УДК 004.043

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ОБРАБОТКИ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ И ОЦЕНКА ЕЁ ПРИМЕНИМОСТИ НА ИНДУСТРИАЛЬНЫХ КЕЙСАХ

Лазебный В. В.[1], Федоров Д. А.[1]
Научный руководитель – канд. техн. наук, доцент Федоров Д. А.[1]
[1]Университет ИТМО
vllazebnyi@gmail.com

Введение

Автоматизация сбора и структурирования информации из разнородных веб-источников (новостные порталы, маркетплейсы, технические логи) стала базовой потребностью для бизнес-аналитики, мониторинга и принятия решений. На практике классический парсинг на XPath/CSS-селекторах оказывается «хрупким»: достаточно небольшого изменения верстки и конвейер ломается, а динамически подгружаемый контент часто вообще не попадает в обработку. Подходы информационного извлечения (IE) на основе машинного обучения могут давать качественный результат, но обычно требуют разметки и настройки под каждый домен, что ведет к дороговизне и сложному масштабированию. Большие языковые модели (LLM) выглядят как естественный способ обработки страницы для вывода структуры, однако их прямое применение в промышленном режиме упирается в риск генерации недостоверных фактов («галлюцинаций») и стоимость вычислений [2, 3]. Поэтому ключевой задачей становится сокращение разрыва между гибкостью LLM и требованиями к надежности, воспроизводимости и автоматизации всего процесса – именно этому посвящено исследование и разработка.

Основная часть

В работе предложен комбинированный метод агрегации неструктурированных данных, который объединяет сильные стороны классического сбора и интеллектуального извлечения в едином многоэтапном конвейере. Метод включает четыре ключевых этапа:

1) Гибридный сбор данных, где Goose3 применяется для извлечения основного текста и метаданных со статических страниц, а Selenium – для источников с динамическим контентом (рендеринг, подгрузка блоков, скрипты). Такой подход снижает зависимость от конкретной разметки и повышает устойчивость к «косметическим» изменениям страниц.

2) Семантическая дедупликация на основе сравнения эмбеддингов, позволяющая отсеивать повторы и близкие по смыслу материалы до запуска дорогостоящих этапов обработки, тем самым экономя вычислительные ресурсы и уменьшая задержки.

3) Извлечение и нормализация с помощью LLM: модель получает текст и строго заданный промпт с целевой JSON-схемой, правилами заполнения и примерами, после чего преобразует содержимое в структурированную запись с нормализованными полями.

4) Двухэтапная верификация результатов – ключевое научное предложение работы. На первом этапе для выявления «галлюцинаций» и нестабильных фрагментов генерируются три варианта структурирования с разными значениями температуры, затем сравниваются эмбеддинги и согласованность ключевых полей (например,

сущностей, числовых атрибутов, дат/цен) для обнаружения аномалий. На втором этапе применяются формальные правила контроля качества: проверка обязательности полей, типов данных, допустимых диапазонов, форматов (даты, валюты, единицы измерения) и соответствия шаблонам. В сумме это снижает долю ручной проверки и делает результат ближе к требованиям промышленной эксплуатации, где важна не только полнота, но и проверяемость полученных данных [3].

Выводы

Разработана автоматизированная система, обеспечивающая устойчивую к изменениям источников агрегацию неструктурированных веб-данных в режиме, близком к реальному времени. Эксперименты на выборке новостных статей показали сохранение 87% семантического содержания исходных материалов после обработки, а также высокую корректность заполнения целевых полей при наличии многоуровневой валидации. Практическая значимость подхода заключается в возможности внедрения в системы мониторинга новостей и медиа, конкурентного анализа, сбора рыночных данных (в т.ч. товарных атрибутов) и анализа логов, где требуется быстро подключать новые источники и форматы без постоянных затрат на поддержку «ручных» парсеров. Дополнительным преимуществом является модульность конвейера: каждый этап может масштабироваться независимо и заменяться альтернативными реализациями без пересборки всей системы

Литература

1. Boegershausen J., Datta H., Borah A., Stephen A. Fields of Gold: Scraping Web Data for Marketing Insights // Journal of Marketing. – 2022. – Vol. 86.
2. Juraj Vladika, Ihsan Soydemir, Florian Matthes. Correcting Hallucinations in News Summaries: Exploration of Self-Correcting LLM Methods with External Knowledge. – arXiv:2506.19607, 2025.
3. Remadi, A., Hage, K., Hobeika, Y., & Bugiotti, F. To prompt or not to prompt: Navigating the use of Large Language Models for integrating and modeling heterogeneous data // Data Knowl. Eng. – 2024. – Vol. 152.