ISSN 2225-9309 (Print)
2024. № 1 (30). С. 16–20

Молодежный
Вестник УГАТУ

http://journal.ugatu.su

УДК 004.89

Система обнаружения аномалий на основе анализа журнала событий

Е. А. Атарская[1], **А. М. Вульфин**[2]

[1] atarskaya.ea@ugatu.su, [2] vulfin.alexey@gmail.com

[1,2] ФГБОУ ВО «Уфимский университет науки и технологий» (УУНиТ)

Аннотация. Значительные объемы накапливаемых данных мониторинга состояния компонентов информационной системы в виде текстовых журналов работы либо не анализируются вовсе, либо подвергаются анализу с помощью ограниченного набора правил. Интеллектуальный анализ журнала работы системы (ЖРС) позволяет выявлять основные типы происходящих событий, отслеживать динамику смены характерных событий во времени, а также выделять нетипичные (аномальные) события для заданного временного окна анализа. Проанализированы основные работы в данном направлении, отмечающие ключевые технологии и возможности построения моделей обнаружения аномалий на основе анализа слабоструктурированных данных ЖРС.

Ключевые слова: обнаружение аномалий; глубокое обучение; анализ логов; парсинг логов; метаданные событий.

ВВЕДЕНИЕ

Значительные объемы накапливаемых данных мониторинга состояния компонентов информационной системы (ИС) в виде текстовых журналов работы (логов) либо не анализируются вовсе, либо подвергаются анализу с помощью ограниченного набора правил.

Более глубокий анализ ограничен возможностями систем разбора (парсеров) и последующего сопряжения с *SIEM* системой с помощью настраиваемых индивидуально программных коннекторов. Дальнейший анализ ориентирован, в первую очередь, на выявление однозначно интерпретируемых записей, вызванных событиями информационной безопасности, помеченными как потенциально опасные для информационной системы. Существенное количество записей журнала работы системы (ЖРС) не имеет разметки принадлежности к потенциально опасным маркерам, характеризующим действия возможного злоумышленника. Анализ ЖРС позволяет между тем выявлять основные типы происходящих событий, группировать их по степени сходства, отслеживать динамику смены характерных для системы типа событий во времени, а также выделять нетипичные (аномальные) события для заданного временного окна анализа. Выделенные в ходе анализа нетипичные события могут быть вызваны как внутрисистемными процессами, характеризующими штатный режим работы, так и сбоями и нарушениями нормального функционирования компонентов информационной системы ввиду ошибок конфигурации, обновления программного обеспечения или внешними причинами. Количество подобных событий относительно общего объема, находящих отражение в ЖРС событий невелико, однако их обнаружение требует значительных усилий по анализу накапливаемых данных мониторинга. Следовательно, повышение эффективности анализа за счет автоматизации разбора, формализации и интеллектуального анализа данных является актуальной задачей, направленной на повышение оперативности выявления аномальных состояний компонент ИС.

Цель работы: совершенствование моделей и алгоритмов анализа текстовых журналов работы системы для повышения оперативности выявления аномальных состояний системы.

Задачи:

1. Анализ методов, моделей и алгоритмов обработки слабоструктурированных данных текстовых ЖРС.

ОБЗОР СУЩЕСТВУЮЩИХ ПОДХОДОВ К АНАЛИЗУ ЖУРНАЛОВ СОБЫТИЙ НА ОСНОВЕ ТЕХНОЛОГИЙ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

ЖРС широко применяются в больших и сложных программно-интенсивных системах; обнаружение аномалий на основе журналов используется для диагностики и устранения неисправностей подобных систем. Существующие методы извлекают последовательности записей событий из журналов в виде временных векторов, сохраняя информацию о времени между событиями, но не позволяют отслеживать причинно-следственные связи между записями ЖРС.

Методы обнаружения вредоносных событий в ЖРС в основном сосредоточены на поведении пользователей и анализируют журналы, фиксирующие их действия в ИС. Большинство таких методов учитывают последовательную связь между записями в журнале, моделируя последовательное поведение пользователей. Однако они игнорируют другие взаимосвязи, что неизбежно приводит к неэффективному обнаружению различных сценариях атак. Предложенный в статье [1] метод *log2vec* является модульным методом, основанным на встраивании графов. Вначале *log2vec* использует эвристический подход, который преобразует записи журнала в гетерогенный граф с учетом разнообразных связей между ними. Затем используется улучшенное вложение графа, соответствующее гетерогенному графу, которое может автоматически представить каждую запись журнала в вектор низкой размерности. Третьим компонентом *log2vec* является практический алгоритм обнаружения, способный разделять вредоносные и обычные записи журнала на различные кластеры. Оценка прототипа *log2vec* показала, что метод превосходит современные подходы, такие как глубокое обучение и скрытая марковская модель, а также *log2vec* демонстрирует способность обнаруживать вредоносные события в различных сценариях атак.

В работе [2] предложена структура под названием *CFDet* для высокоточного обнаружения аномалий в последовательных данных на основе идеи интерпретируемого машинного обучения. *CFDet* способен выявлять аномальные записи в последовательностях без использования каких-либо помеченных аномальных последовательностей/записей для обучения. Структура *CFDet* состоит из модели обнаружения аномальных последовательностей на основе нормальных последовательностей и модели обнаружения аномальных записей с самоконтролем. Основная идея заключается в том, что аномальная подпоследовательность в аномальной последовательности должна находиться далеко от центра кластера нормальных выборок, в то время как остальная нормальная подпоследовательность, рассматриваемая как противоречащая фактической последовательность исходной аномальной последовательности, должна быть близка к центру. Для обнаружения аномальных последовательностей используется подход глубокого описания данных вектора поддержки (*Deep Support Vector Data Description*). Эксперименты на трех наборах данных показывают, что модель может идентифицировать аномальные записи с высокой точностью.

Использование в методах обнаружения аномалий на основе журнала необработанных и неструктурированных записей журнала, т.е. без учета контекста каждого события и метаданных события в журнале, в последствии ограничивает возможности модели глубокого обучения на ранней стадии, что приводит к пропуску аномальных событий и ложным срабатываниям. В статье [3] предложена система обнаружения *DeepSyslog*, основанная на механизме глубокого обучения для обнаружения аномалий и использующая журнал *Syslog*, который регистрирует аномальные события, указывающие на небезопасное состояние компьютерной системы. Ос-

новываясь на последовательном характере потока логов, используется встраивание предложений без наблюдения для извлечения семантической и контекстной информации, скрытой в потоке записей, а не встраивание слов или точечное встраивание, которые фиксируют только сходство между словами журнала. Встраивание предложений интегрируется с метаданными события для формирования полного представления *Syslog*, которое может отличить аномалию, вызванную коррелированными записями журнала и исключительными метаданными события в журнале. Результаты моделирования на широко используемых наборах данных журнала (*BGL* и *HDFS*) показывают, что *DeepSyslog* достигает высокой производительности по сравнению с существующими подходами к обнаружению аномальных событий на основе журнала.

Исследования языковых свойств инструкций по ведению журнала показали, что они содержат обширную информацию, связанную с аномалиями. В статье [4] предложен метод надежного и практичного обнаружения аномалий по ЖРС. Он преодолевает общий недостаток соответствующих работ, т.е. необходимость в большом количестве вручную маркированных обучающих данных, путем построения модели обнаружения аномалий на основе журнальных инструкций. *ADLILog* объединяет информацию, связанную с аномалиями, и данные целевой системы для обучения модели глубокой нейронной сети с помощью последовательной двухфазной процедуры обучения. Обширные экспериментальные результаты на двух наиболее часто используемых эталонных наборах данных показывают, что *ADLILog* превосходит смежные методы: контролируемый на 5-24%, а неконтролируемый на 40-63% по F_1-мере. Эксперименты показывают, что *ADLILog* обладает полезными практическими свойствами, касающимися небольшого размера самой модели и экономии времени при её обновлении.

Отсутствие взаимосвязи между записями журнала приводит к потере логических ассоциаций внутри журнала логов. В статье [5] показан метод обнаружения аномалий в журналах *LogLR*, основанный на механизме логических рассуждений. *LogLR* извлекает логическую связь между временными векторами журналов и повышает точность обнаружения, объединяя логическую тензорную сеть (*LTN*) с *LSTM*. *LogLR* использует *LTN* для выявления логической связи между последовательностями журналов и получает слабые метки для обучения модели *LSTM* с помощью метода оценки слабых меток, что экономит временные ресурсы. *LogLR* продемонстрировал эффективность на двух широко используемых публичных наборах данных.

В ходе эмпирического исследования в статье [6] было обнаружено, что существующие подходы к обнаружению аномалий на основе журналов значительно страдают от ошибок разбора журналов, которые возникают из-за слов, не входящих в словарный запас (*OOV – out-of-vocabulary*), и семантических несоответствий ввиду специфичности конкретной области. Ошибки разбора журнала могут привести к потере важной информации для обнаружения аномалий. Чтобы устранить ограничения существующих методов, был предложен подход к обнаружению аномалий на основе журналов, который не требует разбора журнала. *NeuralLog* извлекает семантический смысл необработанных сообщений журнала с помощью *BERT*-кодера и представляет их в виде семантических векторов. Эти векторы представления затем используются для обнаружения аномалий с помощью модели классификации на основе трансформеров, которая может улавливать контекстную информацию из последовательности журналов. Экспериментальные результаты показали, что предложенный подход может эффективно понимать семантический смысл сообщений журнала и достигать точных результатов обнаружения аномалий. *NeuralLog* достигает значения F_1-меры более 0,95 на четырех публичных наборах данных.

В статье [7] предлагается семантически-ориентированная система представления для анализа журналов *Log2Vec*. *Log2Vec* сочетает в себе метод встраивания слов, специфичных для журнала, для точного извлечения семантической информации из журналов, и процессор слов *OOV* для встраивания слов *OOV* в векторы во время выполнения. Эксперименты по оценке на четырех публичных наборах данных производственных журналов показывают, что *Log2Vec* не только устраняет проблему, связанную с *OOV*-словами, но и значительно повышает производительность двух популярных задач управления сервисами на основе журналов, включая

классификацию журналов и обнаружение аномалий. *Log2Vec* способен назначить «мягкое» представление каждому журналу во время выполнения, чтобы избежать ложных тревог.

При анализе процессов задача сводится к тому, чтобы превратить необработанные данные о событиях в значимые модели, представления или действия. Одной из ключевых проблем анализа процессов на основе данных является высокая размерность данных. В работе [8] решение такой проблемы происходит путем разработки методов обучения представлений для бизнес-процессов. Парадигма обучения представлений применяется к действиям, отпечаткам, журналам и моделям с целью изучения высокоинформативных, но низкоразмерных векторов, часто называемых вкраплениями, на основе архитектуры нейронной сети. Впоследствии эти векторы могут быть использованы для задач автоматического вывода, таких как кластеризация отпечатков, сравнение процессов, их предиктивный мониторинг, обнаружение аномалий и т. д. Основным вкладом работы является предложение архитектур обучения представлений на уровне действий, отпечатков, журналов и моделей, которые могут производить распределенное представление этих объектов и тщательный анализ потенциальных приложений.

ЗАКЛЮЧЕНИЕ

ЖРС являются одним из наиболее ценных источников данных для крупномасштабного управления услугами. При анализе ЖРС задача сводится к тому, чтобы превратить необработанные данные о событиях в значимые модели, представления или действия. Одной из ключевых проблем анализа процессов на основе данных является высокая размерность данных.

Представление журналов, которое преобразует неструктурированные тексты в структурированные векторы или матрицы, служит первым шагом к автоматизированному анализу журналов. Однако существующие методы представления журналов не представляют семантическую информацию журналов, специфическую для конкретной области, и не обрабатывают слова, не входящие в словарный запас новых типов журналов во время выполнения.

Перспективным направлением решения подобной задачи является применение технологий нейросетевой обработки ЖРС с помощью моделей-трансформеров. Проанализированы основные работы в данном направлении, отмечающие ключевые технологии и возможности построения моделей обнаружения аномалий на основе анализа слабоструктурированных данных ЖРС.

СПИСОК ЛИТЕРАТУРЫ

1. **Liu F., Wen Y., Zhang D., Jiang X., Xing X., & Meng D.** Log2vec: A Heterogeneous Graph Embedding Based Approach for Detecting Cyber Threats within Enterprise [Электронный ресурс]. URL: https://www.semanticscholar.org/paper/Log2vec%3A-A-Heterogeneous-Graph-Embedding-Based-for-Liu-Wen/09d634892435e4527eb9da9405458a4db7c6bccf (дата обращения 14.03.2023).

2. **Cheng H., Depeng X., Shuhan Y., Xintao W.** Fine-grained Anomaly Detection in Sequential Data via Counterfactual Explanations [Электронный ресурс]. URL: https://www.researchgate.net/publication/364556797_Fine-grained_Anomaly_Detection_in_Sequential_Data_via_Counterfactual_Explanations (дата обращения 14.03.2023).

3. **Junwei Zh., Yijia Q., Qingtian Z., Peng L., Jianwen X.** DeepSyslog: Deep Anomaly Detection on Syslog Using Sentence Embedding and Metadata [Электронный ресурс]. URL: https://www.researchgate.net/publication/362921618_DeepSyslog_Deep_Anomaly_Detection_on_Syslog_Using_Sentence_Embedding_and_Metadata (дата обращения 14.03.2023).

4. **Bogatinovski J., Madjarov G., Nedelkoski S., Cardoso J., Kao O.** Leveraging Log Instructions in Log-based Anomaly Detection [Электронный ресурс]. URL: https://www.researchgate.net/publication/361831259_Leveraging_Log_Instructions_in_Log-based_Anomaly_Detection (дата обращения 14.03.2023).

5. **Zhang K., Di X., Liu X., Li B., Fang L., Qin Y., Cao J.** LogLR: A Log Anomaly Detection Method Based on Logical Reasoning [Электронный ресурс]. URL: https://www.researchgate.net/publication/365470160_LogLR_A_Log_Anomaly_Detection_Method_Based_on_Logical_Reasoning (дата обращения 14.03.2023).

6. **Le V., Zhang H.** Log-based Anomaly Detection Without Log Parsing [Электронный ресурс]. URL: https://www.researchgate.net/publication/353700482_Log-based_Anomaly_Detection_Without_Log_Parsing (дата обращения 14.03.2023).

7. **Meng W., Liu Y., Huang Y., Zhang Sh., Zaiter F., Chen B., Pei D.** A Semantic-aware Representation Framework for Online Log Analysis [Электронный ресурс]. URL: https://www.researchgate.net/publication/347018960_A_Semantic-aware_Representation_Framework_for_Online_Log_Analysis (дата обращения 14.03.2023).

8. **De Koninck P., vanden Broucke S., Weerdt J.** act2vec, trace2vec, log2vec, and model2vec: Representation Learning for Business Processes [Электронный ресурс]. URL: https://link.springer.com/chapter/10.1007/978-3-319-98648-7_18 (дата обращения 14.03.2023).

ОБ АВТОРАХ

АТАРСКАЯ Елена Андреевна, студент 2 курса магистратуры.

ВУЛЬФИН Алексей Михайлович, доцент кафедры вычислительной техники и защиты информации.

METADATA

Title: Anomaly detection system based on event log analysis.

Authors: E. A. Atarskaya[1], A. M. Vulfin[2]

Affiliation:

[1,2] Ufa University of Science and Technology (UUST), Russia.

Email: [1] atarskaya.ea@ugatu.su, [2] vulfin.alexey@gmail.com

Language: Russian.

Source: Molodezhnyj Vestnik UGATU (scientific journal of Ufa University of Science and Technology), no. 1 (30), pp. 16-20, 2024. ISSN 2225-9309 (Print).

Abstract: The significant volumes of accumulated information system component condition monitoring data in the form of text-based system logs are often analyzed only with a limited set of rules. Data mining of data logs allows identifying the main types of events occurring, monitoring changes of characteristic events over time, and identifying anomalous events for a given time window of analysis. We analysed the main works in this area, indicating key technologies and opportunities to design anomaly detection models based on the analysis of low-structured log data.

Key words: anomaly detection, deep learning, log analysis, log parsing, event metadata.

About authors:

ATARSKAYA, Elena Andreevna, postgraduate student 2 year, Ufa University of Science and Technology.

VULFIN, Alexey Mikhailovich, Associate Professor, Dept. of Computer Engineering and Information Security.