

Формирование атрибутов интеллектуального анализа процессов в разнородных системах

Я. А. Бекенева

Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)
yabekeneva@etu.ru

Аннотация. Интеллектуальный анализ процессов является перспективным методом исследования различных последовательностей событий, которые могут считаться процессом. При мониторинге большого количества параметров возникает сложность выбора параметров для их назначения атрибутам формата журнала событий. От корректности сопоставления атрибутов зависит точность и информативность получаемой модели и, как следствие, качество анализа процесса и выявления аномалий. В статье рассмотрены особенности формирования журналов событий при обработке данных от разнородных источников.

Ключевые слова: интеллектуальный анализ процессов; журналы событий; хес; атрибуты; обработка данных

I. ВВЕДЕНИЕ

Мониторинг процессов с помощью систем, содержащих большое количество разнородных источников, обычно характеризуется большими наборами данных, содержащими множество регистрируемых параметров. При проведении анализа часто приходится оценивать степень влияния разных параметров на итоговый результат, определять наиболее важные параметры, удаляя из рассмотрения другие параметры [1].

Проведение интеллектуального анализа процессов на основе таких наборов данных осложнено жесткой структурой формата журналов событий .xes. Журнал событий содержит всего четыре атрибута: *Case ID*, *Timestamp*, *Resource* и *Activity* [2]. При этом существует также логическая структура формирования журнала событий, т. е. не всегда можно выбрать четыре наиболее важных атрибута из исследуемого набора данных. Кроме того, атрибут *Timestamp* выделен под значение временной метки, без которой построение модели процесса невозможно. Для получения более информативных моделей в распределенных системах мониторинга были предложены метода централизованного [3] и распределенного [4] формирования журналов событий. Ключевой идеей в обоих случаях является комбинация нескольких наиболее значимых атрибутов в качестве атрибута *Activity*, так как именно этот атрибут определяет действие и отображается в блоках модели процесса.

В рамках проведенных исследований не рассматривались вопросы влияния разных атрибутов в составе *Activity* на информативность модели, а также отсутствовало подробное рассмотрение результатов при

выборе разных параметров для атрибутов *Resource* и *Case ID*.

В статье рассматриваются различные варианты формирования атрибутов формата журнала процесса и их влияние на информативность модели в различных случаях.

II. ОПИСАНИЕ ПОДХОДА

A. Особенности формирования атрибута *Activity*

Атрибут *Activity* предназначен для описания действий, выполняемых в рамках наблюдаемого процесса: медицинские процедуры, этапы сделки и т.д. Обычно речь идет о выборе параметра, который бы отвечал за этап выполнения процесса. В случае анализа перемещения объекта в пространстве в качестве атрибута *Activity* может быть выбрано название локации, тогда модель процесса будет представлять собой маршрут, по которому перемещается объект. Такая модель может позволить выявить отклонения в маршруте или недопустимые задержки, однако в ней не будет содержаться больше никакой информации. Локация в качестве атрибута *Activity* может быть выбрана при исследовании задач распространения какого-либо явления. Например, в рамках конкурса VAST 2019 [5] рассматривалась проблема анализа распространения землетрясения и его последствий. Использование локации в качестве *Activity* при анализе исходного набора данных позволила бы проследить проявления землетрясения в виде последовательности разных районов.

В сложных системах анализ маршрута является недостаточным для анализа процессов и выявления аномалий в них. Так, в работах [3, 4] предлагается комбинация атрибутов, указывающих на локацию и тип совершенного действия, что позволяет описать произведенное действие в каждой точке мониторинга. Также было предложено использовать наиболее важные атрибуты, связанные с измерениями какого-либо параметра. В рассматриваемом примере с перемещениями транспортных средств в качестве одного из ключевых атрибутов было выбрано значение веса транспортного средства, полученное в результате проведения процедуры взвешивания в определенных точках мониторинга.

Учет значения веса при анализе может быть полезным при выявлении аномалий, связанных либо с несовпадением значений веса при последовательных измерениях, когда по регламенту эти значения должны совпадать, а также в противоположных случаях. В таком

случае могут быть выявлены аномалии, которые могут указывать на различные недопустимые действия, такие как, например, хищение какой-либо части транспортируемых материалов.

Следует рассмотреть и другие параметры, которые могут иметь значение при анализе процессов.

При анализе процессов, где совместно действуют два субъекта, может возникать необходимость сопоставления этих субъектов. Например, если транспортным средством управляет определенный водитель, то может к числу ключевых параметров может также относиться и идентификатор водителя. Другим способом учета может являться назначение идентификатору водителя атрибуту *Resource*.

При анализе перемещений сотрудников, описанном в [5], разными устройствами системы HVAC регистрируется множество параметров. Выделенные в работе параметры, имеющие наибольшее влияние на качество выявления аномалий, также могут быть рассмотрены для добавления в состав атрибута *Activity*.

В. Формирование атрибутов Case ID и Resource

При формировании атрибутов *Case ID* и *Resource* следует помнить следующую логику.

В качестве атрибута *Case ID* выбирается тот параметр, для которого необходимо построение последовательности событий журнала.

Под атрибутом *Resource* понимаются субъекты, которые выполняют деятельность в рамках исследуемой последовательности событий.

Если необходимо произвести анализ перемещений транспортных средств, то в качестве *Case ID* будет выбран идентификатор транспортного средства. Тогда построенная общая модель будет показывать общие закономерности перемещений, а также возможным будет выделить последовательности событий отдельно для каждого транспортного средства. Это позволит отследить особенности маршрутов и других выбранных параметров. При этом управление транспортными средствами осуществляется водителями, т.е. в качестве атрибута *Resource* будет выбран идентификатор водителей.

Если необходимо отследить последовательность действий водителей, тогда в качестве атрибута *Case ID* будут выбраны идентификаторы водителей. В таком случае в качестве атрибута *Resource* корректно будет выбрать также идентификатор водителей, т.к. они выполняют действия в рамках последовательности событий. Если же есть необходимость исследования последовательность действий водителей в зависимости от управляемых ими транспортных средств.

В некоторых случаях может быть исследована комбинация идентификатора транспортного средства и идентификатора водителя в качестве атрибута *Case ID*, если необходимо построение последовательностей событий для транспортных средств, находящимися под управлением

определенного водителя. Атрибутом *Resource* в таком случае будет являться идентификатор водителя.

С. Исследование распространения явления в пространстве методами интеллектуального анализа процессов

Распространение различных явлений может регистрироваться с помощью различных источников и записываться в различные базы данных. В [6] показан случай, когда в качестве источников информации выступают не устройства, а пользователи социальной сети. В настоящее время социальные сети имеют широкое распространение, и информация о любом событии появляется в виде постов, комментариев в пабликах и на личных страницах пользователей.

Любое сообщение в социальной сети может быть описано с помощью следующих атрибутов:

- идентификатор пользователя (никнейм, id учетной записи, фамилия и имя);
- локация – может содержаться в тексте самого сообщения, в виде геометки или определяться с помощью функций социальной сети;
- время отправки сообщения;
- содержание сообщения.

Текст сообщения может содержать как текстовые данные, так и различные эмодзи, изображения, в том числе gif, а также видеозаписи.

Подход к анализу текстовых сообщений и выделению ключевых слов в соответствии с заданной тематикой представлен в работе [7].

Рассмотрим возможность анализа сообщений с помощью интеллектуального анализа процессов.

Если рассматривать назначение перечисленных выше атрибутов в соответствии с форматом .xes. Однозначным будет соответствие параметров *Timestamp* и «Время отправки сообщения». Однако другие параметры заслуживают более детального рассмотрения.

Если необходимо выстроить последовательности сообщений, написанные определенными пользователями, то в качестве *Case ID* следует выбрать идентификатор пользователя. В качестве атрибута *Resource* следует выбрать этот же параметр, так как действие написания сообщения выполняется пользователями. Атрибутом *Activity* будет являться ключевая информация из сообщений. В таком случае атрибут локации пользователя остается неучтенным, и при необходимости может осуществляться комбинация локации и ключевой информации из сообщения для составления атрибута *Activity*.

Однако если существует возможность перемещения пользователей между различными локациями, в качестве *Resource* может быть выбрана локация.

Если необходимо построения последовательностей событий для определенных локаций, что может представлять наибольший интерес в исследовании, так как позволяет отследить динамику изменения ситуации в отдельных районах, то в качестве атрибута *Case ID* следует выбирать локацию. Атрибутом *Resource* будет являться идентификатор пользователя, а *Activity* – информация из сообщений.

III. ЭКСПЕРИМЕНТЫ

Рассмотрим некоторые варианты построения моделей процессов в соответствии с описанным выше подходом.

На рис. 1 представлена модель процесса, построенная для транспортных средств. При формировании атрибута *Activity* была использована комбинация места наблюдения, типа совершенного действия и измеренного значения веса.

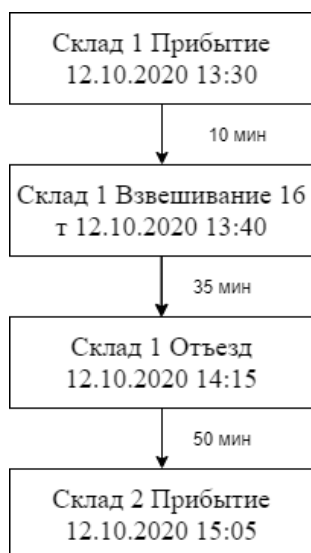


Рис. 1. Фрагмент модели перемещения транспортного средства

Рис. 2 представляет собой аналогичную модель, однако при формировании атрибута *Activity* для типов действий «Въезд» и «Выезд» были также использованы идентификаторы водителей. Это было сделано из следующих соображений: как правило, регистрация пропуска человека выполняется на контрольно-пропускных пунктах. Внутри ограниченной территории пропуск может не регистрироваться. Кроме того, увеличения количества параметров в одном блоке может перегрузить модель, и выбор трех, но разных для разных событий, параметров позволяет получить сбалансированную модель. Помимо этого, в большинстве случаев интерес представляет именно сопоставление водителей на определенном транспортном средстве при перемещении между разными территориями. То есть идентификатор водителя, покинувшего предыдущую территории на определенном транспортном средстве, должен совпадать с идентификатором водителя на этом же транспортном средстве в следующей локации. В противном случае это может указывать на несанкционированную смену водителя и возможное нарушение.

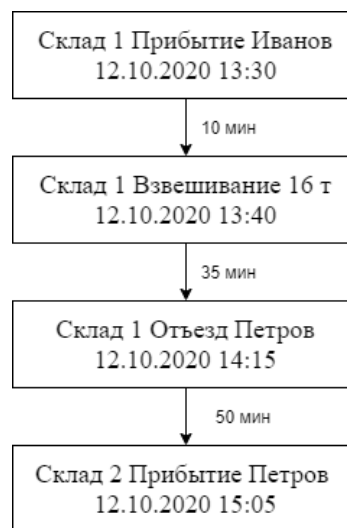


Рис. 2. Фрагмент модели перемещения транспортного средства с учетом фамилии водителя

На рис. 3 представлена укрупненная модель процесса распространения землетрясения в различных районах города. При построении модели в качестве атрибута *Activity* были выбраны локации.

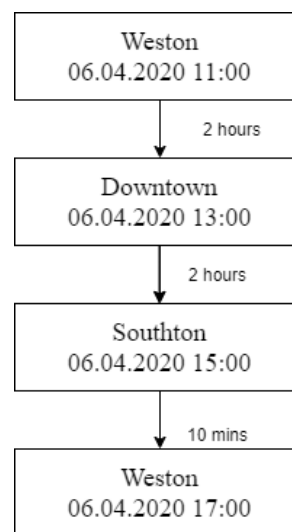


Рис. 3. Распространение эпицентров землетрясения по районам

IV. ЗАКЛЮЧЕНИЕ

В работе проведено исследование формирования атрибутов формата журнала событий при обработке и анализе разнородных данных. Показаны разные варианты назначения и комбинации атрибутов в зависимости от целей проводимого исследования.

В результате можно сделать вывод, что комбинация нескольких атрибутов для формирования атрибута *Activity* позволяет повысить информативность модели, при этом объем данных возрастает незначительно. Тем не менее при выборе ключевых атрибутов для интеграции их в *Activity* следует помнить, что весьма большое количество

атрибутов может перегружать модель и даже сделать затруднительным её визуальное восприятие.

Поэтому следует определять наиболее важные параметры для разных типов событий и осуществлять выбор нужных атрибутов в зависимости от типа события или места наблюдения и, следовательно, состава средств мониторинга.

СПИСОК ЛИТЕРАТУРЫ

- [1] Dakic D., Sladojevic, S., Lolic, T., & Stefanovic, D. Process mining possibilities and challenges: a case study //2019 IEEE 17th International Symposium on Intelligent Systems and Informatics (SISY). IEEE, 2019. C. 000161-000166.
- [2] Van Der Aalst W. et al. Process mining manifesto //International Conference on Business Process Management. Springer, Berlin, Heidelberg, 2011. C. 169-194.
- [3] Bekeneva Y.A. Algorithm for Generating Event Logs Based on Data from Heterogeneous Sources //2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). IEEE, 2020. C. 233-236.
- [4] Bekeneva Y.A. An Approach to the Distributed Generation of Event Logs Based on Data from Heterogeneous Monitoring Devices //2020 9th Mediterranean Conference on Embedded Computing (MECO). IEEE, 2020. C. 1-4.
- [5] Novikova E., Kotenko I., Murenin I. The Visual Analytics Approach for Analyzing Trajectories of Critical Infrastructure Employers // Energies. 2020. T. 13. №. 15. C. 3936.
- [6] VAST Challenge 2019: Disaster at St. Hemark, URL: <https://vast-challenge.github.io/2019/overview.html> (last accessed 28.06.2021)
- [7] Seal D., Roy U.K., Basak R. Sentence-level emotion detection from text based on semantic rules // Information and Communication Technology for Sustainable Development. Springer, Singapore, 2020. C. 423-430.