Кузьмин Сергей Николаевич
магистрант
Научный руководитель
Сластихина Мария Дмитриевна
канд. физ.-мат. наук, доцент
ФГБОУ ВО «Саратовский государственный технический университет им. Гагарина Ю.А.»
г. Саратов, Саратовская область

МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ: ТЕНДЕНЦИИ И ВЕКТОРЫ РОСТА

Аннотация: *современный мир динамически меняющаяся система, в который данные – критически важный ресурс. Существуют тенденции роста как их количества, так и источников их производящих. Были проведены исследования и сравнения тенденций роста данных, и людей, занимающихся их обработкой. Чтобы минимизировать эффект, на сотрудников, создаются новейшие методы в сфере интеллектуального анализа данных. В данной статье были собраны актуальные тенденции методов интеллектуального анализа данных и графовых нейросетей. На основе собранных данных, были сделаны выводы и заявление о готовящемся вкладе автора в эту отрасль.*

Ключевые слова: *интеллектуальный анализ данных, анализ данных, нейронные сети, методы обработки данных, методы оптимизации, анализ рынка.*

Введение

В современном мире данные это критически важный ресурс, который используется повсеместно, на ряду с нефтью или золотом. Однако, в отличие от природных ресурсов, их объём не сокращается, а постоянно растёт. Данные представляют собой не только массивы информации, но и статистику о каждом человеке: его привычках, интересах и поведенческих особенностях. В обобщённом виде они формируют точки роста продуктов и сервисов компаний.

В период с 2010 по 2024 год объём мировых данных увеличился в 73,5 раза – с 2 до 147 зеттабайт. По подсчётам индийского аналитического агентства Statista [1], в период с 2024 по 2025 год объём данных возрастёт до 181 зеттабайта, что соответствует приросту почти на одну четверть. По оценкам агентства IDC [2], к 2028 году объём мировых данных достигнет 393,3 зеттабайта.

Рост объёма хранимых данных в сети носит устойчивый характер и за всю историю наблюдений не опускался ниже 16% в год. В период пандемии, когда онлайн-коммуникации и дистанционные формы взаимодействия достигли максимального распространения, темпы прироста данных продемонстрировали выраженный всплеск.

Одновременно увеличиваются не только объёмы данных, но и количество объектов, генерирующих эти данные. К таким объектам относятся практически все устройства интернета вещей (IoT), имеющие доступ к глобальной сети. Так, в период с 2021 по 2025 год количество IoT-устройств увеличилось почти на 8 миллиардов – с 11,28 до 19,08 млрд единиц. Согласно прогнозу агентства DemandSage и эксперта по данным и статистике Навина Кумара (Naveen Kumar), к 2030 году количество устройств интернета вещей возрастёт примерно на треть и достигнет 29,42 млрд единиц [3].

Прогноз роста рынка

Исходя из данных о росте количества IoT-устройств, можно прогнозировать и дальнейший рост объёмов генерируемых данных, представляющих ценность для бизнеса. Наибольший вклад в генерацию данных вносят смартфоны, планшеты, персональные компьютеры и другие устройства, посредством которых пользователи взаимодействуют с веб-ресурсами и друг с другом. Любое действие пользователя – переход по ссылке, заказ товара в интернете или поисковый запрос – является источником данных о его поведении.

Все собранные данные потенциально представляют собой новые точки роста продуктов, продаж и управленческих решений для компаний, владеющих информацией. Для их обработки бизнес привлекает специалистов в области аналитики данных (Data Analyst) и анализа больших данных (Data Scientist). Однако,

по данным Zipdo [4], в период с 2021 по 2031 год количество вакансий в мире увеличится лишь на 28%. При этом в США, согласно данным Zippia [5], численность аналитиков данных вырастет на 108 тысяч рабочих мест за десятилетие – с 93 471 в 2021 году.

Таким образом, рост объёмов данных и источников их генерации опережает прогнозируемый рост числа специалистов отрасли. В этих условиях становится необходимой как оптимизация существующих методов анализа данных, так и разработка новых подходов.

Технологии интеллектуального анализа данных

Многие компании, помимо применения современных аналитических методов, активно используют нейронные сети – модели, специально обученные для интеллектуального анализа данных. Интеллектуальный анализ данных представляет собой процесс автоматического выявления ранее неизвестных, нетривиальных, практически полезных и интерпретируемых закономерностей в больших объёмах данных с целью извлечения информации, пригодной для принятия решений в различных сферах деятельности.

High-Utility Itemset Mining

В период с 2021 по 2025 год получили развитие новые методы, включая High-Utility Itemset Mining (HUIM) и его модификации, усовершенствованные подходы к обучению графовых нейронных сетей (GNN) [6], методы потоковой обработки данных, а также автоматизацию машинного обучения (AutoML) [7].

Метод HUIM (High-Utility Itemset Mining) ориентирован на выявление наборов элементов не только по частоте их появления, но и по показателям полезности, таким как прибыльность или значимость. В период с 2023 по 2025 годы были предложены различные вариации данного метода и способы его применения, эффективные в различных условиях, включая:

- задачу Top-k High Utility Itemsets, направленную на поиск k наиболее полезных наборов [8];

- инкрементный метод High Average-Utility Itemset Mining (iHAUIM), предназначенный для динамически обновляемых баз данных без полной переобработки [9];

- параллельную реализацию HUIM с использованием GPU и распределённых вычислительных систем [10];

- алгоритмы, адаптированные для баз данных с нестабильной или отрицательной полезностью [11].

В области графовых нейронных сетей были разработаны методы, позволяющие снижать эффект доминирования отдельных групп узлов, возникающий вследствие их высокой связности. Это особенно важно при анализе данных социальных сетей. Ранние примеры автономных нейронных сетей, такие как Tay (разработка Microsoft) и Grok (разработка xAI), показали, что обучение на неструктурированных пользовательских графовых данных может приводить к усвоению радикальных паттернов поведения. Это объясняется высокой плотностью связей внутри радикально настроенных групп, которые в графовом представлении становятся приоритетными источниками информации [12].

Современные методы GNN позволяют снижать влияние таких групп при интеллектуальном анализе данных, улучшать качество контентных рекомендаций и уменьшать эффект информационных пузырей [13]. Эти подходы включают предварительную и последующую обработку данных, введение штрафных механизмов при нарушении принципов справедливости, балансировку графовых связей, обучение инвариантных векторов признаков и использование причинно-следственных ограничений.

Среди классических задач интеллектуального анализа данных сохраняют актуальность методы потоковой обработки данных. К ним относятся как экспериментальные методы многопоточной обработки несбалансированных потоков [14], так и потоковые реализации HUIM [15]. Ограничение классических алгоритмов поиска полезных наборов заключается в их ориентации на статические данные, что приводит к генерации устаревших паттернов в динамичных средах.

Метод Scented Utility Miner

Метод Scented Utility Miner (SUM), применяемый в динамических базах данных, использует стратегию повторной индукции (ReInduction) для отслеживания актуальности полезности наборов в реальном времени. Алгоритм SUM включает карту остатков, мастер-карту, механизм повторной индукции и динамическое вычисление порога полезности.

Экспериментальные результаты показывают, что алгоритм SUM превосходит ULB по эффективности как по времени выполнения, так и по потреблению памяти [16], а также снижает количество неактуальных паттернов. По сравнению с алгоритмом EIHI, SUM демонстрирует меньшие требования к памяти за счёт отказа от trie-подобных структур и повторного использования информации предыдущих итераций [17].

Сравнение с алгоритмами, основанными на модели скользящего окна (naive-FHMDS и FHMDS), показывает, что последние могут превосходить SUM по скорости, однако требуют существенно большего объёма памяти и дополнительных вычислений при увеличении числа окон [18].

Одним из ключевых направлений развития отрасли является автоматизация машинного обучения. Для построения нейронных сетей требуются большие объёмы обучающих данных, вычислительные ресурсы и высококвалифицированные специалисты, что делает AutoML логичным развитием существующих подходов [19]. В системах интеллектуального анализа данных нейронные сети планируется обучать в рамках полного ML-конвейера, включающего выбор методов предобработки, генерации признаков, моделей и гиперпараметров.

Уже сегодня применяется перенос знаний (мета-обучение), позволяющий снижать затраты на разработку новых моделей для сходных задач, хотя при этом возможен и негативный перенос, связанный с особенностями исходных данных.

Предлагаемое решение

Однако, существует и альтернативный подход, использование малых языковых моделей, для анализа поступающих данных. Правильно дообученная малая языковая модель способна решить проблему адаптации модели, за счет возможности работать с контекстом.

Работа с контекстом – важная часть адаптивности, так малая языковая модель способна зафиксировать паттерны логов и использовать их для будущего анализа поступающей информации. Эта же особенность позволяет малым языковым моделям эффективно находить аномалии в потоке данных логов.

Помимо преимуществ адаптивности и возможности работать с контекстом, преимуществом любых языковых моделей – это возможность делать вывод на естественном языке.

Таким образом, эффективным решением проблемы обучаемости, эффективности и интерпретируемости могут стать малые языковые модели.

Заключение

Открытыми остаются проблемы адаптации моделей к изменяющимся данным без ручной настройки, а также интеграции экспертных знаний в процесс автоматической оптимизации. Для решения первой задачи используются методы обнаружения дрейфа данных, в том числе в сочетании с алгоритмами потокового анализа, такими как SUM. Для решения второй задачи применяются системы правил различной степени жёсткости, однако уже существует решение, способное разрешить обе задачи одновременно, и существует оно на базе малых языковых моделей.

Список литературы

1. Statista. Volume of data/information created worldwide from 2010 to 2024. – URL: https://www.statista.com/statistics/871513/worldwide-data-created (дата обращения: 20.10.2025).

2. International Data Corporation (IDC). Worldwide Global DataSphere Forecast. – URL: https://my.idc.com/getdoc.jsp?containerId=US53383425 (дата обращения: 20.10.2025).

3. DemandSage. Big Data Statistics and Trends. – URL: https://www.demandsage.com/big-data-statistics (дата обращения: 21.10.2025).

4. Zipdo. Education Report 2025. – URL: https://zipdo.co (дата обращения: 21.10.2025).

5. Zippia. Job Outlook for Data Analysts in the United States. – URL: https://www.zippia.com/data-analyst-jobs/trends (дата обращения: 22.10.2025).

6. Wu Y., Zhang X., Lin J. High-Utility Itemset Mining: A Survey // arXiv preprint arXiv:2204.09888. – 2022 – URL: https://arxiv.org/abs/2204.09888 (дата обращения: 22.10.2025).

7. Lin J., Gan W., Fournier-Viger P. Advances in High-Utility Pattern Mining // Artificial Intelligence Review. – 2024 – URL: https://link.springer.com/article/10.1007/s10462–024–10726–1 (дата обращения: 22.10.2025).

8. Gan W., Lin J., Fournier-Viger P. Top-k High Utility Itemset Mining // arXiv preprint arXiv:2303.14510. – 2023 – URL: https://arxiv.org/abs/2303.14510 (дата обращения: 22.10.2025).

9. Zhang Y. et al. Incremental High Average-Utility Itemset Mining // arXiv preprint arXiv:2407.11425. – 2024 – URL: https://arxiv.org/abs/2407.11425 ;

10. Scientific Reports. – 2024 – URL: https://www.nature.com/articles/s41598–024–60279–0 ; – 2024 – DOI: 10.1038/s41598–024–60279–0 (дата обращения: 22.10.2025).

11. Lin J., Gan W., Fournier-Viger P. Parallel High-Utility Itemset Mining // IEEE Transactions on Knowledge and Data Engineering. – 2023 – DOI: 10.1109/TKDE.2023.3290371. – URL: https://ieeexplore.ieee.org/document/10167780 (дата обращения: 22.10.2025).

12. Ahmed A., Hussein S. High Utility Mining in Dynamic Databases // Iraqi Journal of Intelligent Computing and Informatics. – 2025 – Vol. 4, No. 2. – P. 172–181. – DOI: 10.52940/ijici.v4i2.126 (дата обращения: 22.10.2025).

13. Dong Y. et al. Tutorial on Fairness in Graph Neural Networks // arXiv preprint arXiv:2204.09888. – 2022 – URL: https://arxiv.org/abs/2204.09888 (дата обращения: 22.10.2025).

14. Dong Y. ICDM 2022 Tutorial: Fair Graph Mining. – URL: https://yushundong.github.io/icdm_tutorial_2022.pdf (дата обращения: 22.10.2025).

15. Gomes R., Bifet A. Streaming High-Utility Pattern Mining // arXiv preprint arXiv:2204.03719. – 2022 – URL: https://arxiv.org/abs/2204.03719 ;

16. Machine Learning. – 2023 – DOI: 10.1007/s10994–023–06353–6 (дата обращения: 22.10.2025).

17. Lin J., Gan W., Fournier-Viger P. Scented Utility Miner: Stream High Utility Itemset Mining // International Journal of Data Science and Analytics. – 2023 – URL: https://link.springer.com/article/10.1007/s41019–023–00229–4 ; – 2023 – DOI: 10.1007/s41019–023–00229–4 (дата обращения: 23.10.2025).

18. Lin J., Gan W., Fournier-Viger P. Scented Utility Miner // International Journal of Data Science and Analytics. – 2023 – Иллюстрации к экспериментам (рис. 16–17) (дата обращения: 24.10.2025).

19. Lin J., Gan W., Fournier-Viger P. Comparison of SUM and EIHI Algorithms // International Journal of Data Science and Analytics. – 2023 – Иллюстрации (рис. 21–22) (дата обращения: 24.10.2025).

20. Lin J., Gan W., Fournier-Viger P. Sliding Window-Based HUIM Algorithms // International Journal of Data Science and Analytics. – 2023 – Иллюстрации (рис. 23–24) (дата обращения: 24.10.2025).

21. Lin J., Gan W., Fournier-Viger P. Advances in High-Utility Pattern Mining // Artificial Intelligence Review. – 2024 – URL: https://link.springer.com/article/10.1007/s10462–024–10726–1 (дата обращения: 24.10.2025).