УДК 004.67

ИНСТРУМЕНТ НОРМАЛИЗАЦИИ ДАННЫХ НОРМАТИВНО-СПРАВОЧНОЙ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ

Розанова Анна Вячеславовна, студент, Базовая кафедра «Аналитика больших данных и методы видеоанализа», Уральский федеральный университет имени первого Президента России Б.Н. Ельцина

Предеин Никита Сергеевич, студент, Базовая кафедра «Аналитика больших данных и методы видеоанализа», Уральский федеральный университет имени первого Президента России Б.Н. Ельцина

Ивлиев Трофим Алексеевич, студент, Базовая кафедра «Аналитика больших данных и методы видеоанализа», Уральский федеральный университет имени первого Президента России Б.Н. Ельцина

Толстов Авдей Тарасович, аспирант Института радиоэлектроники и информационных технологий, Уральский федеральный университет имени первого Президента России Б.Н. Ельцина

Саиф Муджахед Абдулла Хаель, старший преподаватель Базовой кафедры «Аналитика больших данных и методы видеоанализа», Уральский федеральный университет имени первого Президента России Б.Н. Ельцина

Аннотация

Целью исследования является реализация инструмента автоматического обнаружения и дальнейшей нормализации дублирующейся нормативно-справочной информации в учётных системах. Для решения задачи применяется двухуровневая архитектура, которая включает алгоритм «HNSW» для поиска k ближайших соседей, выполняющий первичный отбор кандидатов, и открытую большую языковую модель «Qwen3» для семантического анализа и выработки рекомендаций по нормализации. Разработанная концепция позволит значительно увеличить качество данных, хранимых в ERP/CRM-системах, автоматизировав при этом до 80% ручных операций и сократив операционные расходы. Проведённое исследование вносит существенный вклад в область интеллектуальной обработки данных, предлагая точечное решение одной из наиболее острых проблем цифровой трансформации.

КЛЮЧЕВЫЕ СЛОВА: Цифровая трансформация производственных процессов, интеллектуальный подход к автоматизации, нормативно-справочная информация, нормализация информации, семантический анализ, машинное обучение, большие языковые модели, HNSW.

A TOOL FOR NORMALIZING REGULATORY INFORMATION DATA USING A LARGE LANGUAGE MODEL

Rozanova Anna Vyacheslavovna, Student, Department of Big Data Analytics and Video Analysis Methods, Ural Federal University named after the first President of Russia B.N. Yeltsin

Predein Nikita Sergeevich, Student, Department of Big Data Analytics and Video Analysis Methods, Ural Federal University named after the first President of Russia B.N. Yeltsin

Ivliev Trofim Alekseevich, Student, Department of Big Data Analytics and Video Analysis Methods, Ural Federal University named after the first President of Russia B.N. Yeltsin

Tolstov Avdey Tarasovich, Postgraduate Student of the Institute of Radio Electronics and Information Technology – RTF, Ural Federal University named after the first President of Russia B.N. Yeltsin

Saif Mujahed Abdullah Hayel, Senior Lecturer, Department of Big Data Analytics and Video Analysis Methods, Ural Federal University named after the first President of Russia B.N. Yeltsin

Abstract

The purpose of the study is to implement a tool for automatic detection and further normalization of duplicate regulatory and reference information in accounting systems. To solve the problem, a two-level architecture is used, which includes the "HNSW" algorithm for searching for k nearest neighbors, which performs the initial selection of candidates, and the open large language model "Qwen3" for semantic analysis and development of recommendations for normalization. The developed concept will significantly increase the quality of data stored in ERP/CRM systems, while automating up to 80% of manual operations and reducing operating costs. The research conducted makes a significant contribution to the field of intelligent data processing, offering a targeted solution to one of the most pressing problems of digital transformation.

KEYWORDS: digital transformation of production processes, intelligent approach to automation, normative reference information, normalization of information, semantic analysis, machine learning, large language models, HNSW.

Введение

Мастер-данные – это совокупность справочных сведений, описывающих сущности и обеспечивающие контекст для интерпретации операционных данных [1]. К типичным примерам мастер-данных относятся справочники и классификаторы, которые используются для идентификации и категоризации различных объектов в информационных системах.

Управление основными данными – это система управления мастер-данными, для их централизованного контроля, согласования и поддержания целостности [2]. Основная цель заключается в обеспечении единообразного представления в корпоративных системах, а также для устранения дублирования, несогласованности и несопоставимости информации.

Стоит отметить, что мастер-данные делятся на три основных типа:

- Референсные данные – являются статичными или даже слабо изменяемыми справочниками. В качестве примера, можно привести: страны, валюты или общероссийские классификаторы, которые формируются на основе внешних авторитетных источников.
- Основные данные – характеризуются динамичностью и сложной структурой, включая как линейные, так и иерархические зависимости. К ним можно отнести сведения о клиентах, контрагентах, сотрудниках, абонентах и корпоративных активах, где каждая из записей обладает схожим набором атрибутов, но при этом требует постоянного обновления и управления в процессе деятельности.
- Сложные иерархические справочники – представляют структурированные наборы данных, описывающие разнородные сущности, такие как продукты, товары и материалы. Управляются централизованно и характеризуются большей вариативностью атрибутивного состава на разных уровнях иерархии.

Для создания единых «золотых записей» нужно начинать со сбора всей доступной информации об объектах из различных источников её хранения – CRM/ERP-систем, Excel-файлы и другие [3]. Затем специальные алгоритмы должны будут проанализировать эти данные и

выявить дублирующие записи, определяя, что они относятся к одному объекту учёта. На основе собранной информации формируются эталонные записи, которые объединяют лучшие данные из всех источников. Наконец, унифицированные записи синхронизируются со всеми системами, обеспечивая единообразие и достоверность всей информации там, где она используется.

В современных учётных системах нормативно-справочная информация стала играть ключевую роль в поддержке управленческих процессов и обеспечении эффективной работы предприятия. Однако на практике часто может возникнуть ситуация, когда за их содержание отвечают различные подразделения, приводя к потенциальному появлению дублирующих записей с отличающимися идентификаторами и наименованиями продукции [4]. При отсутствии эффективного контроля за качеством справочных данных нормативная база постепенно теряет структурную согласованность. Это приводит к увеличению числа дублирующих записей, снижению прозрачности бизнес-процессов и ошибкам в операционной деятельности. Чаще всего работник создаёт новую запись вместо того, чтобы искать и корректировать существующую, поскольку точно не знает, есть ли подходящий товар на складе, и не уверен, какую позицию из справочника следует использовать. В результате этого предприятие несёт значительные убытки: продукция простаивает на складах, а сотрудники тратят время на лишние операции.

Ключевой проблемой в нормализации НСИ на протяжении многих лет оставалась зависимость традиционных методов от ручного создания правил и доменно-специфичных словарей. В своих работах многие авторы пытались решить проблему, комбинируя правила с классическим машинным обучением, такие как Random Forest или k-means, что повышало точность, но не снимало ограничения с привязкой к конкретным данным и языкам [5-8]. Дальнейшее развитие методов, например, использование вероятностных сигнатур или GPU-ускорения, было направлено на улучшение производительности, но не решало проблему смысловой неоднозначности и вариативности написания [9, 10].

Целью исследования является разработка гибридного инструмента нормализации нормативно-справочной информации с использованием большой языковой модели, который позволит автоматизировать процесс выявления и устранения дубликатов в системах учёта. Такой подход обеспечит более эффективное и экономное решение по сравнению с ручной нормализацией, учитывая как формальное сходство, так и семантическую эквивалентность.

Методы

В данном исследовании для преодоления указанных ограничений предлагается принципиально иной подход, основанный на комбинации векторного поиска (HNSW) и

семантических эмбеддингов, генерируемых крупными языковыми моделями (LLM). Этот метод позволяет отказаться от рутинного создания правил, заменив его автоматизированным анализом контекста и семантического сходства. Подход не только обеспечивает устойчивость к опечаткам и синонимии, но и обладает высокой масштабируемостью, что критически важно для обработки больших объемов неоднородных данных в современных системах управления НСИ.

В основе нашего решения лежит последовательная цепочка преобразования данных: от исходных текстовых записей через векторное их представление к финальным рекомендациям по их нормализации. Каждый этап процесса включает в себя специализированные инструменты, оптимизированные для работы с большими массивами структурированной и неструктурированной информации.

Традиционные rule-based методы и статистические подходы не способны выявлять семантические дубликаты с различными формулировками, но идентичным смыслом, требуя постоянной ручной адаптации под новые области применения. Системы, дополненные правилами, сохраняют зависимость от экспертных знаний и сложно масштабируются на большие объемы данных. Большие языковые модели, используемые в облаке, создают риски утечки конфиденциальной корпоративной информации и сопровождаются значительными эксплуатационными расходами. Комбинированный подход, включающий в себя «HNSW» с локально развёрнутой «LLM», может решить перечисленные проблемы. Векторный поиск обеспечивает быстрое, а главное масштабируемое определение потенциальных дубликатов, а большая языковая модель сможет провести их смысловую верификацию и нормализацию.

Таблица 1. Преимущества и недостатки методов нормализации данных [5-10, 11]

Метод	Преимущества	Недостатки	Эффективность
Rule-base	Простота интерпретации результатов; хорошая точность на структурированных данных; предсказуемость поведения	Трудоёмкость создания и поддержки правил; низкая адаптивность к данным	Низкая
Классические методы машинного обучения	Автоматизация обработки; умеренная масштабируемость; легко интегрируются в системы	Не учитывают семантику; требуют точные размеченные данные	Средняя
Метод, интегрирующий статистический правила к методам машинного обучения	Повышенная точность в специфичных областях; частичная компенсация недостатков классических ML	Сложность настройки; зависимость экспертизы; ограниченная адаптивность	Средняя/Высокая
Нейросетевые методы (R-Drop, CNN и т. д.)	Высокая точность на текстовых данных; возможность выявлять сложные паттерны	Высокие вычислительные затраты; сложная настройка и обучение; требует больших данных	Средняя

Методы векторного поиска (HNSW)	Высокая масштабируемость; эффективный отбор кандидатов	Недостаточная семантическая глубина; зависимость от контекста	Высокая
Метод с применением большой языковой модели (LLM)	Высокая семантическая точность; контекстный анализ; точечная генерация форм	Высокая стоимость; медленная обработка больших объёмов данных; возможные риски	Высокая
Гибридный метод (HNSW + LLM)	Сочетание скорости векторного поиска и семантической глубины; масштабируемость; улучшенная точность и устойчивость к шуму	Сложность интеграции и настройки конвейера; требования к сложной и дорогой инфраструктуре	Очень высокая

Применение больших языковых моделей – ключевой подход для решения описанной задачи по семантическому поиску дублирующихся данных. Интеграция «LLM» в процесс обработки нормативно-справочной информации позволяет не только повысить точность выявления дубликатов, но и минимизировать необходимость ручного вмешательства, обеспечивая масштабируемость решения для работы с большими объемами корпоративных данных [12]. Данный подход подтверждается свой растущей научной релевантностью, что отражается в увеличении количества исследований и цитируемости работ, посвященных применению языковых моделей в управлении мастер-данными. Таким образом, решено было использовать модель «Qwen3», которая выполнит семантический анализ потенциальных дубликатов, оценив вероятность их дублирования и предлагая рекомендации по их нормализации. В статье Малькова и Яшунина он использовался для приближенного поиска ближайших соседей в векторных пространствах [13]. HNSW (Hierarchical Navigable Small World) - это алгоритм, используемый для быстрого приближенного поиска ближайших соседей в многомерных векторных пространствах [13]. Авторы решили проблему масштабируемости традиционных методов поиска, предложив многослойную графовую структуру, где каждый слой содержит подграфы с масштабами их расстояний. Ключевым преимуществом «HNSW» стало логарифмическое время поиска при помощи разделения связей по характеристическим расстояниям и использованию эвристики выбора соседей, демонстрируя производительность на данных разной размерности и структуры.

Результаты

Результаты исследования показали, что предложенная двухуровневая архитектура эффективно справляется со своей задачей автоматического обнаружения и представления рекомендаций по нормализации дублирующейся нормативно-справочной информации. Так алгоритм «HNSW» обеспечил первичный отбор кандидатов для сравнения, сформировав устойчивые и точные результаты на этапе поиска ближайших соседей, тогда как большая

языковая модель «Qwen3» позволила провести семантический анализ и сформировать на его основе рекомендации, учитывая смысловые особенности наименований.

Ниже, на рисунке 1 были представлены распределения значений по семантической близости для отдельных компонентов и их гибридного объединения на тестовой выборке. Наименее эффективным оказался метод R-Drop, который демонстрирует самый широкий разброс значений с основной концентрацией в зоне низкой семантической близости от 0.25 до 0.45, что указывает на нестабильность формируемых векторных представлений. Метод HNSW показывает существенно лучшую и более стабильную семантическую близость, формируя выраженный узкий пик в области значений около 0.4. Наиболее впечатляющие результаты демонстрируют методы с использованием больших языковых моделей (LLM). Чистый LLM-метод создает очень высокий и крайне узкий пик в зоне значений, превышающих 0.8, что свидетельствует о выдающейся способности формировать семантически близкие векторы. Комбинированный подход HNSW+LLM объединяет преимущества обоих методов, показывая высочайшую плотность в области значений, приближающихся к 0.9, что делает его лидером по точности и согласованности в определении семантической близости.

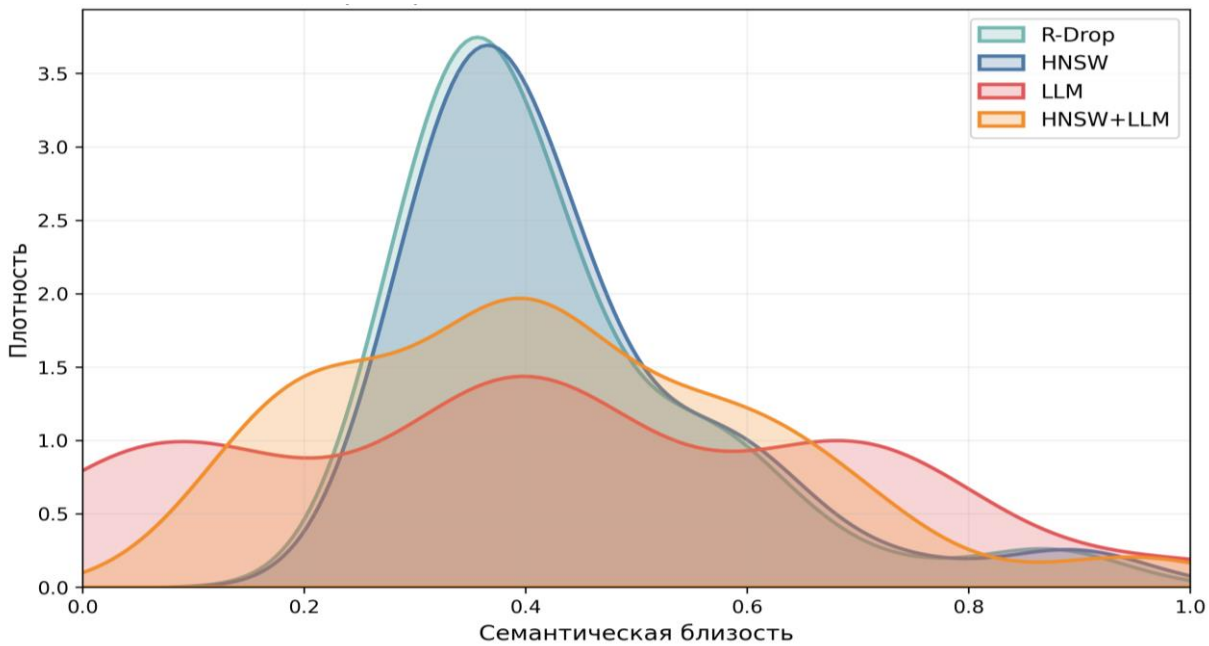

Рис. 1. Плотность распределения значений семантической близости

Анализ показал, что «HNSW» формирует узкий пик, отражающий высокую концентрацию результатов и стабильность поиска, «LLM» создаёт более широкие кривые, учитывающие смысловые нюансы и вариативность формулировок, а «R-Drop» занимает промежуточное положение между ними. Гибридный подход объединяет все сильные стороны этих методов, формируя ровное и информативное распределение, что подтверждает

эффективность для нормализации данных и устойчивость к лексическим шумам в реальных рабочих условиях.

Обсуждение

В ходе исследовательской работы особое внимание было уделено согласованию вычислительных, а также и лингвистических компонентов системы, поскольку именно их взаимодействие полностью определяет итоговое качество нормализации информации.

Основной же сложностью было то, что векторное хранилище и большая языковая модель оперируют разными представлениями: в первом случае – геометрическое, во втором – контекстуальное. Их объединение и обработка потребовала выработки единого механизма интерпретации смысловых признаков, что в значительной степени усложнило настройку, но позволило добиться соответствия семантических структур и их числовых представлений.

Заключение

Проведённое исследование продемонстрировало, что объединение контекстуальных данных, получаемых из большой языковой модели, и геометрических их представлений, хранящихся в индексах типа «HNSW», позволяет построить достаточно эффективный механизм автоматической нормализации и унификации дублирующейся информации. Их сочетание позволяет достичь согласованности без привлечения ручного вмешательства, существенно повышая точность последующей аналитики и управления процессами – при этом сохраняя устойчивость к вариативности формулировок, лексическим шумам и семантическим нюансам, делая подход применимым в реальных и динамичных средах.

Литература

1. Борисов А. М. MDM системы: инновации в управлении мастер-данными и их влияние на бизнес-процессы //Национальная ассоциация ученых. – 2024. – №. 101-1. – С. 8-12.
2. Шарипова В. Д. Роль MDM-системы в деятельности организации и типы данных, использующиеся в MDM-проектах //Системный анализ и логистика. – 2020. – №. 4. – С. 13-20 DOI: 10.31799/2007-5687-2020-4-13-20.
3. Haneem F. et al. Resolving data duplication, inaccuracy and inconsistency issues using Master Data Management //2017 International Conference on Research and Innovation in Information Systems (ICRIIS). – IEEE, 2017. – С. 1-6, doi: 10.1109/ICRIIS.2017.8002453.
4. Li J. et al. EDA-Debugger: An LLM-Based Framework for Automated EDA Runtime Issue Resolution //2025 26th International Symposium on Quality Electronic Design (ISQED). – IEEE, 2025. – С. 1-7, doi: 10.1109/ISQED65160.2025.11014463.

5. Cho, H., Choi, W. & Lee, H. A method for named entity normalization in biomedical articles: application to diseases and plants. BMC Bioinformatics 18, 451 (2017). https://doi.org/10.1186/s12859-017-1857-8
6. Chen L, Fu W, Gu Y, Sun Z, Li H, Li E, Jiang L, Gao Y, Huang Y. Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking. J Am Med Inform Assoc. 2020 Oct 1;27(10):1576-1584. doi: 10.1093/jamia/ocaa155.
7. Вечканова Ю. С., Федосин С. А. Алгоритмы машинного обучения в управлении нормативно-справочной информацией (НСИ). – 2021.
8. Устинов С. М., Устинов М. М., Чистяков В. Ю. Система управления нормативно-справочной информацией как инструмент интеграции информационных систем на уровне данных. – 2021.
9. Zhang Y. et al. Scalable entity resolution using probabilistic signatures on parallel databases //Proceedings of the 27th ACM International Conference on Information and Knowledge Management. – 2018. – С. 2213-2221.
10. Son Y., Kim C., Lee J. FED: Fast and Efficient Dataset Deduplication Framework with GPU Acceleration //arXiv preprint arXiv:2501.01046. – 2025.
11. Izonin, I.; Tkachenko, R.;Shakhovska, N.; Ilchyshyn, B.; Singh,K.K. A Two-Step Data NormalizationApproach for ImprovingClassification Accuracy in theMedical Diagnosis Domain. Mathematics 2022,10, 1942. https://doi.org/10.3390/math10111942
12. Волосников П. Д. Обоснование применения LLM для решения задачи поиска дублей при создании инструмента нормализации данных НСИ / П. Д. Волосников, Э. К. Сагилова. — Текст : электронный // Российские регионы в фокусе перемен : сборник докладов XIX Международной конференции (Екатеринбург, 14–16 ноября 2024 г.). — Екатеринбург : Издательство Издательский Дом «Ажур», 2025. — С. 853-855.
13. Malkov Y. A., Yashunin D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs //IEEE transactions on pattern analysis and machine intelligence. – 2018. – Т. 42. – №. 4. – С. 824-836. DOI: 10.1109/TPAMI.2018.2889473