УДК 004.89

Автоматическая нормализация гетерогенных данных IoT-датчиков на основе языковых моделей с использованием внешних источников

Подморин Д.О. (ИТМО)
Научный руководитель – аспирант, инженер Ковальчук М.А. (ИТМО)

Введение.

Современный Интернет вещей — это быстро развивающаяся сетевая структура, направленная на интеграцию физических устройств в цифровое пространства, используя интернет как среду для передачи данных [1]. Стремительное развитие Интернета вещей (IoT) и других систем, требующих внедрения большого количества устройств, привело к возникновению проблем в области совместимости данных из-за разнообразия форматов и протоколов связи между различными устройствами и системами.

Согласно аналитике от промышленных партнеров, ручная интеграция одного датчика занимает от 2 до 4 часов. Это время уходит на поиск и чтение документации, написание драйверов и проверку работы интегрированного устройства.

Анализируя полученные данные с реальной IoT системы, были обнаружены 1200 уникальных параметров, каждое устройство имеет примерно 3-4 целевых параметра. Таким образом, в систему было интегрировано примерно 350 датчиков. Исходя из этой информации, мы можем посчитать, что компания потратила 1050 человеко-часов на интеграцию или 131 рабочий день.

Таким образом, ручная интеграция новых устройств требует больших трудозатрат на одно устройство, а ввиду большого числа устройств требует автоматизации.

Основная часть. В рамках решения проблемы унификации параметров с датчиков были разработаны 4 алгоритма:

1) Алгоритм, основанный на онтологиях и экспертных знаниях. Данный подход позволяет автоматизировать интеграцию для известных данных путем создания четких правил преобразования. Основным недостатком подхода является необходимость экспертных знаний для создания новых правил.
2) Алгоритм, основанный на больших языковых моделях (БЯМ). Формат входных данных может быть не определен, что требует создания гибкого алгоритма преобразования данных. Языковые модели являются гибким инструментом и даже могут работать с неизвестными форматами, что позволяет применять их для интеграции IoT-устройств в общую систему [2].
3) Гибридный алгоритм, объединяющих онтологии и БЯМ. Основная идея заключается в оптимизации запросов к языковой модели. Использование онтологий позволяет снизить количество токенов, что позволяет избегать переполнения контекста БЯМ.
4) Алгоритм, основанный на гибридном подходе с использованием поиска документации в сети Интернет для повышения качества унификации. Данный подход требует грамотного промптирования поскольку документация также будет тратить контекстное модели [3].

Выводы. Были разработаны 4 алгоритма и проведены эксперименты. Наиболее эффективный алгоритм показал прирост производительности примерно в 18 раз над ручной интеграцией при высоком качестве.

Список использованных источников:

1. Elijah, O., Rahman, T. A., Orikumhi, I., Leow, C. Y., Hindia, M. N. (2018). An overview of Internet of Things (IoT) and data analytics in agriculture: Benefts and challenges. IEEE Internet of Things Journal, 5, 3758–3773
2. Mior, Michael J. "Large Language Models for JSON Schema Discovery." arXiv preprint arXiv:2407.03286 (2024).
3. Clavié, Benjamin Ciceu, Alexandru Naylor, Frederick Soulié, Guillaume Brightwell, Thomas.

(2023). Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification.