УДК 004.056
DOI: 10.15827/0236-235X.126.197-206
Дата подачи статьи: 09.11.18
2019. Т. 32. № 2. С. 197–206

Преобразование данных от разнородных систем мониторинга

Я.А. Бекенева [1], *аспирант, yana.barc@mail.ru*

[1] *Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), г. Санкт-Петербург, 197376, Россия*

В статье представлен подход к подготовке данных, получаемых от разнородных систем мониторинга, для их дальнейшего анализа методами интеллектуального анализа данных. Основной проблемой анализа данных при мониторинге различных процессов являются различия в описании событий для разных типов источников, в том числе формат представления данных. Кроме того, одно и то же событие может быть описано с помощью данных от разных систем мониторинга.

В настоящей работе приведена формальная модель анализируемого процесса, описаны основные проблемы анализа разнородных данных, выделены формальные критерии отнесения записей от разных источников к одному событию. В предлагаемом подходе в качестве источников данных учитываются не только записи, поступающие от различных мониторинговых систем в режиме реального времени, но и учетные базы, используемые для хранения информации. Ключевая идея состоит в том, что движущиеся объекты разных типов могут совершать действия как единое целое в рамках исследуемой задачи (например, транспортное средство и водитель). Использование учетных систем позволяет найти взаимосвязи между подобными движущимися объектами и тем самым повысить точность объединения записей, относящихся к одному событию.

Предложенный подход был опробован на реальных данных, полученных от предприятия. В результате применения всех описанных преобразований удалось существенно сократить избыточную размерность совокупной таблицы с данными, а также значительно снизить количество пропущенных значений. Данные, анализ которых был затруднительным из-за их разного формата, приведены к единому формату и представлены в виде единой таблицы, удобной для дальнейшего исследования методами интеллектуального анализа данных.

Ключевые слова: *разнородные источники, системы мониторинга, преобразования данных, атрибуты, события, формат данных.*

Современные предприятия и организации имеют сложную информационную инфраструктуру, которая, как правило, включает в себя большое количество разнородных источников данных, в том числе связанных с мониторингом процессов [1]. Одной из частных задач мониторинга процессов является мониторинг перемещений движущихся объектов на определенной территории и (или) между распределенными зонами, входящими в состав сложного объекта.

Как правило, для этого используются различные системы мониторинга, включающие системы контроля доступа, средства фото- и видеофиксации и т.п., преимущественно применяющиеся для получения данных о таких процессах, как попадание движущегося объекта на закрытую территорию (вход, въезд), перемещение из одной зоны в другую, а также покидание закрытой территории (выход, выезд). Системы мониторинга разных типов имеют разные базы для записи и хранения данных [2].

Все данные, получаемые от систем мониторинга, направляются на центральный узел сбора данных. Они поступают в разном формате, который зависит от типа средства мониторинга, зафиксировавшего информацию. Как правило, на центральном узле осуществляются некоторые первичные преобразования данных с целью приведения к виду, удобному для представления и хранения, однако их формат по-прежнему остается разным и зависит от типа источника.

Разнообразные средства мониторинга и контроля зачастую фиксируют определенный набор параметров при возникновении того или иного события. При этом несколько средств мониторинга могут одновременно зафиксировать одно и то же событие, каждый со своим набором параметров. Как правило, получаемые от различных датчиков «сырые» данные оказываются непригодными для анализа существующими методами и нуждаются в ряде предварительных преобразований.

При подготовке данных для применения к ним средств интеллектуального анализа данных необходимо решить следующие задачи:

– привести данные к единому формату, пригодному для анализа;

– интегрировать записи, относящиеся к одному событию, в одну запись;

– при большом количестве разных типов событий, характеризующихся разным составом атрибутов, выделить группы событий со схожим составом атрибутов.

В данной работе представлен метод подготовки данных от разнородных источников к единому формату. Предлагаемый метод доработан по сравнению с первоначальной версией, представленной в работе [3]. В качестве источников данных учитываются не только датчики систем мониторинга, но и учетные базы, данные из которых используются для сопоставления субъектов разных типов с целью более точного выделения записей, относящихся к одному событию. Метод был опробован на реальных данных, полученных от предприятия.

Обзор существующих методов

Рассмотрим известные методы подготовки данных, поступивших от разнородных источников.

Задача интеграции таких данных описана в работе [4], где авторы предлагают метод, основанный на онтологии и определении семантической близости концептов. Метод позволяет группировать семантически близкие записи и структурировать БД для более удобной обработки информации, однако не выявляет взаимосвязи между записями, относящимися к одному событию.

В работе [5] предлагается способ приведения XML-данных к единому формату на основе теории графов, деревьев решений и XSTL-преобразований. Метод используется для решения достаточно узкого круга задач (анализа данных, поступающих от веб-сервисов). Кроме того, он решает лишь задачу приведения данных к единому формату.

Задача интеграции разнородных данных, поступающих от дорожных датчиков и GPS-треков, решается в [6]. В работе учитываются взаимосвязи между записями от различных источников, относящимися к одному событию. Авторы ставят задачу моделирования транспортных потоков на основе реальных данных, поэтому задача приведения их к формату, удобному для последующего анализа, в рамках исследования не рассматривается.

В работе [7] описан основанный на гауссовой графической модели метод объединения данных от разнородных датчиков, каждый из которых способен фиксировать разнородные параметры. Метод предлагает сократить размерность данных, однако не предусматривает корреляцию различных записей и предназначен лишь для сбора информации и ее удобного представления.

Алгоритмы корреляции событий [8, 9] решают задачу объединения данных от разнородных источников и установления взаимосвязей между ними, однако, как правило, в контексте задачи обеспечения безопасности вычислительных сред разнородными источниками являются различные компоненты сети, при этом данные имеют единый формат и не требуют дополнительных преобразований для дальнейшей обработки и анализа.

Формальная модель процесса

Приведенная далее формальная модель процесса является доработанной и уточненной версией модели, представленной в [10]. Добавлено описание зон, введены подмножества движущихся объектов, подвергаемых наблюдению, а также объектов мониторинга. Уточнены параметры атрибутов и формальное описание событий с помощью атрибутов.

Производственные объекты, как распределенные, так и расположенные в пределах определенной территории, как правило, представляют собой конечное множество отдельно взятых зон: $Z = \{z_1, z_2, \ldots, z_n\}$.

Зоны могут быть выделены изначально в зависимости от их расположения, функционального назначения или же заданы в процессе решения конкретной задачи в зависимости от ее условий.

На производственном объекте могут происходить различные фиксируемые события. Их можно представить конечным множеством: $E = \{e_1, e_2, \ldots, e_n\}$.

В контексте решаемой задачи событиями считаются разного рода перемещения движущихся объектов, составляющих множество SB: $SB = \{sb_1, sb_2, \ldots, sb_k\}$.

Можно выделить различные типы движущихся объектов, подвергаемых мониторингу: транспортные средства (автомобили, автобусы, поезда и т.д.), люди (сотрудники организации,

студенты учебных заведений и т.д.) и др. Поэтому в зависимости от процесса и участвующих в нем перемещающихся объектов можно выделить разные подмножества движущихся объектов: $SB = \{SB^1, SB^2, \ldots, SB^m\}$.

Для фиксации событий используются различные средства мониторинга SM: $SM = \{sm_1, sm_2, \ldots, sm_j\}$.

В зависимости от характера наблюдаемых процессов на предприятиях используются средства мониторинга различных типов, которые составляют подмножества: $SM = \{SM^1, SM^2, \ldots, SM^b\}$.

Различные системы мониторинга, как правило, распределены среди всех анализируемых зон или некоторых из них. Каждый отдельно взятый элемент системы мониторинга (конкретная фото- или видеокамера, датчик контроля пропускной системы и пр.) имеет свой уникальный идентификатор. При анализе данных этот параметр может быть использован для определения места, где было зафиксировано событие.

В качестве источников данных DS могут выступать как средства мониторинга SM, так и различные виды документации или дополнительные БД: $DS = \{DS^{doc}, DS^{bd}, \ldots, DS^{sm}\}$.

В некоторых DS может содержаться информация от отдельно взятого средства мониторинга (sm = const), в то время как в других DS консолидироваться информация от разных средств мониторинга ($sm = var$).

Информация об одном и том же событии может содержаться в нескольких DS и, таким образом, представляться в виде нескольких различных записей.

Каждая запись содержит определенный набор атрибутов из множества атрибутов A и значений этих атрибутов. Набор атрибутов обычно определяется типом средства мониторинга SM.

Как правило, набор атрибутов в любой записи содержит следующую информацию:

– временные характеристики события (t);

– тип совершенного действия ($type$);

– средство мониторинга (sm), также указывающее на место совершения события (z);

– сущность, совершившая действие (sb);

– объект, над которым было совершено действие (sv);

– дополнительные параметры ($p_g g$).

Таким образом, множество атрибутов можно условно разбить на подмножества: $A = A^t \cup A^{type} \cup \cup A^{sm} \cup A^z \cup A^{sb} \cup A^{sv} \cup A^{p_g g}$.

В одних случаях атрибуты, указывающие на место совершения события (A^z), могут быть вынесены в отдельный атрибут, в других – определяться только принадлежностью средства мониторинга sm конкретной зоне z. Как правило, данные от средств мониторинга sm должны позволять определять моменты вхождения в ту или иную зону, нахождения внутри зоны и выхода из нее.

Зачастую наименования атрибутов, описывающих один и тот же параметр, могут отличаться в разных типах источников данных. Кроме того, разные средства мониторинга могут фиксировать различные параметры, идентифицирующие сущность, совершившую действие: $a^{sb} = \langle a^{sb\_\alpha}, a^{sb\_\beta}, \ldots, a^{sb\_\gamma} \rangle$.

В ходе мониторинга также могут возникать непредвиденные ситуации, в результате которых определение некоторых параметров может быть затруднено, то есть возникает проблема пропуска данных или их нечеткость.

При фиксации одного и того же события разными средствами мониторинга необходимо принимать во внимание возможные временные задержки, возникающие из-за различных особенностей выполнения исследуемых действий, размещения объектов мониторинга и иных факторов, обусловливающих неодновременность создания записей. В связи с этим для каждого типа выполняемых действий необходимо ввести допустимые временные задержки τ.

Таким образом, отдельно взятое событие e_i может быть описано с помощью набора атрибутов:

$$e_i = (\{a_1^t, \ldots, a_x^t\}, \{a_1^{type}, \ldots, a_y^{type}\}, \{a_1^{sm}, \ldots, a_j^{sm}\}, \{a_1^{sb}, \ldots, a_k^{sb}\}, \{a_1^{sv}, \ldots, a_w^{sv}\}, \{a_1^{p_g g}, \ldots, a_u^{p_g g}\}).$$

При анализе такого рода данных возникают следующие проблемы:

– определение типа события в зависимости от состава атрибутов;

– синхронизация временных параметров с целью точной группировки данных для описания отдельно взятого события;

– дублирование, нечеткость или пропуск данных.

Таким образом, для анализа событий и выявления среди них возможных нарушений необходимо решить следующие задачи:

– выполнить общие преобразования, позволяющие привести данные к виду, пригодному для дальнейшего анализа; так, для методов Data Mining необходим набор в виде

матрицы (таблицы), где каждая строка представляет собой описание события, а столбцы – атрибуты этого события;

– записи, поступившие от разных систем мониторинга и относящиеся к одному и тому же событию, должны быть сгруппированы, для чего, прежде всего, необходимо определить критерии, по которым следует выделять и объединять такие записи.

Общие преобразования

Любое событие e_i происходит в определенной точке пространства в определенный момент времени [5]. При этом событие может быть инициировано одним субъектом или группой субъектов. В рамках данной работы будут рассмотрены события, инициированные одним субъектом sb_k.

Таким образом, каждое событие имеет набор атрибутов A, относящихся к субъекту, инициировавшему событие (A^{sb}), месту (A^z) и времени совершения события (A^t). Такие атрибуты, а также атрибуты, повторяющиеся для всех без исключения типов событий, можно назвать общими и выделить их в группу Common (рис. 1). Разные средства мониторинга могут описывать события с помощью разных наборов атрибутов. Поэтому описания событий могут отличаться друг от друга и зависеть от типа события и средств, которые фиксируют такие типы событий. Вариативные атрибуты могут быть выделены в группу Variable.

Особенностью такого рода баз является различное представление событий. В таблице Common, как правило, каждое событие, зафиксированное той или иной системой мониторинга, описывается одной строчкой, а атрибуты расположены по столбцам. В таблице Variable, напротив, каждая строка описывает отдельно взятый атрибут и соответствующее ему значение.

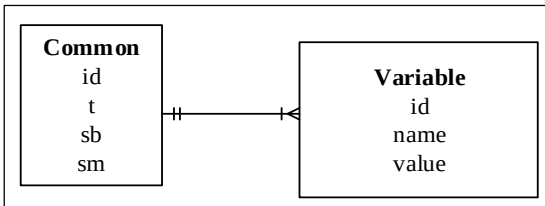

Рис. 1. Структура хранения данных об инцидентах

Fig. 1. Incident data storage structure

Для комплексного анализа событий необходимо интегрировать данные, полученные от всех типов источников, в единую таблицу. Для БД, где данные хранятся так, как показано на рисунке 1, прежде всего необходимо привести две разнородные таблицы к общему виду. К таблицам из группы Variable предлагается применить преобразование *pivot*. Это позволит расположить атрибуты по столбцам и записать соответствующие им значения в виде одной строки. Далее следует объединить таблицы Common и Variable по атрибуту, идентифицирующему событие, как показано на рисунке 2.

В результате будет получена общая таблица UDS_SM, содержащая множество записей от всех возможных средств мониторинга. Каждая строка такой таблицы соответствует одной записи, полученной от какого-либо источника данных, а столбцы описывают все возможные атрибуты, фиксируемые всеми имеющимися средствами фиксации инцидентов.

Таблица, тем не менее, является лишь промежуточным вариантом обработанных данных, так как при таком представлении данных имеется ряд нерешенных проблем.

Во-первых, каждая строка такой таблицы представляет собой единственную запись от конкретной системы мониторинга для конкретного события. Если одно событие может быть одновременно зафиксировано несколькими системами мониторинга, то в таблице оно будет

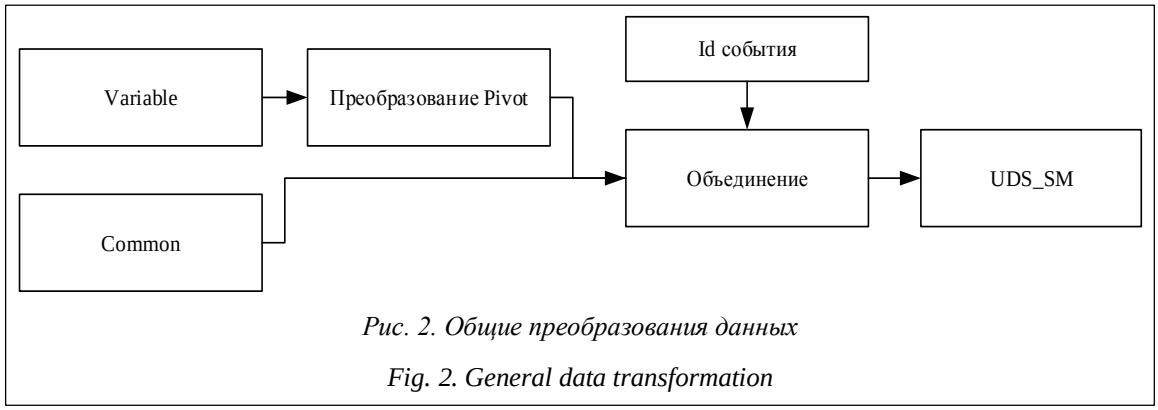

Рис. 2. Общие преобразования данных

Fig. 2. General data transformation

описано с помощью нескольких строк, при этом число строк будет равным числу систем мониторинга, зафиксировавших событие.

Во-вторых, для каждой системы мониторинга может существовать свой набор наименований одних и тех же параметров, то есть обозначения параметров могут отличаться в зависимости от типа системы мониторинга. Таким образом, в подобного рода таблице будет присутствовать большое количество атрибутов, имеющих разные названия, но дублирующих друг друга по смыслу, что приведет к большому количеству пропущенных значений.

Для решения описанных проблем следует сделать ряд преобразований, позволяющих объединить записи, относящиеся к одному событию, тем самым снизив избыточность представления данных и значительно уменьшив число пропущенных значений.

Объединение записей, описывающих одно и то же событие

Если событие e_i может быть описано с помощью нескольких записей от разных систем мониторинга, то такие записи должны быть сформированы в одно и то же время источниками данных, расположенными в одном и том же месте, и указывать на один и тот же субъект, инициировавший событие.

Пусть некоторый субъект sb_1 совершает действия, фиксируемые системами мониторинга sm_1, sm_2, sm_3. Каждая из этих систем генерирует записи с определенным набором параметров. Допустим, данные, формируемые системой sm_1, описываются таким набором атрибутов, как $< a_1^t, a_2^{type}, a_1^{sm}, a_4^{sb} >$. Аналогично данные, формируемые системой sm_2, описываются набором атрибутов $< a_3^t, a_1^{type}, a_2^{sm}, a_k^{sb}, a_1^{sv} >$, а данные, формируемые системой sm_3, – $< a_1^t, a_1^{type}, a_3^{sm}, a_4^{sb}, a_1^{sv}, a_{add} >$.

Объединение всех записей представлено в таблице 1.

Таблица 1

Формат данных после объединения записей в таблицу UDS_SM

Table 1

Data format after merging records into UDS_SM table

a_1^t	a_2^{type}	a_1^{sm}	a_4^{sb}	a_3^t	a_1^{type}	a_2^{sm}	a_k^{sb}	a_1^{sv}	a_3^{sm}	a_{add}
X	X	X	X							
				X	X	X	X	X		
X			X		X			X	X	X

В таблице показан вариант, когда в полученном файле имеются одновременно два атрибута, идентифицирующих субъект, инициировавший событие. Формат записи может как отличаться (один из атрибутов может указывать на субъект в виде слова, другой – на идентификатор в виде номера и пр.), так и быть одинаковым для всех трех колонок. В некоторых случаях исходная строка может содержать и два атрибута, указывающих на субъект: например, цифровой идентификатор и соответствующая ему фамилия, номер транспортного средства и пр. Таким образом, на данном этапе необходимо сначала найти и объединить атрибуты, описывающие один и тот же смысловой параметр.

В первую очередь, нужно объединить все повторяющиеся атрибуты, указывающие на один и тот же параметр и имеющие общий формат записи, при этом сами значения должны совпадать.

Например, при получении таблицы подобного вида необходимо выделить все атрибуты, указывающие на субъект, и объединить их в один общий атрибут. Для этого можно как создать совершенно новый атрибут, идентифицирующий субъект, так и выбрать один из уже существующих. В приведенном примере будет создан новый атрибут *value_sb*. Аналогично должны быть проанализированы и преобразованы все атрибуты с общим смыслом.

При анализе отдельно взятых событий или отдельных типов событий важно понимать, из каких процедур, фиксируемых средствами мониторинга, состоит это событие. Необходимо определить последовательность этих процедур и временные задержки между ними. Например, при входе в офисное здание сотрудник организации прикладывает пропуск на входе, а через несколько секунд его лицо попадает в объектив камеры наблюдения. Следовательно, необходимо понимать, что временные атрибуты записей, относящихся к одному событию, могут отличаться на определенное значение, а не быть одинаковыми. Поэтому на данном этапе временные атрибуты следует оставлять неизменными.

В некоторых случаях процесс может включать в себя события, совершаемые разными типами субъектов (например, автобусы и поезда), действия которых фиксируются разными средствами контроля. Параметры записей при этом могут существенно различаться для разных типов субъектов. В таких случаях количество ат-

рибутов, идентифицирующих субъект, будет равно количеству типов субъектов.

Результаты преобразования представлены в таблице 2.

Таблица 2

Формат данных после объединения одинаковых по смыслу атрибутов

Table 2

Data format after combining attributes with the same meaning

a_1^t	*type*	a_1^{sm}	*value_sb*	a_3^t	a_2^{sm}	a_1^{sv}	a_3^{sm}	a_{add}
X	X	X	X					
	X		X	X	X	X		
X	X		X			X	X	X

Очевидно, что при большом количестве атрибутов с одинаковым смысловым значением таблица такого вида является более наглядной и удобной. Она не перегружена лишними параметрами, а количество пропущенных значений будет существенно меньше, чем в таблице 1. Тем не менее, в такой таблице еще имеется некоторое количество пропущенных значений, что по-прежнему создает неудобства при ее анализе.

Как было сказано ранее, любое событие имеет три наиболее важные характеристики:

- время совершения события (*t*);
- место, где событие было совершено (*z*);
- субъект, инициировавший событие (*sb*).

В общем случае при анализе и группировке некоторого количества (*d*) записей, которые лишь частично описывают одно и то же событие, важно следовать трем основным правилам.

1. Временные атрибуты события должны совпадать или разница между ними не должна превышать допустимую задержку τ:

$$\max(a_1^t, ..., a_d^t) - \min(a_1^t, ..., a_d^t) = \Delta t \le \tau.$$

2. Должны совпадать пространственные атрибуты события: $a_1^{sm} \to z_v, ..., a_d^{sm} \to z_v$.

3. Субъект должен быть одним и тем же: $a_1^{sb} = a_1^{sb_k}, ..., a_d^{sb} = a_d^{sb_k}$.

Однако не всегда можно однозначно соотнести данные от разных источников, особенно, если в некоторых записях отсутствуют временные параметры или же зафиксированы идентификаторы разных объектов наблюдения (например, номер автомобиля и идентификатор пропуска, записанный на определенного водителя). В таком случае необходимо сопоставлять информацию от дополнительных источников данных, например, учетных систем.

В общем виде алгоритм группировки данных может быть таким, как представлен на рисунке 3. После выполнения данного преобразования каждый отдельно взятый набор данных (табл. 2) будет иметь вид, представленный в таблице 3.

Таблица 3

Формат данных после преобразований

Table 3

Data format after transformation

time	*type*	*zone*	*value_sb*	a_1^{sv}	a_{add}
X	X	X	X	X	X

Полученные данные пригодны для дальнейшего анализа методами Data Mining.

Пример применения подхода

Для проверки применимости алгоритмов интеллектуального анализа данных к данным, преобразованным с помощью предложенного метода, были проведены эксперименты в среде RapidMiner версии 8.001 [11]. Среда RapidMiner представляет собой удобный инструмент анализа данных с простым и понятным пользовательским интерфейсом. Процесс обработки данных реализуется в виде набора блоков, выбираемых пользователем в зависимости от решаемой задачи. При необходимости настройки каждого блока могут быть изменены. Несомненными преимуществами использования среды RapidMiner являются простота и удобство реализации экспериментов, так как от пользователя не требуется знание языков программирования, а блоковая структура процесса позволяет быстро внести изменения в процесс.

Для проведения экспериментов использовалась выборка данных, полученная от предприятия. Источниками данных являлись такие системы мониторинга, как системы контроля доступа, видеокамеры, измерительные системы, кроме того, имелись учетные таблицы о соответствии номеров пропусков фамилиям сотрудников и о соответствии фамилий водителей номерам управляемых ими транспортных средств. Выборка представляла собой две таблицы из БД. Таблица incident содержала данные о зафиксированных событиях, то есть являлась аналогом таблицы Common (рис. 1), таблица incident_attributes содержала наименования вариативных атрибутов и их значения, то есть являлась аналогом таблицы Variable (рис. 1).

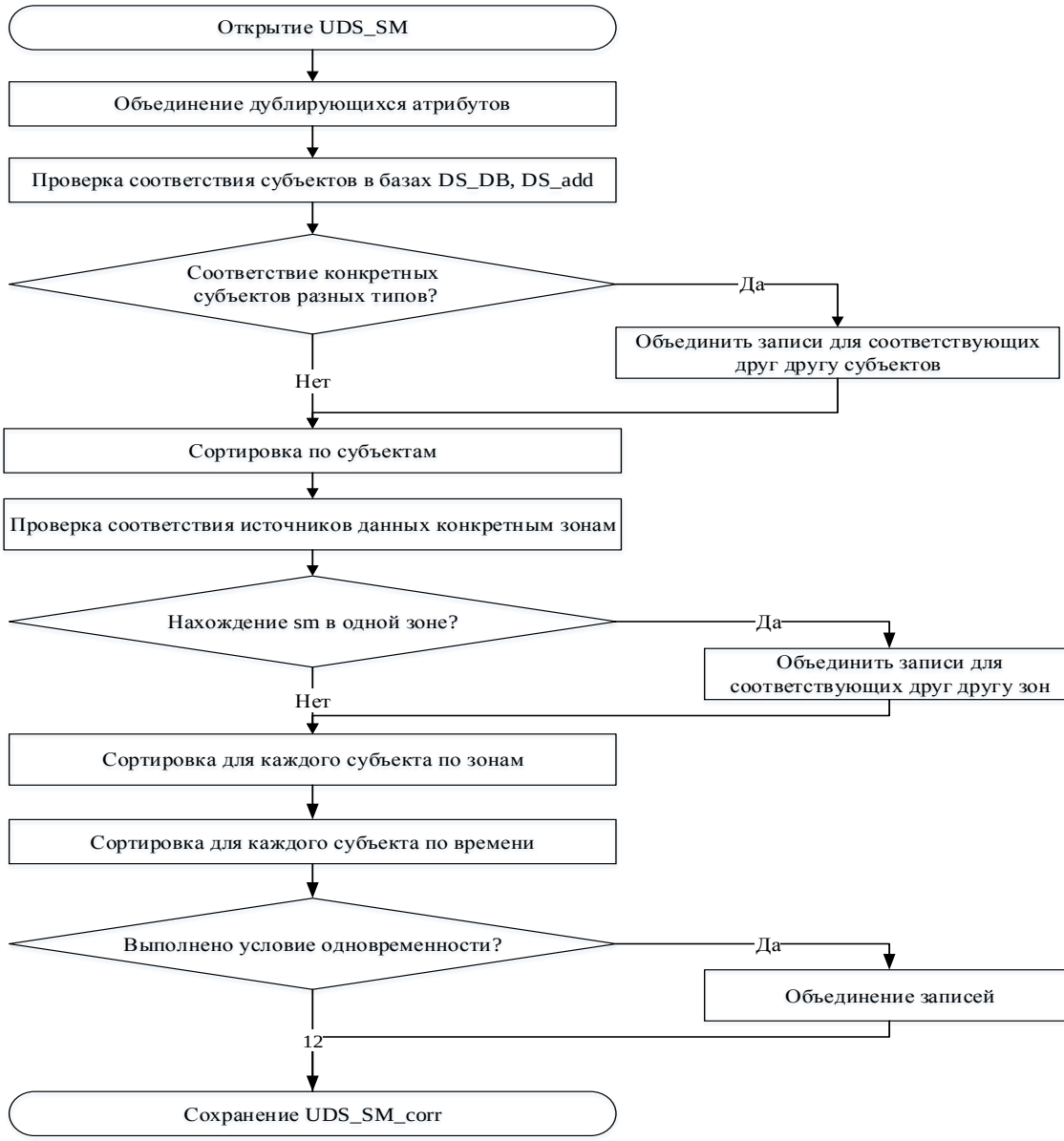

Рис. 3. Алгоритм объединения записей, относящихся к одному событию

Fig. 3. The algorithm for combining records related to one event

Реализация данного алгоритма в среде RapidMiner показана на рисунке 4.

Операторы *Read CSV* используются для чтения исходных таблиц incident и incident_attributes.

После выбора необходимых атрибутов *(incident_id, value, name)* в блоке *Select Attributes* к таблице incident_attributes было применено преобразование *Pivot*, после чего преобразованная таблица была объединена (оператор *Join*) с таблицей incident по идентификатору инцидента *incident_id*. В результате получена общая таблица, состоящая из 5 493 строк, соответствующих инцидентам, и 92 столбцов, соответствующих атрибутам. В таблице представлены данные, описывающие разные типы событий, инициированные разными типами субъектов.

Далее был выполнен ряд операций, связанных с группировкой атрибутов. Сначала выбраны атрибуты (*Select Attributes*), имеющие общее смысловое значение, но записанные в разные столбцы общей таблицы. В некоторых случаях было осуществлено переименование атрибутов (*Rename*), а потом созданы новые атрибуты (*Generate Attributes*), при этом каждый

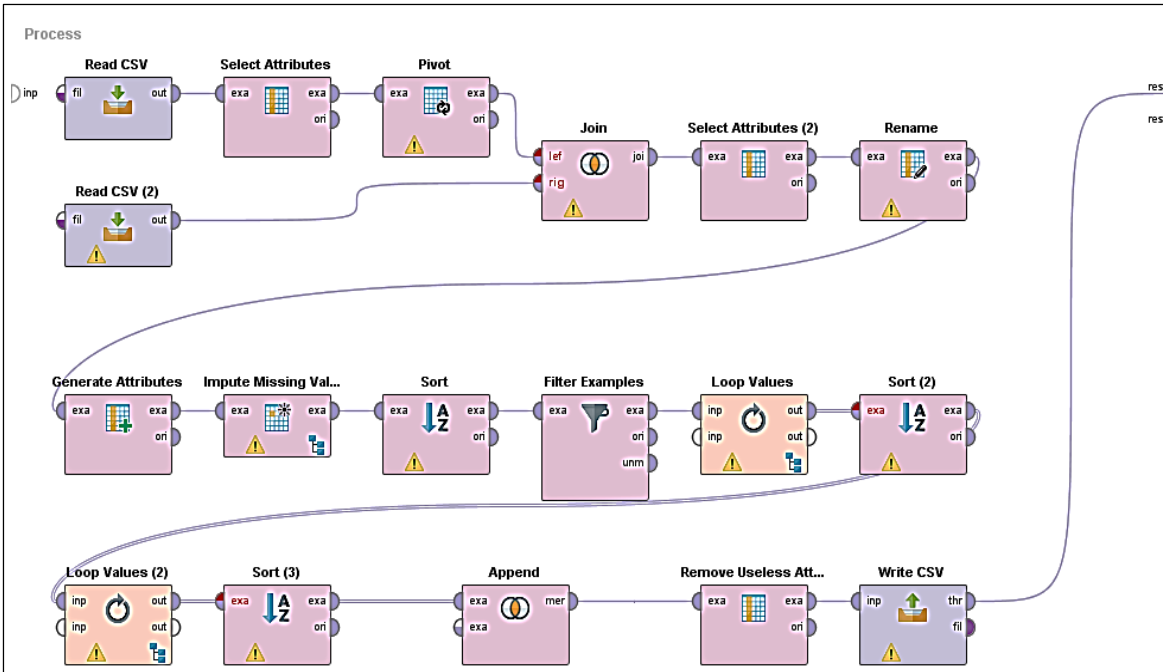

Рис. 4. Реализация подхода в среде RapidMiner

Fig. 4. The implementation of the approach in RapidMiner

из них представлял собой совокупность значений заданных атрибутов, имеющих общий смысл. После проведения этого этапа была получена таблица, число столбцов которой сократилось до 47.

Затем была проведена группировка записей об одном событии. Для этого таблица, полученная на предыдущем шаге, была подана на вход оператора *Impute Missing Values*. Внутри этого блока было задано использование таблицы соответствия номеров транспортных средств и фамилий водителей в качестве опорных данных для заполнения пропущенных значений атрибутов. С помощью оператора *Filter Examples* были произведены фильтрация записей, относящихся к исследуемому субъекту, и сортировка записей по идентификатору субъекта (*Sort*). Далее с помощью циклов *Loop Values* последовательно произведена сортировка (*Sort*) по зоне наблюдения и по времени совершения действия соответственно. Затем с помощью оператора *Append* объединены записи, удовлетворяющие условию одновременности.

В заключение были удалены изначально не несущие смысловой нагрузки атрибуты или атрибуты с общим смысловым значением, объединенные ранее в один общий атрибут (*Remove Useless Attributes*). После всех описанных этапов осуществлена запись итогового файла с помощью оператора *Write CSV*. Полученная таблица имела размерность 2 598 строк и 32 столбца.

Заключение

В данной работе рассмотрены основные проблемы анализа данных о мониторинге процессов с помощью разнородных систем мониторинга. Предложен подход к объединению данных от разных источников и их представлению в едином формате, удобном для дальнейшего анализа. Выделены критерии объединения записей от различных источников при описании одного и того же события. Показано, что предложенный подход помогает существенно снизить избыточность представления данных, а также снизить число пропущенных значений в совокупной таблице за счет объединения одинаковых по смыслу атрибутов и объединения записей, относящихся к одному событию. С помощью предложенного подхода может быть получена единая таблица, каждая строка которой наиболее полно описывает отдельно взятое событие. Такая таблица может быть проанализирована доступными методами интеллектуального анализа данных с целью выявления аномальных действий или грубых нарушений регламента.

Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации в рамках государственного задания «Организация научных исследований», задание № 2.6113.2017/6.7.

Литература

1. Dumas M., La Rosa M., Mendling J., Reijers H. Fundamentals of business process management. Heidelberg: Springer, 2013, vol. 1, p. 2.

2. Bastani K., Rao P.K., Kong Z. An online sparse estimation-based classification approach for real-time monitoring in advanced manufacturing processes from heterogeneous sensor data. IIE Transactions, 2016, vol. 48, no. 7, pp. 579–598. DOI: 10.1080/0740817X.2015.1122254.

3. Кузнецов Д.П., Ржеуцкая С.Ю. Метод интеграции онтологий разнородных источников данных в АСУП // Вестн. Череповец. гос. ун-та. 2013. Т. 1. № 4. С. 51.

4. Сапунов Н.О. Интеграция разнородных источников данных посредством xml Web-сервисов при организации управления транспортным процессом // Вестн. гос. ун-та морск. и реч. флота им. адм. С.О. Макарова. 2010. № 4. С. 8.

5. Старожилец В.М., Чехович Ю.В. Комплексирование данных из разнородных источников в задачах моделирования транспортных потоков // Машинное обучение и анализ данных. 2016. Т. 2. № 3. С. 260–275. DOI: 10.21469/22233792.2.3.01.

6. Dong C., Xiuquan Q., Gelernter J., Li Xiaofeng, Meng Luoming. Mining data correlation from multi-faceted sensor data in the Internet of Things. China Communications, 2011, vol. 8, no. 1, pp. 132–138.

7. Mirheidari S.A., Arshad S., Jalili R. Alert correlation algorithms: a survey and taxonomy. Proc. 5th Intern. Symp. on Cyberspace Safety and Security (CSS 2013). LNCS, 2013, vol. 8300, pp. 183–197.

8. Valdes A., Skinner K. An Approach to Sensor Correlation. Proc. of the Recent Advances in Intrusion Detection (RAID-2000), Touluse, 2000. URL: https://www.researchgate.net/profile/Alfonso_Valdes/publication/228523518_An_approach_to_sensor_correlation/links/56d4437d08ae868628b24ba8.pdf (дата обращения: 27.10.2018).

9. Kholod I.I., Bekeneva Ya.A., Novikova E.S., Shorov A.V. Intellectual model for violations detection in the business process. In Young Researchers in Electrical and Electronic Engineering (EIConRus), Proc. IEEE Conf. of Russ., 2018, pp. 313–317. DOI: 10.1109/eiconrus.2018.8317095.

10. Bekeneva Ya.A., Lebedev S.I., Kholod I.I., Shorov A.V., Novikova E.S. Method for transformation of data from heterogeneous monitoring devices for violations detection. Proc. XXI Intern. Conf. (SCM), 2018, pp. 753–756.

11. RapidMiner: Data Science Platform. URL: https://rapidminer.com/ (дата обращения: 27.10.2018).

Software & Systems
DOI: 10.15827/0236-235X.126.197-206

Received 09.11.18
2019, vol. 32, no. 2, pp. 197–206

Transformation of data from heterogeneous monitoring systems

Ya.A. Bekeneva [1], *Postgraduate Student, yana.barc@mail.ru*

[1] *St. Petersburg Electrotechnical University "LETI", St. Petersburg, 197376, Russian Federation*

Abstract. The paper presents an approach to the preparation of data obtained from heterogeneous monitoring systems for their further analysis by data mining methods. The main problem of data analysis in monitoring various processes is the difference in the description of events for different types of sources, including a data presentation format. In addition, one event might be described using data from different monitoring systems.

The paper presents a formal model of the analyzed process, describes the main problems of analyzing heterogeneous data, and highlights formal criteria for assigning records from different sources to a single event. In the proposed approach, a source of data is not only real-time records from various monitoring systems, but also account databases used for storing information. The main idea is that moving objects of different types can perform actions as a unit within the framework of the task being studied (for example, a vehicle and a driver). Account systems allow finding relationships between such moving objects and thereby increase the accuracy of combining records related to one event.

The proposed approach has been tested on real data obtained from an enterprise. After applying all described transformations, it has become possible to significantly reduce the excess dimension of an aggregate data table, as well as significantly reduce the number of missing values. When data analysis was difficult due to their different formats, such data were brought to a single format and presented in the form of a single table

that is convenient for further research using data mining methods.

Keywords: heterogeneous sources, monitoring systems, data transformations, attributes, events, data format.

Acknowledgements. *The research has been financially supported by the Ministry of Education and Science of the Russian Federation within the framework of the state task "Organizing scientific research", task no. 2.6113.2017/6.7.*

References

1. Dumas M., La Rosa M., Mendling J., Reijers H. *Fundamentals of Business Process Management*. Heidelberg, Springer Publ., 2013, vol. 1, p. 2.
2. Bastani K., Rao P. K., Kong Z. An online sparse estimation-based classification approach for real-time monitoring in advanced manufacturing processes from heterogeneous sensor data. *IIE Trans.* 2016, vol. 48, no. 7, pp. 579–598. DOI: 10.1080/0740817X.2015.1122254.
3. Kuznetsov D.P., Rzheutskaya S.Yu. Method of integration of ontologies of heterogeneous data sources in automated control systems. *Bulletin of Cherepovets State Univ*. 2013, vol. 1, no. 4, p. 51 (in Russ.).
4. Sapunov N.O. Integration of heterogeneous data sources through xml Web services when organizing the management of the transport process. *Bulletin of the State Univ. of Maritime and River Fleet n.a. Admiral S.O. Makarov*. 2010, no. 4, p. 8 (in Russ.).
5. Starozhilets V.M., Chekhovich Yu.V. Integrating data from disparate sources in traffic flow modeling problems. *Machine Learning and Data Analysis*. 2016, vol. 2, no. 3, pp. 260–275 (in Russ.). DOI: 10.21469/22233792.2.3.01.
6. Dong C., Xiuquan Q., Gelernter J., Li Xiaofeng, Meng Luoming. Mining data correlation from multi-faceted sensor data in the Internet of Things. *China Communications*. 2011, vol. 8, no. 1, pp. 132–138.
7. Mirheidari S.A., Arshad S., Jalili R. Alert correlation algorithms: a survey and taxonomy. *Proc. 5th Intern. Symp. on Cyberspace Safety and Security (CSS 2013). LNCS*. 2013, vol. 8300, pp. 183–197.
8. Valdes A., Skinner K. *An Approach to Sensor Correlation. Proc. of the Recent Advances in Intrusion Detection (RAID-2000)*. Toulouse, 2000. URL: https://www.researchgate.net/profile/Alfonso_Valdes/publication/228523518_An_approach_to_sensor_correlation/links/56d4437d08ae868628b24ba8.pdf (accessed October 27, 2018).
9. Kholod I.I., Bekeneva Ya.A., Novikova E.S., Shorov A.V. Intellectual model for violations detection in the business process. *2018 IEEE Conf. of Russ. Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pp. 313–317 (in Russ.). DOI: 10.1109/eiconrus.2018.8317095.
10. Bekeneva Ya.A., Lebedev S.I., Kholod I.I., Shorov A.V., Novikova E.S. Method for transformation of data from heterogeneous monitoring devices for violations detection. *Proc. 2018 21st Intern. Conf. on Soft Computing and Measurements (SCM)*. St. Petersburg, 2018, pp. 753–756 (in Russ.).
11. *RapidMiner: Data Science Platform*. URL: https://rapidminer.com/ (accessed October 27, 2018).

Для цитирования

Бекенева Я.А. Преобразование данных от разнородных систем мониторинга // Программные продукты и системы. 2019. Т. 32. № 2. С. 197–206. DOI: 10.15827/0236-235X.126.197-206.

For citation

Bekeneva Ya.A. Transformation of data from heterogeneous monitoring systems. *Software & Systems*. 2019, vol. 32, no. 2, pp. 197–206 (in Russ.). DOI: 10.15827/0236-235X.126.197-206.