

УДК 004.89

Стратегическое чанкирование как механизм формализации бизнес-сущностей в корпоративных системах управления знаниями

Шестакова Дарья Николаевна darieshestakova02@gmail.com

МГТУ им. Н.Э. Баумана, Москва, Россия

Аннотация. Рассмотрено стратегическое чанкирование как способ смысловой сегментации корпоративных документов для формализации бизнес-сущностей. Представлен подход, в котором чанки формируются по содержательным границам, обеспечивая логическую целостность и снижение фрагментации знаний. Показано, что такой принцип разбиения формирует более связанное векторное пространство, повышает точность семантического поиска и устойчивость retrieval-ориентированных систем. Установлено, что смысловые чанки точнее отражают структуру процессов и ролей, обеспечивая лучшую интерпретируемость данных. Сделаны выводы, что стратегическое чанкирование можно применять как управленческий инструмент повышения качества корпоративной информационной среды и эффективности интеллектуальных сервисов, работающих с большими массивами документации.

Ключевые слова: стратегическое чанкирование, бизнес-сущности, управление знаниями, семантическое разбиение, эмбединги, векторное пространство, retrieval-ориентированные системы, RAG, корпоративная документация, структурирование информации, семантический поиск

Strategic chunking as a mechanism for formalizing business entities in corporate knowledge management systems

Shestakova Daria Nikolaevna darieshestakova02@gmail.com

BMSTU, Moscow, Russia

Abstract. Strategic chunking is considered as a method of semantic segmentation of corporate documentation aimed at formalizing business entities. An approach is presented in which text fragments are formed according to meaningful boundaries rather than formal size constraints, ensuring contextual integrity and reducing information fragmentation. It is shown that semantic chunking leads to a more coherent vector space, improves the accuracy of semantic search, and increases the robustness of retrieval-oriented systems. It is established that meaning-based chunks more accurately reflect the structure of organizational processes and roles, providing higher interpretability of knowledge representations. The conclusions are drawn that strategic chunking can be applied as a managerial instrument for improving the quality of the corporate information environment and enhancing the effectiveness of intelligent services based on vectorization and text analysis.

Keywords: strategic chunking, business entities, knowledge management, semantic segmentation, embeddings, vector space, retrieval-oriented systems, RAG, corporate documentation, semantic search

Введение. Современные организации сталкиваются с проблемой семантической разрозненности корпоративной документации, что затрудняет доступ к знаниям и снижает эффективность цифровых систем [1]. Традиционные методы чанкирования, основанные на механическом разбиении текста, не отражают бизнес-логику и ведут к потере контекста, дублированию информации и снижению точности семантического поиска, несмотря на развитие эмбедингов и векторных моделей. В этих условиях требуется стратегический подход, при котором чанкирование рассматривается как управленческая политика формализации бизнес-сущностей. Такой подход позволяет создавать семантический слой над корпоративной документацией, усиливая ее связность и пригодность для интеллектуальных систем. Цель исследования — обосновать модель стратегического чанкирования, позволяющую повысить качество управления знаниями и эффективность корпоративных KMS.

Методы и материалы; результаты. Исследование опиралось на анализ корпоративной документации, структурированной по функциональным направлениям и представленной в виде регламентов, инструкций и проектных материалов. Семантическое пространство формировалось с использованием современных моделей эмбедингов, что позволяло сравнивать две стратегии разбиения текста: техническую и смысловую. При выборе смысловых границ учитывались свойства *content-defined chunking*, обеспечивающие устойчивость структуры данных и снижение потерь информации [2]. Этот подход согласуется с требованиями *retrieval*-ориентированных систем, где качество поиска определяется правильностью сегментации и полнотой контекста [3], а также подтверждается результатами анализа русскоязычных *retrieval*-моделей, чувствительных к неоднородности сегментации [4]. Смысловое разбиение соответствует принципам управления знаниями, предполагающим структурированность, однозначность и доступность информации как условие эффективности корпоративных KMS [5]. Эффективность подходов оценивалась по метрикам семантической близости и плотности векторного пространства, а также экспертным проверкам соответствия фрагментов бизнес-сущностям.

Смысловое разбиение обеспечило более четкую и устойчивую структуру корпоративных знаний по сравнению с техническим подходом. Векторные представления таких чанков формировали плотные смысловые группы, лучше соответствующие бизнес-сущностям, что снижало фрагментацию и потери информации, полностью согласуясь с преимуществами контент-ориентированного разбиения [2]. Улучшенная согласованность фрагментов положительно отражалась на работе *retrieval*-компонентов: повышалась точность поиска, устойчивость контекста и качество ответов, что соответствует требованиям RAG-систем к корректности сегментации [3]. На русскоязычных документах стратегическое разбиение обеспечивало более стабильные показатели релевантности и позволило минимизировать ошибки, возникающие из-за лингвистической вариативности [4]. Полученная структура также согласовывалась с принципами корпоративного управления знаниями, предпо-

лагающими логичность, ясность и доступность содержания как основу эффективной работы с документами [5].

Обсуждение полученных результатов. Полученные результаты демонстрируют, что стратегическое чанкирование формирует более устойчивую и интерпретируемую структуру корпоративных знаний, чем разбиение текста фиксированными фрагментами. Смысловые границы позволяют сохранять целостность бизнес-сущностей, благодаря чему векторное пространство становится более связным, а тематические области — лучше очерченными. Это приводит к снижению потерь информации, уменьшению фрагментации и повышению качества последующего анализа.

Улучшение работы retrieval-компонентов объясняется тем, что смысловые чанки содержат достаточный объем контекста для точного извлечения и сопоставления информации. В результате система реже сталкивается с неполными или разорванными фрагментами, что повышает точность поиска и устойчивость ответов. Особенно заметным оказалось влияние смыслового разбиения там, где исходные документы отличаются стилевой неоднородностью и сложной терминологией: стратегически сформированные фрагменты обеспечивали более стабильные показатели качества и снижали количество нерелевантных результатов.

Стратегическое чанкирование также согласуется с принципами эффективного управления знаниями, предполагающими ясность, структурированность и доступность корпоративной информации. Формирование логически завершенных смысловых блоков облегчает навигацию, способствует согласованности между документами и повышает качество использования знаний внутри организации.

Стратегическое чанкирование можно рассматривать как ключевой инструмент повышения качества корпоративной информационной среды. Оно влияет не только на технические характеристики векторных моделей, но и на общий уровень управляемости знаниями, делая информационные ресурсы организации более целостными, точными и пригодными для дальнейшего анализа.

Заключение. Проведенное исследование показало, что стратегическое чанкирование обеспечивает более логичную и согласованную структуру корпоративных знаний по сравнению с традиционным разбиением текста. Формирование фрагментов по смысловым границам позволяет точнее отражать бизнес-сущности, снижать уровень фрагментации и обеспечивать полноту контекста, что напрямую повышает качество семантического поиска и работу retrieval-ориентированных систем.

Смысловое разбиение способствует созданию устойчивого векторного пространства, где тематические области представлены более четко, а связи между ними фиксируются корректнее. Это не только улучшает результаты автоматизированных моделей, но и повышает удобство работы пользователей с документацией, делая знания более доступными и структурированными.

Стратегическое чанкирование следует рассматривать не как технический процесс, а как управленческую практику, влияющую на качество корпора-

тивной информационной среды. Полученные результаты подтверждают, что использование смысловых границ является эффективным инструментом повышения управляемости знаний и может служить основой для дальнейшего развития корпоративных систем поиска, анализа и интеграции информации.

Список источников

- [1] Зимова Н.С. Особенности внедрения системы управления знаниями в российских компаниях. *Научный результат. Социальные и гуманитарные исследования*, 2019, т. 5, № 3. <https://doi.org/10.18413/2408-9338-2019-5-3-0-7>
- [2] Gregoriadis M., Balduf L., Scheuermann B., Pouwelse J. A thorough investigation of content-defined chunking algorithms for data deduplication. URL: <https://arxiv.org/abs/2409.06066> (accessed 24.09.2025).
- [3] Чураков И.И., Егоров А.Н., Апаева П.Д. Использование метода RAG для наполнения базы данных цифрового помощника на базе LLM. URL: https://elibrary.ru/download/elibrary_83017794_44819466.pdf (дата обращения 24.09.2025).
- [4] Мельников А.В., Николаев И.Е., Русанов М.А., Аббазов В.Р. Сравнительный анализ методов RAG для построения русскоязычных интеллектуальных сервисов. *Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника»*, 2025, т. 25, № 2, с. 5–18. <https://doi.org/10.14529/ctcr250201>
- [5] Алорайни Л.И.М. Сущность, цель и задачи управления знаниями в международной компании: основные тенденции в современных условиях. *Форум молодых ученых*, 2024, № 6(94), с. 88–93.