

АВТОМАТИЗИРОВАННОЕ ФОРМИРОВАНИЕ ЖУРНАЛОВ СОБЫТИЙ НА ОСНОВЕ НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДАННЫХ ДЛЯ ЗАДАЧ PROCESS MINING

Рассматривается подход к автоматизированному формированию журналов событий из неструктурированных данных для задач Process Mining. Разработан программный прототип, реализующий извлечение и структурирование событий с использованием генеративных моделей и формирование event log.

Введение

Для задач Process Mining необходимы журналы событий, содержащие информацию о действиях, временных характеристиках и участниках процесса [1–2]. На практике такие сведения часто представлены в неструктурированном виде, что требует их предварительного извлечения и формализации [3]. Для решения этой задачи могут использоваться генеративные модели, учитывающие контекст текста и обеспечивающие извлечение и структурирование событий [4].

I. Подход к формированию журналов событий

В работе предложен подход к автоматизированному формированию журналов событий на основе неструктурированных текстовых данных для задач Process Mining. Предлагаемое решение ориентировано на преобразование текстовой информации в формализованное представление, пригодное для последующего анализа процессов.

Подход реализует конвейер обработки, включающий получение текстовой информации, извлечение событий, нормализацию результатов и формирование event log в формате XES. Для извлечения и структурирования событий используются генеративные модели, учитывающие контекст текста. Разработан программный прототип, реализующий основные этапы данного конвейера.

II. Результаты извлечения событий

В демонстрационном сценарии использовались новости о кибератаке на компании. Результатом обработки является структурированное представление событий в формате JSON, на основе которого формируется журнал событий. Примеры извлечённых событий приведены в таблице 1.

Полученные результаты показывают возможность преобразования неструктурированных данных в структуру для анализа процессов.

Таблица 1 – Примеры извлечённых событий

Дата	Событие	Результат
2023-10-01	Обнаружена подозрительная активность	Начато расследование
2023-10-10	Критические серверы зашифрованы	Системы заблокированы
2023-10-25	Выполнено восстановление систем	Частичная потеря данных

III. Выводы

Разработанный прототип демонстрирует возможность автоматизированного преобразования неструктурированных текстовых данных в журналы событий для задач Process Mining. К ограничениям подхода относится зависимость качества извлечения от структуры исходного текста и используемой генеративной модели. Дальнейшее развитие работы может быть связано с количественной оценкой качества извлечения и расширением набора обрабатываемых источников.

Список литературы

1. Marin-Castro, H. M. Event Log Preprocessing for Process Mining: A Review / H. M. Marin-Castro, E. Tello-Leal // Applied Sciences. – 2021. – Vol. 11, No. 22. – Art. 10556. DOI: 10.3390/app112210556.
2. Pegoraro, M. Mining Uncertain Event Data in Process Mining / M. Pegoraro, W. M. P. van der Aalst // Proceedings of the International Conference on Process Mining (ICPM). – Aachen, 2019. – P. 89–96. DOI: 10.1109/ICPM.2019.00023.
3. Adnan, K. An Analytical Study of Information Extraction from Unstructured and Multidimensional Big Data / K. Adnan, R. Akbar // Journal of Big Data. – 2019. – Vol. 6, Issue 1. – P. 1–12. DOI: 10.1186/s40537-019-0254-8.
4. Patel, R. A Scoping Review of Large Language Model-Based Approaches for Information Extraction from Radiology Reports / R. Patel, L. Green, J. White // npj Digital Medicine. – 2024. DOI: 10.1038/s41746-024-01219-0.

Воронова Кристина Дмитриевна, магистрант кафедры информационных технологий в бизнесе НИУ-ВШЭ-Пермь, kdvoronova@edu.hse.ru.

Научный руководитель: Лядова Людмила Николаевна, доцент кафедры информационных технологий в бизнесе НИУ ВШЭ, кандидат физико-математических наук, доцент, llyadova@hse.ru.