2021

Данные логов и предварительная обработка

КИРСАНОВА СОФЬЯ

Оглавление

Данные логов

Лог – это файл с хронологической информацией, который хранится на компьютере или сервере пользователя и содержит данные о действиях программ или пользователей.

- Логи могут понадобиться, если нужно узнать статистику по сайту. Логи сайтов отображают следующую информацию:
    - статистику посещаемости
    - моменты входа и выхода с сайта
    - поисковые запросы, по которым приходят пользователи, и наиболее популярные страницы сайта
    - поисковики, страны и браузеры посетителей
    - сайты, которые ссылаются на этот ресурс
- В случае вирусов или атаки логи помогут быстрее выяснить причину и соответственно помочь устранить ее
- Для восстановления доступов используются логи авторизации, которые собирают данные о попытках входа
- В случае ошибок в работе определенного ПО, устройства или ОС, когда необходимо определить источник проблемы

Логи используются практически везде, где ведется запись и прослеживание истории программного процесса. Лог файл показывает события и его непосредственный источник. Причиной события может быть:

- действия пользователя системы
- программная ошибка
- действия со стороны ОС
- действие со стороны используемого оборудования

Предобработка данных логов

Метод предварительной обработки – функция, которая получает журнал событий и возвращает предварительно обработанный журнал событий.

Как правило, существует две категории последовательности аномалий: последовательные и количественные аномалии. Программы обычно выполняются в соответствии с фиксированными потоками, а журналы представляют собой последовательность событий, создаваемых этими исполнениями.

Последовательная аномалия возникает, если последовательность журналов отклоняется от нормальных шаблонов программных потоков. Выполнение программы имеет некоторые постоянные линейные зависимости, которые могут быть зафиксированы количественными соотношениями журналов и всегда должны выполняться при различных рабочих нагрузках.

Возникает количественная аномалия, если эти отношения нарушаются для набора журналов.

Существующие подходы к автоматическому обнаружению аномальных лог-последовательностей, которые можно разделить на две категории: подходы,

основанные на счетчике сообщений журнала, интеллектуальный анализ для выявления количественных аномалий и подходы, основанные на глубоком обучении, для изучения последовательных шаблонов из последовательностей журналов.

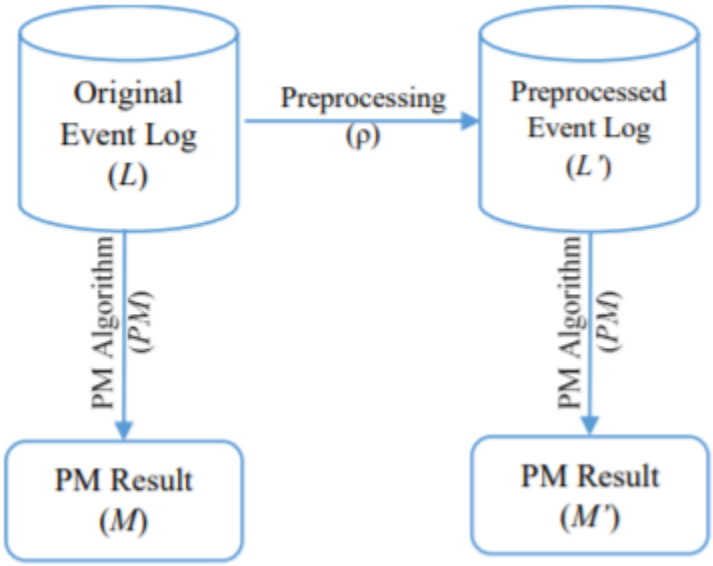
Уменьшение размера данных

При работе с большими журналами событий невозможно использовать стандартные алгоритмы. Простой подход к решению этой проблемы заключается в уменьшении размера данных о событиях с помощью выборки. Кроме того, во многих случаях не требуется полный журнал событий, и аномалию в процессе уже можно обнаружить, используя лишь небольшую часть данных.

В статье [2] (описание датасета приведено в статье) и [3] (ссылка на датасет в разделе источников) рассматриваются различные методы выбора подмножеств и оценивается их эффективность на основе данных о реальных событиях.

Значит анализ можно применять к предварительно обработанному журналу событий без необходимости его изменения. Чтобы оценить эффективность методов предварительной обработки, мы можем рассмотреть эффективность методов майнинга или качество результатов.

Для повышения производительности вычислений мы учитываем время PM (на иллюстрации) на L и сравниваем его со временем вычисления на L' + время предварительной обработки (однако для некоторых приложений мы можем игнорировать время предварительной обработки).



В целом, с помощью методов предварительной обработки мы уменьшаем сложность данных о событиях.

На рисунке 2 – различные подходы к предварительной обработке, которые можно использовать для уменьшения сложности и размера журналов событий. Для упрощения рассмотрим журнал событий как табличные данные, строки которых соответствуют логам, а столбцы отображают действия.

Источники

1. Adriano Augusto, Raffaele Conforti, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, Andrea Marrella, Massimo Mecella, Allar Soo: Automated Discovery of Process Models from Event Logs: Review and Benchmark // IEEE Transactions on Knowledge and Data Engineering (Volume: 31, Issue: 4, April 1 2019)
2. Mohammadreza Fani Sani, Sebastiaan J. van Zelst, Wil M. P. van der Aalst: The Impact of Event Log Subset Selection on the Performance of Process Discovery Algorithms // ADBIS 2019: New Trends in Databases and Information Systems
3. Mohammadreza Fani Sani: Preprocessing Event Data in Process Mining // Process and Data Science Chair, RWTH Aachen University, Aachen, Germany, 2020