection in graphs. // Physics Reports. – 2010. – T. 486, № 3–5. – P. 75–174.
- 26 Traag V. A., Waltman L., van Eck N. J. From Louvain to Leiden: guaranteeing well-connected communities. // Scientific Reports. – 2019. – T. 9. – Article 5233.
- 27 Bouma G. Normalized (pointwise) mutual information in collocation extraction. // Proceedings of GSCL. – 2009.
- 28 Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. // Journal of Computational and Applied Mathematics. – 1987. – T. 20. – P. 53–65.
- 29 Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. // arXiv preprint arXiv:1301.3781. – 2013.