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 Alasalı 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. – С. 71–81.

12 Xu D., Tian Y. A comprehensive survey of clustering algorithms. // Annals of Data Science. – 2015. – Т. 2, № 2. – С. 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. – Т. 33, № 11. – С. 6247–6306.

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

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

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

18 Nagpal A., Jatain A., Gaur D. Review based on data clustering algorithms. // In: 2013 IEEE Conference on Information & Communication Technologies. – 2013. – С. 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. – Т. 8, № 4. – С. 56.

20 Ahmad A., Khan S. S. Survey of state-of-the-art mixed data clustering algorithms. // IEEE Access. – 2019. – Т. 7. – С. 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. – Т. 2, № 3. – С. 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

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, $PPMI^2$). 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 разреженны, 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 1 представлена хронологическая динамика за последние 15 лет.

Таблица 1 – 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 2.

Таблица 2 – 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 2

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 (дендрограмму), 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 связные 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 разреженностях 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 3 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 3 – Сравнение репрезентативных методов для описанных подходов в контексте поисковых запросов

Подход	**Метод**	**Преимущества**	**Недостатки**
Векторный	k-means	Быстрый базовый уровень; эффективен на компактных признаках/эмбеддингах [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 3

Подход	**Метод**	**Преимущества**	**Недостатки**
Графовый	Leiden (детектирование сообществ)	Естественны для совстречаемости; интерпретируемы; масштабируемы на разреженных графах [21]	Зависимость от построения графа и весов; эффект разрешения; необходимость фильтрации рёбер 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 зависимы 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 совстречаемости, 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 компании, 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-

штабируется 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 практичнее 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 совстречаемости 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 Alasalı 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. – С. 71–81.

12 Xu D., Tian Y. A comprehensive survey of clustering algorithms. // Annals of Data Science. – 2015. – Т. 2, № 2. – С. 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. – Т. 33, № 11. – С. 6247–6306.

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

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

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

18 Nagpal A., Jatain A., Gaur D. Review based on data clustering algorithms. // In: 2013 IEEE Conference on Information & Communication Technologies. – 2013. – С. 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. – Т. 8, № 4. – С. 56.

20 Ahmad A., Khan S. S. Survey of state-of-the-art mixed data clustering algorithms. // IEEE Access. – 2019. – Т. 7. – С. 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. – Т. 2, № 3. – С. 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.