Применение вероятностных моделей для уточнения процесса восстановления и оценки качества моделей бизнес-процессов в условиях ограниченного размер журнала событий

Ивлев К.Г, Затевалов А.С.

Аннотация **— целью данной работы является исследование возможностей построения и уточнение оценок качества моделей бизнес-процессов в условиях неполного исходного журнала событий информационной системы за счет применения вероятностных моделей для искусственной генерации журналов событий произвольного объема. Неполнота журнала событий заключается в том, что в нем не представлены все возможные ветви выполнения бизнес-процесса. Данный факт приводит к получению модели процесса низкого качества и осложняет последующий анализ модели экспертом соответствующей предметной области с целью принятия точных решений. В данной работе предложен способ, который позволяет дополнить журнал событий информационной системы экземплярами бизнес-процесса с недостающими ветвями выполнения из оригинальной модели процесса. Способ заключается в первоначальном обучении вероятностной модели (в качестве примера взята цепь Маркова), и использовании обученной модели для искусственной генерации журнала. Также в данной работе предложен способ, который позволяет оценить данный подход на основе общепринятых метрик, и доказывает его применимость в решении прикладных задач, связанных с анализом моделей, восстановленных из журналов информационных систем.**

Ключевые слова**—Process Mining, вероятностная модель, интеллектуальный анализ процессов.**

I. Введение

Process Mining или интеллектуальный анализ процессов – это подход из дисциплины процессного управления, который позволяет анализировать бизнес-процессы, основываясь на журналах событий информационных систем. Основные задачи, решаемые при помощи интеллектуального анализа процессов – это извлечение (или восстановление) модели бизнес-процесса из журнала информационной системы, проверка соответствия модели бизнес-процесса журналу и дополнение модели знаниями из журнала. Каждая из перечисленных задач для успешного решения требует определенного понимания предметной области, процессно-ориентированного подхода и структурированного журнала событий достаточно хорошего качества. Перед тем, как начинать работать с событийным журналом информационной системы его необходимо привести к специально оговоренной структуре.

Интеллектуальный анализ процессов является достаточно молодой и очень перспективной областью знаний, так как позволяет устранить разрыв между реальными событийными данными бизнес-процесса и управлением операционными процессами, тем самым провоцируя постоянное улучшение и оптимизацию процессной модели бизнес-подразделения. Также интеллектуальный анализ процессов позволяет аналитикам получать более актуальную информацию о фактическом функционировании процессов для принятия более точных и обоснованных тактических решений. [1]

При решении вышеперечисленных основных задач приходится сталкиваться с такими проблемами, как:

- «шумы» в журнале событий: алгоритмы восстановления, плохо реагирующие на «шум» могут выдавать модели непригодные для дальнейшего анализа. Может получиться так, что модель невозможно будет «проиграть», то есть заложить в информационную систему с целью автоматизации процесса, так как модель имеет некорректный вид. Также затруднения могут присутствовать при проверке соответствия моделей журналу. Даже если модель абсолютно точно соответствует журналу событий, то наличие «шума» в журнале приведет к отклонению показателей;
- наличие сложных структур в журнале событий. Одним из строгих способов представить модель процесса в информационной системе является сеть Петри (математический аппарат для моделирования динамических дискретных систем). У данного аппарата имеются ограничения по сложности моделируемых структур. Если подобные структуры содержатся

Статья получена 17 декабря 2013 г. 2013. Работа является результатом магистерской диссертации «Применение вероятностных моделей для уточнения процесса восстановления моделей бизнес-процессов в условиях ограниченного размера журнала событий».

Ивлев К. Г., Московский Государственный Университет

Затевалов А.С., Московский Государственный Университет

в журнале событий, то алгоритмы по работе с моделями бизнес-процессов усложняются. Неадаптированные алгоритмы приводят к получению моделей недостаточного качества;

- сложность восстанавливаемой модели. Слишком сложная модель не представляет ценности для аналитика, так как человеку невозможно извлечь пользу из большой диаграммы с большим количеством переходов, циклов и повторяющихся типов событий;
- большой объем данных. На большом объеме данных при промышленном, а не при учебном использовании, возникает проблема слишком долгой работы алгоритмов (например, генетические алгоритмы требуют большого количества итераций для нахождения оптимальной модели; [2]
- «неполнота» журнала событий. Под «неполнотой» понимается отсутствие в журнале событий всех ветвей выполнения бизнес-процесса, которые есть в модели, на основе которой бизнес-процесс был реализован. Данная проблема может возникнуть, например, если информационная система эксплуатируется в течение достаточно небольшого количества времени, или часть журнала была утеряна, или по каким-либо другим причинам часть бизнес-процесса не исполнялась.

Для того чтобы оценить полученную модель применяются общеизвестные метрики качества (ниже приведено неформальное описание метрик) [3]:

- «пригодность» (fitness) – насколько хорошо модель описывает журнал. Чем больше экземпляров бизнес-процесса из журнала описывается моделью, тем выше мера пригодности;
- «обобщенность» (generalization). Метрика, по смыслу являющаяся противоположностью «точности». Можно построить такую модель, в которой каждая отдельная ветка будет предназначена для отражения каждого отдельного экземпляра бизнес-процесса из журнала. Подобная модель характеризуется высокой мерой «пригодности», но низкой мерой «обобщенности», так как не будет позволять ни малейшей модификации экземпляров процесса в журнале;
- простота (simplicity). Сложность модели бизнес-процесса приводит к ее бесполезности с точки зрения последующего анализа. Существует класс алгоритмов восстановления процессов, которые выдают хорошее значение метрики «простота», но за счет снижения показателей «пригодности» и «точности»;
- точность (precision). Очень легко построить универсальную простую модель, которая может воспроизвести любую конечную последовательность событий указанных типов. У такой модели будет высокая «пригодность», но низкая «точность», так как очень много вариантов выполнения процесса, которые допускаются моделью, но не присутствуют в журнале.

Не имеет смысла анализировать вышеперечисленные метрики качества модели бизнес-процесса в отдельности. Из комментариев к каждой метрике видно, что значительное улучшение одной метрики может привести к значительному ухудшению другой. Поэтому стоит принимать во внимание только средневзвешенную сумму метрик при минимально заданном пороге каждой отдельно взятой метрики.

В силу того, что интеллектуальный анализ процессов появился несколько лет назад, то существует еще много открытых вопросов и нерешенных проблем, которые ждут своего часа для решения практических задач анализа бизнес-процессов.

В рамках работы будет рассмотрена проблема «неполноты» журнала событий. Данная проблема является важной, так как:

- при отсутствии ее решения полученная модель является неточной;
- в этой модели не будут представлены все те ветви выполнения, которые есть в исходной модели процесса. Под исходной моделью процесса следует понимать модель, которая была реализована в информационной системе. А журнал, с которым ведется работа, пишется развернутой информационной системой;
- полученную недостоверную модель сложно анализировать;
- если журнал используется для оценки качества уже существующей (или построенной вручную) модели, то характеристики могут быть неточными и значительно отклоняться от реального положения дел.

II. Начальные сведения

В данном разделе будут приведены определения и алгоритмы, применяемые в дальнейшем исследовании.

A. Сеть Петри

Сеть Петри - аппарат для моделирования динамических дискретных систем. Сеть Петри определяется как четверка <P, T, I, O>, где P и T - конечные множества позиций (кружочки, см. рис. 4) и переходов (квадраты), I (дуги, направленные от позиций к переходам) и O (от переходов к позициям) множества входных и выходных функций. Сеть Петри - двудольный ориентированный граф. Пересечение P и T пусто.[4]

Каждая позиция может содержать в себе маркер. Распределение маркеров по событиям называется маркировкой. Маркеры могут перемещаться в сети. Каждое изменение маркировки называется событием, причем каждое событие связано с определенным переходом.

Каждому условию в сети Петри соответствует определенная позиция. Совершению события соответствует срабатывание перехода, при котором маркеры из входных позиций этого перехода перемещаются в выходные позиции. Последовательность событий образует моделируемый

процесс. Переход срабатывает, если для каждой из его входных событий выполняется условие, при котором число маркеров в событии больше либо равно числу дуг, ведущих из события в переход. После срабатывание перехода в каждом выходном событии появляется столько маркеров, сколько дуг ведет из перехода в данное событие.

Таким образом, простое представление системы сетью Петри основано на двух понятиях: событиях и условиях. События это действия, имеющие место в системе. Возникновением событий управляет состояние системы. Условие может принимать либо значение “истина”, либо значение “ложь“. Так как события являются действиями, то они могут происходить. Для того чтобы событие произошло, необходимо выполнение соответствующих условий. Эти условия называются предусловиями события. Возникновение события может вызвать нарушение предусловий и может привести к выполнению других условий - постусловий.

Важные свойства сети Петри:

- асинхронная природа (поддерживается параллельное выполнение событий).
- выполнение сети Петри это последовательность дискретных событий. Порядок появления событий является одним из возможных, допускаемых основной структурой. Данный факт приводит к недетерминированности.
- при помощи сети Петри можно смоделировать конечный автомат.

B. Алгоритм восстановления сети Петри на основе языковой теории регионов

Существует группа алгоритмов, при помощи которых можно восстановить сеть Петри из журнала событий, под названием “теория регионов“. Группа алгоритмов основывается на поведении, которое могло бы быть у существующей сети. [5] В начале из журнала выделяются состояния и переходы между состояниями. Далее строится ориентированный граф переходов из состояния в состояние. Минимальные группы состояний, обладающими определенными свойствами (все дуги у состояний из данной группы являются либо входящими, либо выходящими для каждого конкретного состояния) объединяются в регион. Регион становится местом в сети Петри, которая будет впоследствии построена. В качестве расширения алгоритма был предложен следующий подход: считать регионом только то множество переходов, которое не ограничивало бы описываемые процессы, если в регион добавить еще одно состояние, полученное специальным образом. Входящими дугами в состояние будут все входящие в регион дуги, выходящими дугами из состояния будет все выходящие из региона дуги. Минус подхода к построению сети Петри на основе регионов – «взрывное количество состояний» («state explosion problem»). Логичным следствием является попытка разбить состояния на некоторые приблизительно независимые группы (abstract domain), и работать с каждой группой по отдельности.[6]

Указанный в названии раздела алгоритм строится на теории регионов, описанной выше. Задача теории регионов – определить количество и места позиций в искомой сети Петри процесса. Языковая теория таким же способом ищет позиции, строя регионы, но на вход получает не ориентированный граф переходов из состояния в состояние, а некоторый «язык». Этот язык строится из алфавита, где в качестве символов фигурируют события из журнала, а в качестве правил используются зависимости между событиями журнала.

Рассмотрим некоторый журнал информационной системы, в котором представлен набор событий. Для данного журнала можно сконструировать сеть Петри N_0, в которой будут только переходы без позиций. Такая сеть Петри полностью может воспроизвести любой из экземпляров бизнес-процесса, приведенных в журнале.

Идея алгоритма заключается в том, что добавляя дополнительные позиции в такую сеть Петри, можно только ограничить поведение, т.е. уменьшить потенциальное количество возможных экземпляров бизнес-процессов, которые данная модель может воспроизвести. Задачей алгоритма является добавление позиций в сеть Петри таким образом, чтобы получившаяся сеть Петри могла воспроизвести все экземпляры бизнес-процесса из журнала. В базовой реализации алгоритма извлечения на основе языковой теории регионов итоговый набор регионов (на основе которых создаются позиции) подбирается как пространство положительных целочисленных решений системы линейных неравенств. Неравенства для системы формируются на основе такого утверждения, что разность количества маркеров, находящихся на некоторой возможной позиции искомой сети Петри перед наступлением k-ого события рассматриваемого экземпляра процесса, и количества маркеров, которые необходимо затратить для наступления k-ого события должно быть неотрицательным. Если записать такие неравенства для каждого события каждого экземпляра процесса в журнале, и решить получившуюся систему, то можно получить набор позиций, которые можно добавить в искомую сеть Петри. Полученная сеть Петри по определению сможет воспроизводить журнал, потому что проверено, что каждая добавляемая позиция не нарушает последовательность выполнения экземпляров рассматриваемого процесса.

Модель, полученная в результате применения данного подхода к восстановлению:

- по определению имеет наивысшее значение метрики «пригодности» («fitness») – 100%;
- метрика «точность» может значительно пострадать, если на вход процесса восстановления подается неполный относительно наличия всех возможных ветвей выполнения процесса журнал информационной системы;

По причине недостающей точности, способ извлечения модели на основе теории регионов был выбран в качестве алгоритма для проведения экспериментов с целью исследования возможности построения и уточнения оценок качества моделей бизнес-процессов в условиях ограниченного исходного

журнала событий

III. Исследование и построение задачи

A. *Общее описание эксперимента*

В рамках данной работы для решения проблемы с неполнотой исходного журнала событий информационной системы было решено применить вероятностную модель. Задача вероятностной модели – дополнить журнал событий недостающим поведением для получения скорректированной модели процесса. Таким образом, вначале построим вероятностную модель на основе имеющегося журнала событий. Далее используем вероятностную модель для искусственной генерации новых экземпляров процесса. Так как вероятностная модель имеет гибкость, то на выходе можно получить экземпляры процесса, которые не были представлены в исходном журнале бизнес-процесса.

Факт появления дополнительного поведения в журнале информационной системе можно использовать для решения двух задач: улучшения модели и проверки качества модели. Ниже представлено описание двух подходов для решения указанных задач.

1) Получение модели

Используя неполный журнал событий можно получить модель бизнес-процесса довольно низкого качества, что затруднит работу аналитика. Применим сгенерированный журнал для восстановления другой модели и покажем, что качество такой модели будет выше, чем исходной.

2) Проверка качества модели.

Процедуру по проверке качества восстановленной модели можно проводить, если аналитик:

- используя некоторый журнал событий, провел операцию по восстановлению модели бизнес-процесса
- не уверен в качестве журнала событий с точки зрения его полноты (т.е. наличия все возможных экземпляров бизнес-процесса)
- хочет получить объективную оценку восстановленной модели

Процедура заключается в получении значений характеристик восстановленной модели на основе сгенерированного журнала событий (а не исходного, на основе которого модель была построена).

B. *Доказательство корректности эксперимента*

В реальной ситуации предложенные подходы невозможно проверить на корректность без привлечения эксперта в предметной области. Поэтому было принято решение поставить искусственный эксперимент, который позволит убедиться в применимости предложенного подхода. С этой целью нужно взять некий изначально полный журнал бизнес-процесса информационной системы, и сымитировать реальную ситуацию, искусственно сделав его неполным. То есть убрать из него часть экземпляров бизнес-процесса. Далее построить вероятностную модель (или модели, если экземпляры бизнес-процесса разные и имеют сложную структуру). Сгенерировать пропорционально количеству соответствующих вероятностным моделям экземпляров процесса новые экземпляры процесса.

Для решения первой задачи построим две модели процесса: по искусственно полученному неполному журналу - M_1, и по сгенерированному журналу событий - M_2. Далее, чтобы проверить корректность первого подхода, проведем проверку соответствия этих моделей исходному полному журналу событий. Если совокупные характеристики модели M_2 будут лучше, чем у модели M_1, то этот факт позволит удостовериться в применимости данного подхода на практике, так как в результате работы можно получить модель процесса, которая будет превосходить по характеристикам модель процесса, полученную без применения данного подхода.

Для решения второй задачи построим модель бизнес-процесса на основе искусственного полученного неполного журнала событий. Это та модель, которую необходимо точно оценить. Далее проведем две процедуры проверки соответствия этой модели исходному полному журналу событий и сгенерированному журналу событий. Если полученные значения характеристик будут схожи, то этот факт доказывает корректность подхода и возможность применения его на практике.

C. *Инструментарий*

ProM (или ProM-framework) – это платформа с открытым программным кодом, предоставляющая средства для интеллектуального анализа процессов. Платформа предоставляет площадку для развертывания подключаемых модулей, в которых реализованы алгоритмы интеллектуального анализа процессов. Последняя версия платформы 6.3 поддерживает основной набор подключаемых модулей, покрывающих все основные задачи интеллектуального анализа процессов. Представлены модули для извлечения модели, проверки соответствия и улучшения модели на основе журналов информационных систем. Ниже перечислены модули, которые будут использоваться при постановке экспериментов в рамках данной работы.

1) Модуль «Sequence clustering»

На основе имеющегося журнала событий необходимо обучить вероятностную модель, которая будет инкапсулировать в себе поведение, отраженное в журнале. Основными требованиями, предъявляемыми к вероятностной модели, являются:

- простота, которая не позволит получить «переобученную» вероятностную модель. Переобученная модель будет бесполезна, так как с ее использованием можно будет получить сходный журнал без внесения в него дополнительного поведения;
- сохраняет базовые отношения порядка между элементами пространства событий, что позволит избежать добавления излишнего «шума» в искомую модель процесса (т.е. не будет состояния «недообученности»). Добавления «шума» может только ухудшить показатели метрик модели процесса, и добавит дополнительную работу по фильтрации заведомо некорректных экземпляров процесса в журнале.

В качестве вероятностной модели, удовлетворяющей всем перечисленным характеристикам, была выбрана цепь Маркова первого порядка. С одной стороны такая модель довольно проста по способу построения, с другой – сохраняет базовые отношения порядка между событиями по способу своего обучения.

Подключаемый модуль «Sequence Clustering» используется для кластеризации представленных в журнале информационной системы экземпляров процесса. В качестве алгоритма и информационной сущности, представляющей кластеры, используется цепь Маркова первого порядка.

Цепь Маркова первого порядка – это случайный процесс с дискретными состояниями и дискретным временем. Для данного случайного процесса выполняется единственное правило: для любого момента времени условная вероятность каждого состояния системы в будущем зависит только от состояния системы в настоящем и не зависит от того, когда и как система пришла в это состояние. В нашем случае в качестве дискретных состояний берутся события из журнала информационной системы, а в качестве дискретного времени – очередность наступления событий. Важна только очередность наступления событий. [7]

«Sequence Clustering» работает следующим образом. На вход модулю подается журнал информационной системы и желаемое число кластеров *K*, которое аналитик желает получить в результате работы алгоритма. «Sequence Clustering» работает по принципу «Expectation-Maximization» (последовательного приближения к цели):

- случайным образом инициализируем *K*-цепей Маркова первого порядка (случайным образом выставляются вероятности переходов из состояния в состояние);
- для каждого экземпляра процесса из журнала событий высчитывается вероятность принадлежности экземпляра каждому кластеру. Кластер, к которому экземпляр процесса относится с наибольшей вероятностью, будет являться кластером, к которому экземпляр принадлежит;
- на основе экземпляров, которые отнесены к кластерам на предыдущем шаге, строим цепь Маркова для каждого кластера;
- повторяем два предыдущих шага до тех пор, пока принадлежность любого взятого экземпляра из журнала информационной системы к кластерам не будет меняться. Т.е. экземпляры процесса перестанут «переходить» из кластера в кластер.

«Sequence Clustering» часто применяется в тех случаях, когда модель, извлекаемая из журнала информационной системы, слишком сложна для анализа. Аналитик, разбив журнал на поджурналы, в результате получит несколько более простых моделей, поддающихся анализу, и, соответственно, представляющих большую ценность. По факту может получиться так, что в разных кластерах будут содержаться экземпляры разных бизнес-процессов, представленных в одном и том же журнале.

На выход подключаемого модуля «Sequence Clustering» выдается несколько цепей Маркова в специальном формате. Данный формат избыточен, так как предназначен, в том числе и для визуализации полученной цепи Маркова в ProM.

2) Модуль «ILP Miner»

Модуль является реализацией алгоритма, основанного на языковой теории регионов. Модуль «ILP miner» был выбран в качестве инструмента для восстановления модели бизнес-процесса, потому что проблема неполноты журнала событий является наиболее актуальной при его использовании. В результате использования языковой теории регионов по способу построения алгоритма получается наивысшее значение метрики «пригодность», но из-за «недообученности» восстановленная модель процесса допускает дополнительное поведение (т.е. экземпляры бизнес-процесса), которых заведомо нет в исходном журнале. Дополнительное поведение снижает оценку метрики «точность» («precision»). Предполагается, что, в рамках первой задачи, применение подхода по генерации искусственного журнала позволит без снижения (или с незначительным снижением) значения «пригодности» получить повышение значения «точности».

3) Модули проверки соответствия

Для проверки соответствия модели журналу событий используются следующие составные части ProM:

Модуль «Replay a Log on Petri Net for Conformance Analysis Plug-in» служит для построения «выравниваний» между каждым экземпляром бизнес-процесса в журнале событий и моделью процесса, представленной в виде сети Петри. «Выравнивание» показывает, насколько хорошо журнал может быть «проигран» на основе модели процесса, и насколько много несоответствий между журналом и моделью. В качестве результата данный модуль выдает значение меры «пригодность».

Модуль «Measure Precision/Generalization» конвертирует модель из представления сети Петри в конечный автомат и позволяет получить значения метрик «точности» и «обобщенности» для модели бизнес-процесса.

IV. Практическая часть

A. Описание исходных данных

Для того чтобы решать любую задачу, связанную с интеллектуальным анализом процессов, необходимо ознакомиться с представленной в журнале информацией. На сайте processmining.org находится специальный эталонный журнал событий реальной информационной системы repair_example.mxml, который подходит для первичной проверки качества алгоритмов извлечения модели из журнала. Журнал является достаточно простым, т.к. в нем содержится небольшое абсолютное число типов событий.

B. Формат входных данных.

Журнал repair_examlple.mxml представлен в специальном *.xml формате, который позволяет удобно

работать с данными. Корневым элементом журнала является элемент «WorkflowLog», который указывает на то, что данный файл представляет собой журнал событий в формате *.mxml. Корневой элемент содержит два дочерних. Первый – «Source» имеет атрибут «program», в котором отражено имя информационной системы или подсистемы, в которой журнал был создан. Второй – «Process» представляет собой тип бизнес-процесса, журнал исполнения которого представлен внутри. Внутри элемента «Process» содержатся элементы «ProcessInstance» – конкретные экземпляры бизнес-процесса. Каждый экземпляр имеет атрибут id в качестве уникального числового порядкового идентификатора экземпляра данного процесса. События бизнес-процесса представлены элементом «AuditTrailEntry». Среди вложенных элементов следует отметить следующие:

- «WorkflowModelElement»: значение является типом события (например, Register или Analyze Defect)
- «TimeStamp»: временная метка (когда данное событие произошло)
- «Originator»: кем событие порождено (например, System, tester3)

Загрузив в ProM файл журнала можно увидеть базовую информацию о событиях и экземплярах бизнес-процесса. В рассматриваемом журнале событий находится 11855 событий в 1104 цепочках событий.

C. Фильтрация исходного журнала

Анализ журнала, выполненный при помощи ProM, показал, что не все экземпляры данного бизнес-процесса, представленного в журнале, содержат формальные признаки завершения (отклонение выявлено на основе того, что не все экземпляры заканчиваются одинаково). В экземпляре процесса под номером 1001 последним событием является «тестирование: завершение», а должно быть событие «архивирование».

Для корректности поставленного эксперимента необходимо провести процедуру очистки исходного «эталонного» журнала событий с помощью подключаемого модуля «Simple Heuristics (Filter Log using Simple Heuristics)». В качестве результата будет получен журнал, в котором отсутствует «шум», связанный с незавершенностью бизнес-процесса. Этот факт позволит поставить «чистый» эксперимент, проанализировав результаты восстановления и проверки соответствия в рамках описываемого в данной работе подхода для решения проблемы «неполноты» журнала событий без влияния «шума» в нем.

D. Кластеризация.

Далее проведем, запланированную ранее равномерную выборку экземпляров процесса из отфильтрованного «эталонного» журнала событий в пропорции 5%, 14%, 20%, 25%, 33%, 50%. Используя модуль «Sequence Clustering» выполним процедуру кластеризации журнала событий. На вход модулю подадим искусственно созданный «неполный журнал», и укажем в качестве желаемого количества кластеров – 10. По результатам работы алгоритма было сформировано 10 цепей Маркова.

E. Генерация журнала

Полученные на шаге кластеризации цепи Маркова сохраняем в файлы для выполнения шага по искусственной генерации нового журнала. Файлы именуются следующим образом: *markov_xx_y_zzz.dot*, где *xx* – общее число сгенерированных кластеров, *y* – порядковый номер кластера, *zzz* – число экземпляров бизнес-процесса, которые отнесены к кластеру *y*. Для искусственной генерации журнала был создан программный продукт, которое использует все полученные цепи Маркова и генерирует пропорционально количеству принадлежащих каждому кластеру экземпляров процесса новый журнал событий. Общее число экземпляров нового журнала событий равно числу экземпляров бизнес-процесса «эталонного» журнала, в данном случае 1000. Результат работы приложения выводится в консоль и в файл и имеет csv-формат следующей структуры: «порядковый номер экземпляра бизнес-процесса; наименование события;».

Чтобы использовать искусственный журнал бизнес-процесса в таком виде, нужно его преобразовать в XES-формат. Такая процедура выполняется при помощи модуля «Convert Key/Value Set To Log».

F. Восстановление моделей

Далее, как и было описано в главе «Исследование и построение задачи» при помощи подключаемого модуля «ILP-miner» восстанавливаем две модели:

- модель на основе искусственного журнала бизнес-процесса
- модель на основе «неполного» (искусственно сымитированного) журнала событий для проверки корректности подхода.
- Такую пару моделей строим для каждой выборки из исходного «эталонного» журнала.

G. Подсчет значений основных метрик качества бизнес-процесса

Далее, каждую из полученных моделей проверим на соответствие исходному полному «эталонному» журналу событий. Проверка будет производиться в два этапа:

- для получения значения метрики «пригодность» (fitness) необходимо использовать модуль «Replay a Log On Petri Net For Conformance Analysis». На вход данного модулю подается модель бизнес-процесса и журнал событий, на основе которого будет проведена проверка бизнес-процесса.
- результат использования модуля «Replay a Log On Petri Net For Conformance Analysis» необходимо подать на вход модулю «Measure Precision/Generalization» для подсчета метрик «обобщенности» (generalization) и «точности» (precision).

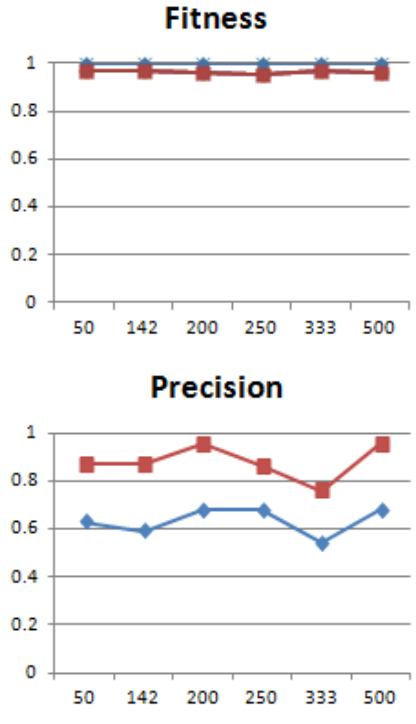
Результаты подсчета метрик были собраны в таблицу (см. табл. 1) для анализа.

Таблица 1. Результаты измерений основных метрик «пригодность», «точность» и «обобщенность»

моделей, построенных по «неполному» журналу и искусственно сгенерированному журналу бизнес-процесса.

"Пригодность" (fitness)		
Размер выборки	**С.Петри (выборка из исходного журнала)**	**С.Петри (искусственный журнал)**
5%	0.999	0.967
15%	0.999	0.971
20%	0.999	0.961
25%	0.999	0.957
33%	1	0.967
50%	0.999	0.961
"Точность" (precision)		
Размер выборки	**С.Петри (выборка из исходного журнала)**	**С.Петри (искусственный журнал)**
5%	0.629	0.871
15%	0.592	0.87
20%	0.678	0.953
25%	0.68	0.86
33%	0.541	0.759
50%	0.68	0.953
"Обобщенность" (generalization)		
Размер выборки	**С.Петри (выборка из исходного журнала)**	**С.Петри (искусственный журнал)**
5%	0.999	0.999
15%	0.996	0.999
20%	0.988	0.999
25%	0.994	0.999
33%	0.998	0.999
50%	0.994	0.999

Для удобства анализа по таблице были построены графики (см. рис. 1).



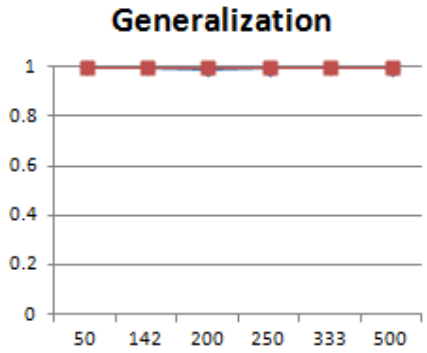


Рис. 1. Графическое представление результатов измерений основных метрик моделей, построенных по «неполному» журналу и искусственно сгенерированному журналу бизнес-процесса.

На рис. 1 красной линией представлены значения основных метрик модели, восстановленной на основе искусственно сгенерированного журнала событий, синей – модели, восстановленной на основе «неполного» журнала событий (по горизонтали указывается размер выборки из исходного «полного» журнала).

Таким образом, цель эксперимента по первой задаче достигнута, модель построенная на основе сгенерированного журнала событий имеет значительно большую точность, чем модель, построенная по «неполному» журналу.

В рамках проведения экспериментального исследования по второй задаче, заключающейся в уточнении оценок качества модели бизнес-процесса в условиях неполного исходного журнала событий, использованы:

- исходный «эталонный» журнал событий;
- модели (сети Петри), на основе «неполного» (сымитированного) журнала событий,
- искусственный журнал событий.

Целью данного эксперимента является исследование возможной схожести метрик качества модели бизнес-процесса, полученных при проверке соответствия «эталонного» и искусственного журнала событий с моделью бизнес-процесса.

Для построения сетей Петри был использован модуль ProM «ILP Miner». Для построения искусственного журнала событий необходимо было задействовать модуль «Sequence Clustering» для построения цепей Маркова и сохранения их в файлы. С помощью созданного программного продукта данные файлы были интерпретированы в матрицы вероятностей переходов и сгенерирован искусственный журнал событий, по объему соответствующий исходному «эталонному». Для проверки соответствия исходного «эталонного» и искусственного журнала событий по модели бизнес-процесса, построенной по неполному журналу, были последовательно использованы два модуля платформы ProM: «Replay a Log On Petri Net For Conformance Analysis» и «Measure Precision/Generalization».

Результаты подсчета метрик для анализа были собраны в таблицу (см. табл. 2).

Таблица 2. Результаты проверки соответствия

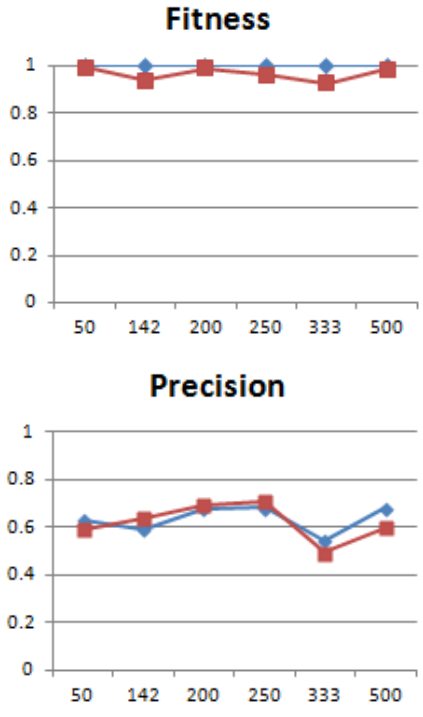
моделей, построенных по «неполному» журналу, и исходного «эталонного» и искусственного журнала событий бизнес-процесса.

"Пригодность" (fitness)		
Размер выборки	**Исходный «эталонный» журнал**	**Искусственный журнал**
5%	0.999	0.992
14%	0.999	0.966
20%	0.999	0.984
25%	0.999	0.980
33%	1.000	0.958
50%	0.999	0.978

"Точность" (precision)		
Размер выборки	**Исходный «эталонный» журнал**	**Искусственный журнал**
5%	0.629	0.617
14%	0.592	0.585
20%	0.678	0.677
25%	0.680	0.655
33%	0.541	0.508
50%	0.680	0.709

"Обобщенность" (generalization)		
Размер выборки	**Исходный «эталонный» журнал**	**Искусственный журнал**
5%	0.999	0.993
14%	0.996	0.997
20%	0.988	0.998
25%	0.994	0.998
33%	0.998	0.985
50%	0.994	0.998

Для удобства анализа по таблице были построены графики (см. рис. 2).



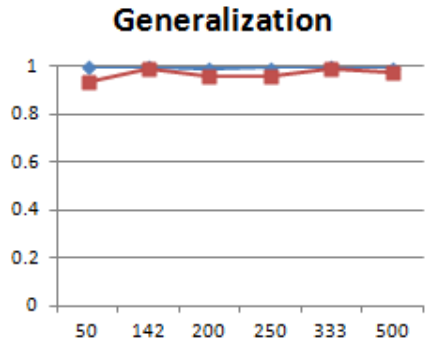


Рис. 2. Графическое представление результатов измерений основных метрик проверки соответствия моделей, построенных по «неполному» журналу, и исходного «эталонного» и искусственного журнала событий бизнес-процесса..

На рис. 2 синей линией представлены значения основных метрик проверки соответствия исходного «эталонного» журнала событий и модели, построенной по «неполному» журналу событий, а красной линией – представлены значения метрик проверки соответствия искусственного журнала событий и модели, построенной по «неполному» журналу событий. Диапазон отклонений по метрике fitness составляет [0.4,4.2]%, по метрике precision [0.1,3.3]%, по метрике generalization [0.1,1.3]%.

Таким образом, цели эксперимента по второй задаче достигнуты. Метрики искусственного и «эталонного» журнала схожи и почти идентичны, что позволяет в реальной ситуации, в отсутствии «эталонного» журнала событий, объективно и адекватно оценивать восстановленную модель бизнес-процесса.

V. Заключение

Подведем итоги:

- разработан метод искусственной генерации событий бизнес-процессов с использованием вероятностной модели (цепь Маркова первого порядка), обученных на журналах событий ограниченного размера;
- была разработана методика построения моделей бизнес-процессов, использующих искусственно сгенерированные журналы событий;
- была разработана методика оценки качества моделей бизнес-процессов, построенных в условиях ограниченного объема журналов событий, основанная на искусственной генерации журналов событий;
- было проведено экспериментальное исследование на примере прикладной задачи, показавшее возможную применимость предложенной методики для улучшения показателя “точность” восстановленной модели бизнес-процесса;
- было проведено экспериментальное исследование на примере прикладной задачи по оценке модели бизнес-процесса, показавшее схожесть и почти идентичность проверки

соответствия искусственных и «эталонных» журналов событий по модели бизнес-процесса, построенной по «неполному» журналу.

Библиография

[1] Wen, L., Wang, J., van der Aalst, W.M.P., Wang, Z., Sun, J.: A Novel *Approach for Process Mining Based on Event Types. BETA Working Paper Series*, WP 118, Eindhoven University of Technology, Eindhoven (2004)

[2] A.K. Alves de Medeiros, A.J.M.M. Weijters, and W.M.P. van der Aalst *Genetic Process Mining: A Basic Approach and Its Challenges*, 2006

[3] J.C.A.M. Buijs, B.F. van Dongen, and W.M.P. van der Aalst *On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery*

[4] Питерсон Дж. *Теория сетей Петри и моделирование систем*, 1984.

[5] Robin Bergenthum, Jörg Desel, Robert Lorenz and Sebastian Mauser *Process Mining Based on Regions of Languages*.

[6] J. Carmona and J. Cortadella *Process Mining Meets Abstract Interpretation*, 2010.

[7] Gabriel M. Veiga and Diogo R. Ferreira *Understanding Spaghetti Models with Sequence Clustering for ProM*

Application of probabilistic models for refinement of the business process discovery and quality assessment with limited size of event log

Ivlev K.G, Zatevalov A.S.

***Abstract* — the purpose of this paper is research application of probabilistic models for refinement of the business process discovery and quality assessment of business process model with limited size of event log. Log incompleteness leads to low quality of process discovery, because there are no some branches of business process presented in log. In this paper the method of addition missing instances of business process instances is proposed. The method is training of probabilistic model (Markov chain), and then using this model for synthetic log generation.**

***Keywords*—Process Mining, probabilistic model.**