Algorithm for Generating Event Logs Based on Data from Heterogeneous Sources

Yana A. Bekeneva
Saint Petersburg Electrotechnical University "LETI"
Saint Petersburg, Russia
yana.barc@mail.ru

Abstract— **In recent years, process mining algorithms are widely used for process analysis. As input data for process mining algorithms, .xes files are used. This format has a limitation for a number of attributes; therefore, in case of registering a single event with several monitoring devices, there is problem of generating event logs based on heterogeneous data. In this paper, an algorithm for generating event logs based on data from heterogeneous monitoring devices is proposed. The most important parameters for the analysis of events are taken into account. Examples of the formation of event logs when choosing a different set of source data are given, the influence of the number and composition of the selected attributes on the result of building business process models is analyzed.**

Keywords—process mining; log files; .xes files; event logs

I. INTRODUCTION

Among the topics mentioned by Gartner [1] as important trends process mining was named as a future progressive trend. Using this technique current business processes can be built on the basis of real data, deviations in business processes are revealed, and a forecast of possible deviations during the current process is also made.

Event logs act as initial data, where each row corresponds to one event, and the columns unambiguously describe certain attributes.

Usually, event logs can conditionally be represented as follows:

- Case id: stores the cases (objects) for which the sequence of events is built.
- Activity name: stores actions performed as part of log events.
- Timestamp: stores the date and time of logging events.
- Resource: stores the main actors of the events of the journal (those who perform actions within the framework of the journal events).
- Other (other data): other information by which events are described.

The event log files [2] in the .xes format are used as input data for the process mining algorithms. Such files contain the listed attributes of the event logs, with the exception of other data. To build a business process model, attributes such as case identifier, activity, timestamp, and resource are used. Therefore, any source data set should be reduced to the event log format with a clear relation of one attribute from the source set to one attribute from event log set. So, for example, to build a business model of the behavior of moving objects in the source data there should be only one attribute identifying the moving object, and each line should be filled with a value corresponding to this event. Similarly, if there are several attributes in the source data that indicate a timestamp (for example, the time the event started and its end time), then when preparing the data, only the most significant attribute will be selected.

Depending on the goals of the process mining and the key characteristics of the process, it is necessary to select 4 main attributes and assign each of them some role provided in the .xes format.

II. APPROACH DESCRIPTION

The formation and analysis of events sequences can be performed for moving objects, places of the event, or a set of objects and places where the event occurred.

For forming a business process for moving objects, a process characterizes all actions sequentially performed by this object in different observation zones, including geographically distributed (spatially remote from each other).

When forming business processes for individual zones, a process characterizes the sequence of actions performed in the same zone with the participation of different moving objects.

Finally, the analysis of processes by the totality of a moving object and place can be used to build processes that characterize the behavior of a single moving object within a particular territory.

The basic principles of attribute assignment for different processes are presented in table 1.

TABLE I. BASIC ATTRIBUTE ASSIGNMENT

Attributes	Case id	Timestamp	Resource	Activity
By moving object	Moving object	Time	Moving object	Type
By zone	Zone	Time	Moving object	Type

This work was supported by the Russian Federation President Award SP-2581.2019.5.

978-1-7281-5761-0/20/$31.00 ©2020 IEEE

Attributes	Case id	Timestamp	Resource	Activity
By moving object in zone	Moving object + zone	Time	Moving object	Type

When forming data sets for building business processes, the researcher should keep in mind the following features, which are clearly visible from the table.

1. By default, the attribute that is responsible for the *case id* will always be understood as the attribute for which the process is being built. If the process should describe the sequence of actions performed by a moving object, then the attribute identifying the object should be selected as the corresponding attribute. Usually, as a result of integration of records about one event in the data set, there is an attribute that uniquely identifies a moving object and is filled with values for all events. If the data set includes information on the movements of different types of moving objects that are not interconnected, then the processes are built separately for each type of moving objects. In the case when it is required to build the process of moving an object in a certain area, a certain attribute should act as an identifier of the case, which can simultaneously identify a moving object and the place of the event. In this case, a similar attribute should be created optionally by merging the necessary fields.

2. By the attribute that is responsible for the *resource*, the default will always be understood as the identifier of the entity that triggers the event.

3. The attribute responsible for *activity* will always be understood as the type of event.

If there are a large number of event parameters that can be important in detecting deviations in processes, a number of difficulties may arise associated with the selection of several attributes from a large data set.

1. When building a business process using a moving vehicle, the subject identifier should be used both for the *case id* attribute (since this is the key attribute for which the sequence of events is built) and for the *resource* attribute (since this is an attribute pointing to the object that implemented event). Consequently, with such a construction of processes, there is no attribute that could contain an indication of the zone where the event occurred.

2. Various types of events can be characterized by important parameters that will remain unaccounted for when constructing business processes. For example, when transporting goods, the most important parameter is the quantitative and sometimes qualitative characteristics of the cargo. In addition, when analyzing the course of the process, such characteristics are one of the most important indicators in identifying deviations. So, for example, a certain parameter must have a constant value within the same sequence of events, or vice versa, change in accordance with some regularity.

For more informative data sets, the following solution is proposed.

Usually, the type of event is presented as a value that determines the type of action taken. Moreover, the action is characterized by some parameters, some of which are the most important. For moving objects the type of event is inextricably linked to the territory. For example, if the direction of movement ("Entry" / "Departure") is indicated as the type of event, then they can also be represented using the indication of the zone ("Entry to the warehouse" / "Departure from the reception department", etc.) . Thus, the following principles should be followed when determining the activity parameter.

If it is necessary to build a sequence of visits to different zones by the selected object, then the activity attribute should be selected not as an event type, but as a parameter indicating the visiting zone. In this case, business processes will be built where the zones sequentially visited by the same object will be indicated in chronological order.

If the parameter indicating the type of event is also important in the analysis, then, first of all, it is necessary to concatenate the fields indicating the type of event and the zone, which will more specifically describe the movements of the object. In this case, the activity attribute will look like, for example, "Entrance Warehouse" or "Departure Reception Hall".

Often, some additional events are also important, which may be different for different types of events. Therefore, to build visual models of business processes, some preliminary data preparation should be carried out.

First of all, you should define key parameters for each type of event. All key parameters can be described as a subset of the A^{key} attributes. Next, concatenate the fields that indicate the type of event and the most important parameters related to each type of event individually. In this way, new attributes describing the event can be obtained. For example, if the measured value of the weight is the most important for the event type "Weighting", then an attribute containing the name of the event and the measured weight: "Weighting 120" will be received as a new attribute.

All the most significant attributes can be similarly combined into a single attribute, which will be used as an attribute of activity for building a business process.

III. ALGORITHM DESCRIPTION

In general, the sequence of preliminary data preparation for applying the methods of process intellectual analysis to them is presented in Fig. 1. In addition, separate sequences of events can be constructed for different purposes of analysis. For example, models of moving a moving object within different zones and models of moving objects to analyze certain characteristics can be built independently of each other.

For example, if during the transportation of goods it is important to control the quantitative characteristics of the goods along the entire route, then the sequence of events can be built according to the following parameters:

- for the case id attribute identifier of vehicle will be presented;

- for the resource attribute the zone where the weight measurement took place will be presented;
- for the activity attribute the measured value of the quantity of the cargo characteristic will be presented.

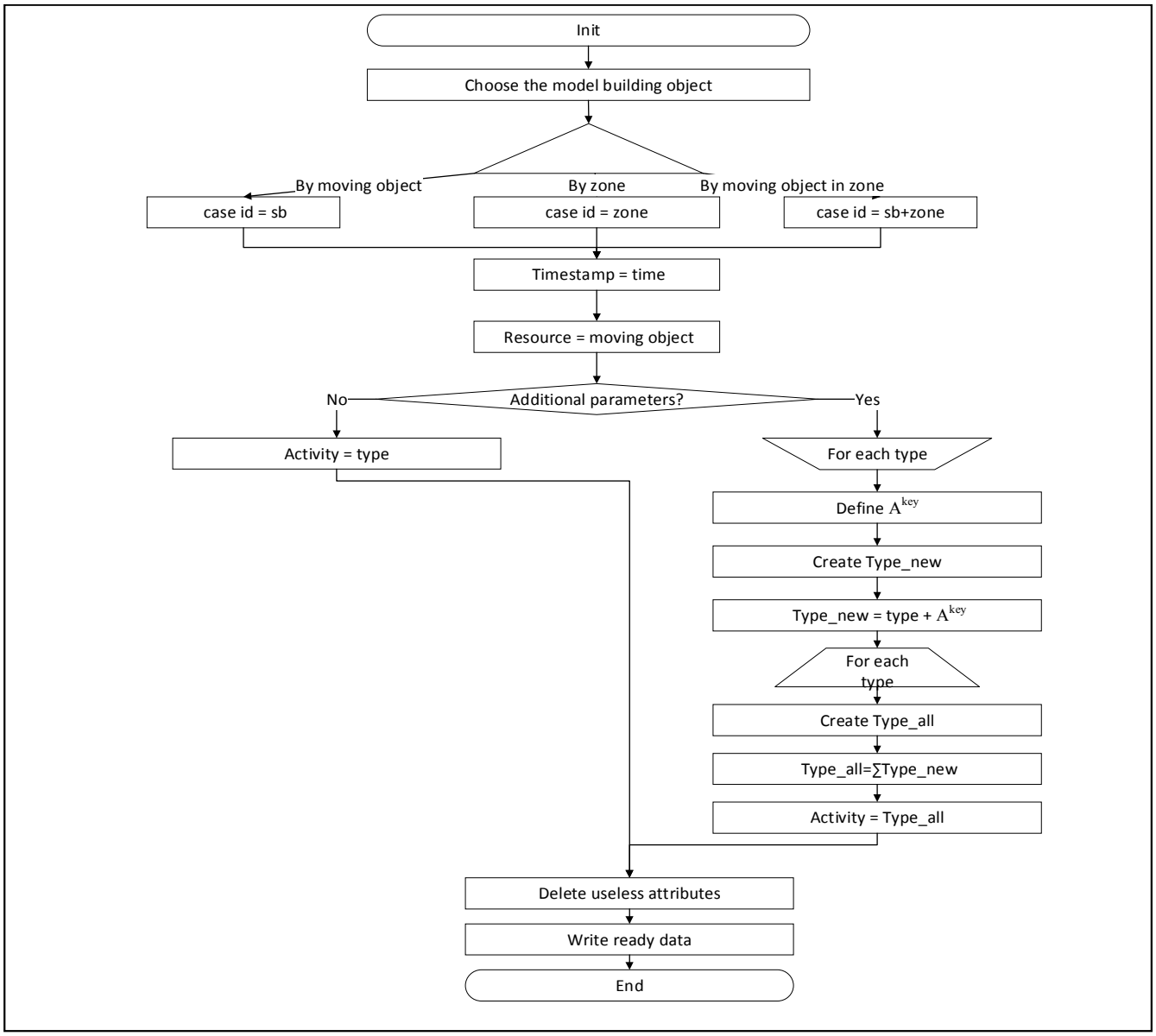

Fig. 1. Algorithm of event logs preparation

Thus, the researcher can get a visual representation of the change or preservation of a certain value of the selected parameter. In this case, deviations can be identified that are directly related to the actions that are directly related to this event parameter.

The standard business process model is represented as a set of blocks illustrated activities and connections between them. These blocks are placed according to action sequence. The name of activity is placed inside of blocks. It means that the researcher sees a sequence of activities. Another attributes are used for building the model. The main idea of the proposed algorithm is to create a complicated Activity attribute consisting of several parameters that are important for the current process analysis.

Using the presented algorithm it is possible to build datasets according to the aims of process analysis.

The view of a final dataset is shown in table 2.

TABLE II. PROPOSED ATTRIBUTE ASSIGNMENT

Attributes	Case id	Timestamp	Resource	Activity
By moving object	Moving object	Time	Moving object	Type+Zone+ p_gg/ Zone/ Zone+ p_gg
By zone	Zone	Time	Moving object	Type + p_gg
By moving object in zone	Moving object + zone	Time	Moving object	Type + p_gg

If it was decided to build various business processes to track the most key characteristics separately, then the preparation of data, the construction of training models and further comparison with them of data on the current process will be carried out independently.

IV. BUILDING A PROCESS MODEL

For building a process model the data from real enterprise were used. The data are obtained from heterogeneous monitoring devices such as cameras, access control systems and so on. The control objects are trucks moving between several distributed zones.

For data preparation RapidMiner Studio [3] was used. For process model building a Fluxicon Disco tool was used [4].

The figure 2 shows a process model where only action type was used for the *Activity* attribute.

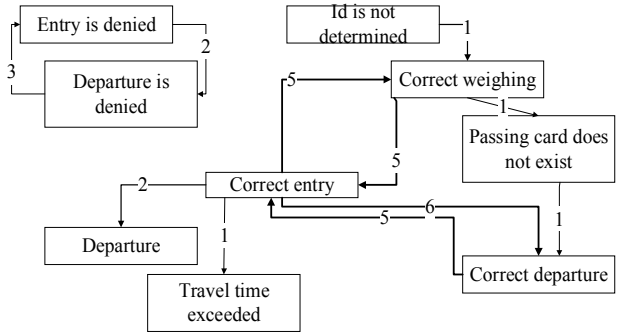

Fig. 2. Process model for type of event

Looking at the picture we can conclude that it is impossible to determine the zones in which events were recorded. The presented model allows only to evaluate the possible sequence of event types without reference to a specific zone. However, different sequences of events may occur in different zones. Thus, the introduction of an additional attribute is appropriate.

The figure 3 shows a process model where action type and zone were used for the Activity attribute.

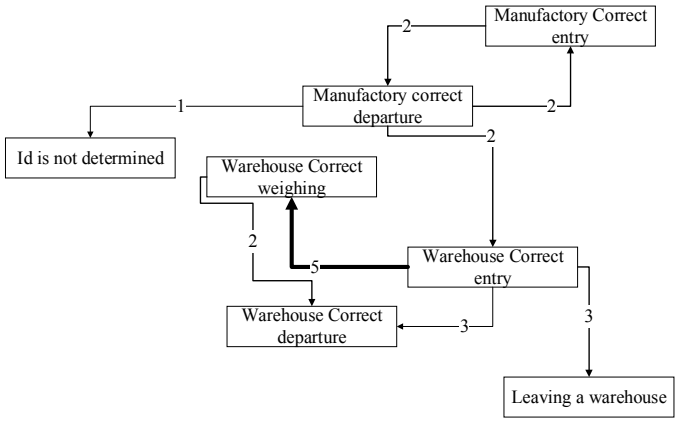

Fig. 3. Process model for type of event

The process model on the figure 3 is more clear and informative that the model on the figure 2.

The study of processes by methods of process intelligence allows you to visualize various scenarios of object behavior. When comparing data with the obtained standard models, it is determined whether an event corresponds to the scenarios presented in the model. During the analysis of processes as a sequence of movements of moving objects, deviations such as omissions of necessary actions or zones, non-compliance with a typical route, atypical duration of stay in a certain zone, etc. can be detected. Adding additional parameters allows you to identify whether the current process corresponds or does not correspond to a typical process, taking into account various characteristics describing the event.

V. CONCLUSION

The study of processes by methods of process mining makes possible to visualize various scenarios of object behavior. When comparing data with the obtained standard models, it is determined whether an event corresponds to the scenarios presented in the model. During the analysis of processes as a sequence of movements of moving objects, deviations such as omissions of necessary actions or zones, non-compliance with a typical route, atypical duration of stay in a certain zone, etc. can be detected. Adding additional parameters allows you to identify whether the current process corresponds or does not correspond to a typical process, taking into account various characteristics describing the event.

REFERENCES

[1] Gartner Tech Trends for 2019 – With Process Mining into the Future, URL: https://lanalabs.com/en/gartner-trends-for-2019/

[2] Aalst W. M. P., van der. Extracting event data from databases to unleash process mining //BPM-Driving innovation in a digital world. – Berlin Heidelberg: Springer, Cham, 2015. – P. 105-128.

[3] RapidMiner: Data Science Platform. URL: https://rapidminer.com/

[4] Process Mining and Automated Process Discovery Software for Professionals – Fluxicon Disco. URL: https://fluxicon.com/disco/