

УДК 004.89

Автоматизированное извлечение и классификация терминов в динамическом графе знаний

Жиров Дмитрий Викторович¹

DVZhirov@sberbank.ru

Лебедев Аркадий Ильич^{2(*)}

lebedevai2@student.bmstu.ru

Костиков Константин Олегович²

kostickov.costya@yandex.ru

Исмаилов Адиль Назимович²

hustle.adik69@yandex.ru

¹ ПАО «Сбербанк», Москва, Россия² МГТУ им. Н.Э. Баумана, Москва, Россия

Аннотация. Рассмотрен подход к автоматизированному извлечению и классификации терминов с использованием динамического графа знаний. Представлен конвейер, включающий распознавание терминов в тексте, определение их принадлежности к иерархической структуре классов и автоматическое расширение графа при появлении новых понятий. Показано, что предложенный метод обеспечивает универсальность обработки текстов различных предметных областей, позволяет адаптивно формировать таксономию и поддерживать актуальность графа знаний без ручного редактирования. Сделан вывод о применимости разработанного решения для систем интеллектуального анализа данных и поддержки принятия решений.

Ключевые слова: граф знаний, извлечение терминов, классификация, автоматизация, динамическая таксономия

Введение. В современных условиях объем текстовой информации растет экспоненциально, что требует разработки инструментов для ее систематизации и автоматической структуризации. Одним из эффективных решений выступает построение графов знаний, в которых термины связываются отношениями принадлежности и иерархии [1]. Однако классические подходы страдают статичностью: изменение предметной области влечет за собой ручную переработку структуры графа [2]. Для преодоления этого ограничения предлагается метод динамического формирования графа знаний, основанный на автоматизированном извлечении и классификации терминов. Цель работы заключается в создании непрерывного конвейера обработки текстов, способного не только извлекать термины, но и интегрировать их в существующую систему классов, а при необходимости — создавать новые элементы графа. Это обеспечивает универсальность подхода и возможность его применения в различных областях знания.

Методы и материалы; результаты. Предложенный конвейер включает несколько этапов. На первом этапе осуществляется распознавание терминов в тексте с использованием методов машинного обучения [3]. Текст разбивается на слова, убираются символы, не относящиеся к алфавитным или знакам препинания. Дообученная модель распознавания естественного языка BERT

совершает контекстное выделение потенциальных терминов из текста и размечает в формате BIO. Полученные термины проходят дополнительную фильтрацию при помощи LLM-эксперта.

Далее выполняется проверка принадлежности термина к существующему дереву классов. Для этого используется иерархическая классификация, позволяющая последовательно уточнять позицию термина на разных уровнях таксономии. Если уверенность классификации оказывается недостаточной, инициируется процесс создания нового класса. Его структура формируется автоматически, что обеспечивает адаптивное расширение графа знаний.

Экспериментальная проверка показала, что предложенный метод обеспечивает стабильное качество извлечения ($F1\text{-score}=0.849$, $\text{Precision}=0.835$, $\text{Recall}=0.86$) и классификации терминов, а также корректное расширение таксономии.

Заключение. Разработан метод автоматизированного извлечения и классификации терминов в динамическом графе знаний. В отличие от традиционных решений, предложенный подход обеспечивает возможность адаптивного расширения структуры классов без участия эксперта. Система универсальна и может быть использована для анализа текстов различных предметных областей.

Список источников

- [1] Dehal R.S., Sharma M., Rajabi E. Knowledge Graphs and Their Reciprocal Relationship with Large Language Models. *Machine Learning and Knowledge Extraction*, 2025, vol. 7, no. 2, art. no. 38. <https://doi.org/10.3390/make7020038>
- [2] Al Khatib H.S. et al. Patient-centric knowledge graphs: a survey of current methods, challenges, and applications. *Frontiers in Artificial Intelligence*, 2024, vol. 7, art. no. 1388479. <https://doi.org/10.3389/frai.2024.1388479>
- [3] Пимешков В.К., Никонорова М.Л., Шишаев М.Г. Комбинированный метод извлечения терминов для задачи мониторинга тематических обсуждений в социальных медиа. *Информатика и автоматизация*, 2024, т. 23, № 4, с. 1110–1138.