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.

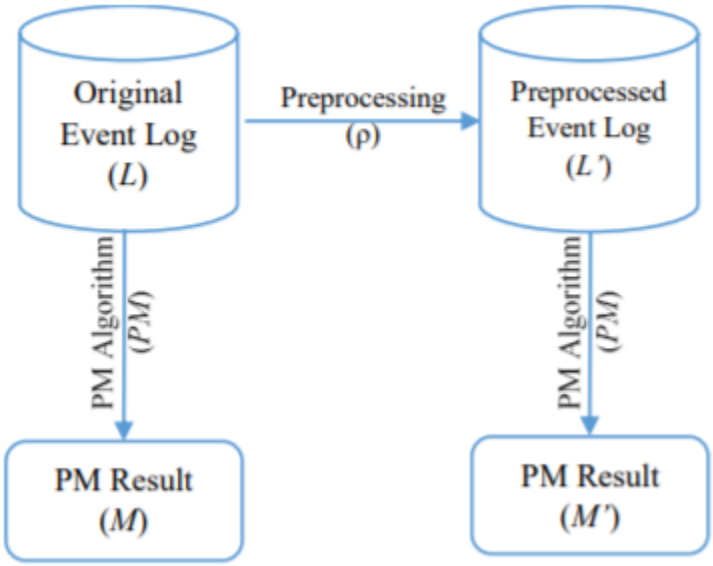
Уменьшение размера данных

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.

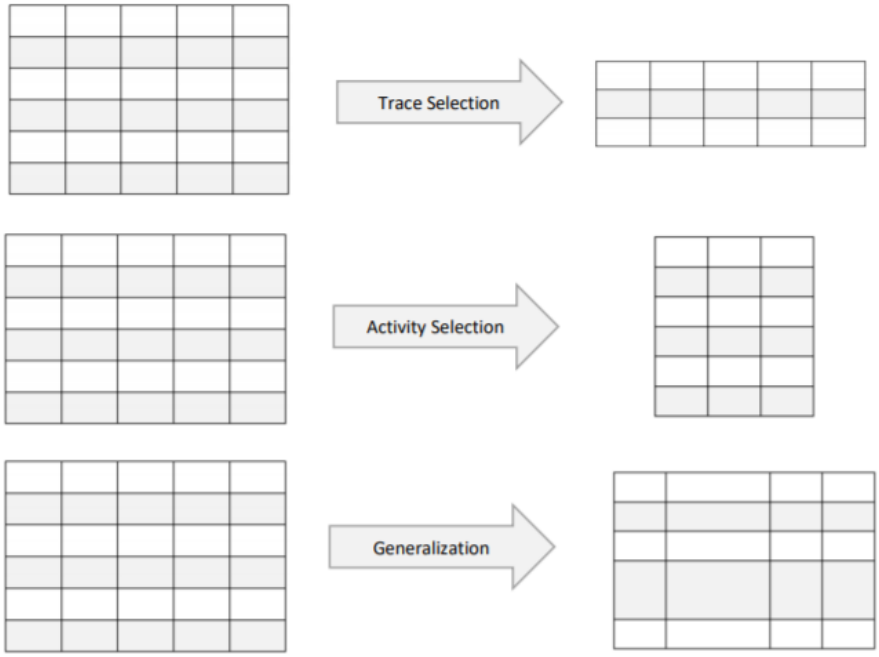
Значит анализ можно применять к предварительно обработанному журналу событий без необходимости его изменения. Чтобы оценить эффективность методов предварительной обработки, мы можем рассмотреть эффективность методов майнинга или качество результатов.

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.



Для каждого подхода возможны различные методы предварительной обработки.

- Первый подход, который охватывает большинство методов предварительной обработки, направлен на то, чтобы выбрать некоторые строки и поместить их в том виде, в каком они есть, в журнал предварительно обработанных событий. Другими словами, они выбирают некоторые логи исходного журнала событий и помещают их в предварительно обработанные журналы событий.
- Для повышения производительности результатов интеллектуального анализа процессов (а иногда и их качества одновременно) также можно выбрать некоторые действия с логами и провести выборку по ним, из которой потом собрать журнал уже обработанных логов
- Наконец, можно объединить аномалии или действия с более общими признаками. Так мы уменьшаем сложность логов, сокращая количество аномалий в них. Мы можем применить этот подход только на уровне уникальных вариантов выборки, только на уровне активности и на обоих из них.

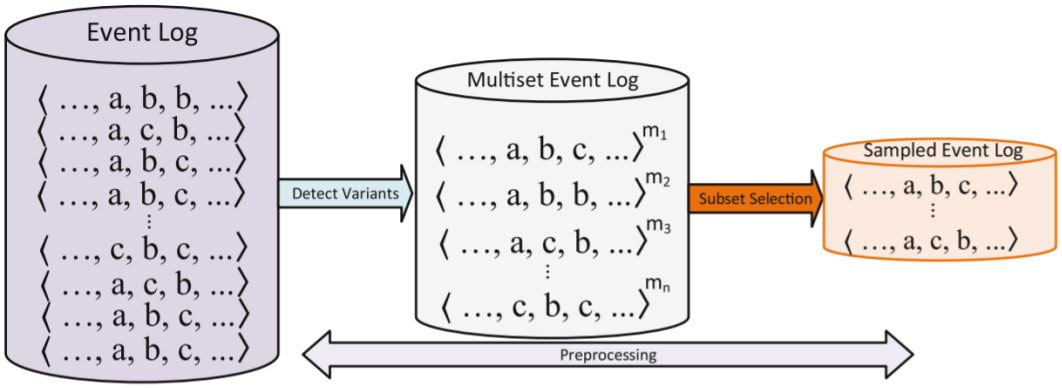
Методы фильтрации эффективно снижают размер данных о событиях, используемых алгоритмами обнаружения. Однако иногда необходимое время для применения этих алгоритмов фильтрации превышает время обнаружения аномалии. Кроме того, эти методы не имеют точного контроля над размером выборочного журнала событий.

Для выборки могут использоваться различные поведенческие элементы журнала событий, например события, непосредственно следующие за какими-либо

особенными отношениями, но не все из них полезны для целей обнаружения аномальных процессов.

Выбрав события, можно рассмотреть события из разных частей процесса, чтобы максимально полно охватить места, в которых могут встретиться ошибки.

Общая схема выбора подмножеств:

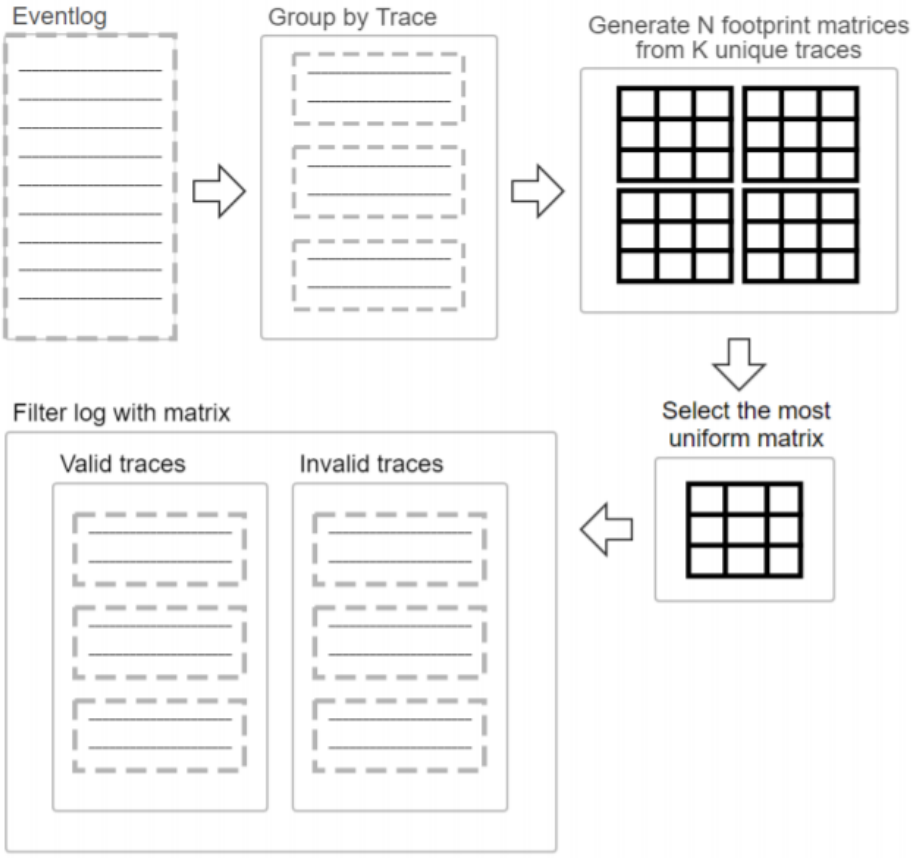


Таким образом, мы создаем подмножество журналов событий только на основе самых важных логов, которые непосредственно следуют или связаны с какими-то особенными местами в программе. Следовательно, эти методы выбора подмножества принимают журнал событий в качестве входных данных и возвращают выборочный журнал событий.

Поход с помощью матриц (статья [4])

Еще один метод основан на сгенерированных матрицах отпечатков (footprint matrix) случайно взятых логов. Находя наиболее похожие матрицы для проверки всего журнала, можно исключить или выделить логи, представляющие необычное поведение. Инструмент был реализован с помощью Python 3, NumPy и Pandas.

Этот предлагаемый инструмент и его алгоритм представляют собой подход к задаче автоматической очистки, которая может находить наиболее частые аномалии в журнале событий. Вся методика основана на концепции матриц отпечатков, которые были введены с помощью альфа-алгоритма. Это алгоритм, который извлекает информацию для каждого журнала и создает так называемые матрицы отпечатков для дальнейшего анализа. Они представляют поток событий для одного или нескольких случаев. Эти матрицы можно использовать для поиска закономерностей и сходств в журнале логов. Метод может очищать журналы событий от неожиданного поведения (т.е. редких последовательностей событий) и позволяет инвертировать результат и выделять редкие случаи.



Все обращения в журнале группируются в зависимости от их идентификатора обращения, чтобы получить всю трассировку (группу). Для каждой из N сгенерированных матриц случайным образом выбирается K групп (т. е. каждая группа используется только для генерации ровно одной матрицы следа). После вычисления сходства каждой матрицы со всеми остальными для проверки журнала событий используется наиболее однородная из них. Окончательный журнал содержит только те трассировки, которые могут быть воспроизведены выбранной матрицей.

Пример:

Процесс управления штрафами за дорожное движение: набор данных представляет собой журнал событий в реальной жизни из системы управления штрафами за дорожное движение местной полиции Италии. Этот набор данных имеет свои особенности с большим количеством неполных трассировок и содержит 150,370 трассировок, включающих 561,470 событий. Количество событий на трассировку варьируется от 2 до 20. Журнал без какой-либо предварительной обработки содержит 231 различных вида вариантов процесса.

THE RESULT AFTER APPLYING A VARIANT FILTER, MANUAL FILTERING AND THE PROCESS PRUNER

	Variants	Traces	Events
Original	231	150370	561470
Variant Filter 56%	131	150270	560551
Process Pruner	29	146147	534594
Manual Filter	8	121833	470119

Фильтр, который отобрал только варианты, повторяющиеся больше одного раза, получил 131 вариант следов. После ручной фильтрации осталось 8, а после применения подхода с матрицами – 29.

Методы обработки с помощью естественного языка (статья [5])

Основная идея заключается в том, что большинство системных журналов являются неструктурированными текстами, “распечатанными” определенными процедурами систем, так что методы обработки естественного языка могут быть применены или улучшены для обнаружения аномалий в журнале.

Что необходимо для этой реализации:

1. Эффективный метод представления шаблонов для точного извлечения семантической и синтаксической информации из шаблонов журналов. Он фиксирует не только контекст слова, но и семантическую информацию, включая синонимы и антонимы (журналы с антонимами обычно указывают на разные события).

2. Механизм для объединения новых шаблонов в случае логов, которые похожи синтаксически и семантически.

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
   Ссылка на датасет из статьи: https://data.4tu.nl/repository/collection:event logs
4. David Baumgartner, Andreas Haghofer, Martin Limberger: Process Pruner: A Tool for Sequence-Based Event Log Preprocessing // Department of Data Science and Engineering University of Applied Sciences Upper Austria, 2020
   Ссылка на датасет из статьи: De Leoni, M. (Massimiliano) and Mannhardt, F. (Felix), “Road traffic fine management process,” 2015. [Online]. Available: https://data.4tu.nl/repository/uuid:270fd440-1057-4fb9-89a9-b699b47990f5
5. Weibin Meng, Ying Liu, Yichen Zhu, Shenglin Zhang , Dan Pei, Yuqing Liu, Yihao Chen, Ruizhi Zhang , Shimin Tao, Pei Sun and Rong Zhou: LogAnomaly: Unsupervised Detection of Sequential and Quantitative Anomalies in Unstructured Logs // *IJCAI* (Vol. 19, No. 7, pp. 4739-4745), 2019
   Датасеты логов: BGL dataset [Oliner and Stearley, 2007], HDFS dataset [Xu et al., 2009], их описания приводятся в статье.

Datasets	Duration	# of logs	# of anomalies
HDFS	38.7 hours	11,175,629	16,838(blocks)
BGL	7 months	4,747,963	348,460(logs)