На правах рукописи

Ходырев Иван Александрович

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ДИНАМИКИ ПОКАЗАТЕЛЕЙ ДЕЯТЕЛЬНОСТИ ПРЕДПРИЯТИЯ НА ОСНОВЕ ЖУРНАЛОВ СОБЫТИЙ ИНФОРМАЦИОННЫХ СИСТЕМ

Специальность 05.13.18 – Математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2014

Работа выполнена в Санкт-Петербургском национальном исследовательском университете информационных технологий, механики и оптики.

Научный руководитель: Бухановский Александр Валерьевич, доктор технических наук

Официальные оппоненты: Гергель Виктор Павлович ,

доктор технических наук, профессор, декан факультета вычислительной математики и кибернетики Нижегородского государственного университета им. Н. И. Лобачевского

Холод Иван Иванович ,

кандидат технических наук, доцент Санкт-Петербургского государственного электротехнического университета им.В.И. Ульянова (Ленина)

Ведущая организация: Учреждение Российской академии наук Институт проблем передачи информации им. А.А. Харкевича РАН (ИППИ РАН)

Защита состоится 22 декабря 2014 г. в 17:00 часов на заседании диссертационного совета Д212.227.06 при Санкт-Петербургском национальном исследовательском университете информационных технологий, механики и оптики по адресу: 197101, Санкт-Петербург, Кронверкский пр., д. 49, конференц-зал, ЦИО.

С диссертацией можно ознакомиться в библиотеке Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики по адресу: 197101, Санкт-Петербург, Кронверкский пр., д.49 и на сайте fppo.ifmo.ru .

Автореферат разослан « 22 » ноября 2014 года.

Ученый секретарь диссертационного совета Лобанов И.С.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы диссертации. Современные информационные системы в различных областях деятельности человека являются источниками косвенных данных о реальных процессах, представляемых в форме так называемых журналов событий. Системы поддержки ведения бизнеса ERP, WFM, CRM позволяют проектировать течение процессов, регулировать и контролировать их исполнение, а также в автоматическом режиме накапливать объективную информацию о событиях, определяющих динамику показателей деятельности предприятия.

Важной задачей для предприятия является прогнозирование динамики показателей его деятельности. Для решения этой задачи применяются разнообразные программные средства: Rockwell Arena, Custom Simulations Sigma, ARIS и др. Однако в реальных ситуациях возможности применения таких средств часто ограничены фокусировкой на прогнозировании устойчивых состояний бизнес-процесса, а не конкретных динамических ситуаций («ориентированность на дизайн»); необходимостью ручной настройки модели, используемой для прогнозов; упрощенными моделями ресурсов. В результате полученные прогнозы часто бывают недостоверными.

Один из способов преодоления указанных проблем – использование журналов событий информационных систем предприятия как основы для построения и идентификации прогностических моделей. Первые работы по прогнозированию с использованием журналов событий появились в конце прошлого десятилетия, в рамках сравнительно новой дисциплины «process mining» извлечения знаний из журналов событий. Прогнозированием на основе журналов событий занимались А. Розинат, В. В. Д. Аалст, Р. Манс, М. Сонг, Д. Накатумба, Х. Шоненберг, Л. Инг, М. Посписил, М. Полато, Г. Вагнер, А. Сендерович и др. Однако методы, предложенные в их работах, имеют ряд ограничений: неочевиден формальный аппарат моделирования потока работ по журналам событий (используются простые инструменты, точность которых невысока), не исследованы проблемы, связанные с зависимостями элементов бизнес-процесса от контекста его исполнения, отсутствует понятие об устаревании данных. В ряде практически важных случаев пренебрежение этими факторами может существенно ухудшить качество прогнозов, что и определяет актуальность темы диссертации.

Объектом исследования являются математические модели бизнес-процессов.

Предметом исследования являются методы автоматизированного построения и использования прогностических имитационных моделей на основе журналов событий.

Целью диссертационного исследования является разработка и обоснование метода дискретно-событийного моделирования бизнес-процессов предприятий на основе журналов событий информационных систем.

Для достижения этой цели решаются следующие **задачи**:

- разработка метода, автоматизирующего процедуры построения динамической модели бизнес-процессов на основе данных, извлекаемых из журналов событий, с учетом их устаревания;
- разработка процедуры идентификации и моделирования характерных элементов динамической модели бизнес-процессов в зависимости от внешних факторов;
- разработка метода имитационного моделирования динамики бизнес-процессов в условиях изменчивости внешней среды для прогнозирования показателей деятельности предприятия;
- разработка алгоритмического и программного обеспечения для экспериментальных исследований предложенных методов и моделей;
- экспериментальное сравнение прогностических свойств предложенного метода имитационного моделирования со свойствами существующих аналогов; апробация на прикладных задачах моделирования деятельности предприятий.

Методы исследования. Для решения поставленных задач использованы методы инженерии знаний, машинного обучения, математической статистики, теории сетей Петри, имитационного моделирования.

Научная новизна исследования заключается в разработке метода дискретно-событийного моделирования бизнес-процессов, позволяющего создавать базовую имитационную модель на основе обучения по журналу событий информационной системы с учетом устаревания данных и влияния внешних факторов, а также применении метода для прогнозирования динамики показателей предприятия.

Достоверность результатов диссертационного исследования обусловлена обоснованностью применения математического аппарата, сравнением экспериментальных данных по эффективности разработанных алгоритмов с эффективностью аналогов (метод Розинат-Аалста), а также апробацией на практических примерах показателей деятельности предприятий.

На защиту выносятся:

- метод численного моделирования динамики показателей предприятия с использованием журналов событий, учитывающий состояние бизнес-процессов, наличие неактивных экземпляров процессов, влияние контекстных признаков и устаревание данных;
- процедура определения параметров и оценки качества моделей бизнес-процессов по журналам событий.

Практическую значимость работы определяют:

- методика прогнозирования показателей предприятия с использованием журналов событий на основе математического моделирования его бизнес-процессов, обеспечивающая повышение точности среднесрочного (один–шесть месяцев) прогнозирования;
- программное обеспечение прогнозирования показателей предприятия, реализующее разработанную методику.

Апробация работы. Основные положения диссертационной работы докладывались и обсуждались на XI Международной научно-технической конференции «Информационно-вычислительные технологии и их приложения» (Пенза, 2009); XXIII Международной научно-технической конференции «Математические методы и информационные технологии в экономике, социологии, образовании» (Пенза, 2009); XXIV Международной научно-технической конференции «Математические методы и информационные технологии в экономике, социологии, образовании» (Пенза, 2009); XXXVIII Международной научно-практической конференции «Неделя науки СПбГПУ» (ч. XVIII; Санкт-Петербург, 2009); 14 International Conference on Computational Science (Cairns, Australia, 2014).

Внедрение результатов работы. Результаты работы использованы в процессе развития и коммерциализации проекта «Создание высокотехнологичного производства комплексных решений в области предметно-ориентированных облачных вычислений для нужд науки, промышленности, бизнеса и социальной сферы» (в рамках Постановления №218 Правительства РФ, 2010-2013 гг.), при выполнении НИР «Создание и апробация методики семантического поиска и оценки информации, направленной на выявление перспективных направлений научных исследований в России и за рубежом на примере сферы «Автоматическое управление» по данным из открытых источников на основе потребностей реального сектора экономики и обеспечения конкурентных позиций отечественных производителей на перспективных рынках инновационных товаров и услуг и созданных научно-технических заделов», «Разработка системы мониторинга социокультурных процессов в киберпространстве», «Предсказательное моделирование экстремальных явлений и оценка рисков устойчивого развития сложных систем» по заказу Министерства образования и науки Российской Федерации.

Личный вклад автора заключается в сопоставлении и анализе основных алгоритмов построения моделей процессов, разработке метода извлечения моделей процессов из журналов событий с использованием эвристики устаревания данных, разработке усовершенствованного алгоритма поиска развилок в моделях процессов, формировании классификации признаков, влияющих на процесс (с учетом зависимости между его элементами), модернизации алгоритма деревьев классификации и регрессии для учета устаревания данных, а также разработке целостной методики прогнозирования показателей предприятия с использованием журналов событий на основе математического моделирования его бизнес-процессов, учитывающего устаревание данных и внешние факторы, влияющие на ход бизнес-процессов.

Публикации. Основные теоретические и практические результаты диссертации опубликованы в 9 работах, среди которых 5 работ в ведущих рецензируемых изданиях, рекомендуемых ВАК, и 1 работа – в материалах международных конференций.

Структура и объем диссертации. Диссертация состоит из введения, 4 глав и заключения (144 стр., 59 рисунков, 30 таблиц), список литературы включает 90 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, отмечена практическая значимость полученных результатов.

В первой главе представлен обзор состояния исследований в части прогнозирования показателей деятельности организаций, рассмотрен метод прогнозирования Розинат-Аалста, основанный на извлечении знаний из журналов событий, и его разновидности; описаны пути развития дискретно-событийного моделирования бизнес-процессов с использованием журналов событий.

Существенная часть данных, характеризующих течение бизнес-процессов, порождаемых информационными системами уровня предприятия, может быть формализована особым образом и представлена в виде единого формата журналов событий. Журналы представляют собой набор информации о последовательности событий, происходящих в рамках выполнения определенного экземпляра бизнес-процесса, а также о контексте, который их сопровождает. В работе рассмотрены два формата журналов событий (MXML и Open XES), отмечены их достоинства и недостатки.

Для прогнозирования бизнес-процессов рассмотрены различные виды моделирования: аналитическое, комбинированное, а также имитационное (включая системно-динамическое, дискретно-событийное и агентное моделирование). Дискретно-событийное моделирование наиболее перспективно для прогнозирования в ближайшей перспективе, к его недостаткам можно отнести следующее: в реальных ситуациях прогнозы, сделанные с помощью этих систем, часто оказываются неточными, что затрудняет их практическое использование. Выделяют три основные причины этого: ориентированность на дизайн (фокус на прогнозировании устойчивых состояний бизнес-процесса, а не конкретных ситуаций, в которых возникает потребность прогнозирования), ручная настройка существенного числа элементов имитационной модели и упрощенные модели ресурсов.

Рассмотрены альтернативные методы, полученные в рамках дисциплины «process mining», которая фокусируется на извлечении знаний из журналов событий информационных систем (рис. 1).

Рисунок 1 – Построение и использование моделей процессов на основе дисциплины «process mining»

Рассмотренные методы перспективны, но не лишены недостатков, которые проиллюстрированы на примере метода Розинат-Аалста. Также указаны пути развития методов моделирования: усовершенствование алгоритмов извлечения модели потоков работ из журналов событий, учет неактивных экземпляров процесса, уточнение модели процесса с использованием зависимостей от контекстных признаков и учет устаревания данных. Эти положения послужили основой исследований, описываемых в следующих главах.

Во **второй главе** сформулированы основные понятия, используемые в работе, а также представлен метод дискретно-событийного моделирования.

К основным понятиям, используемым в работе, относятся: «событие», «экземпляр процесса», «журнал событий», «модель процесса», «маркированная сеть Петри», «последовательность срабатывания переходов» и «развилка процесса». Базовой единицей информации, используемой в области анализа процессов, является «*событие*». Под событием (e) понимается изменение состояния процесса, *фиксируемое* информационной системой, $e \in E$, где E – множество всех событий, C – множество строковых идентификаторов классов событий, а $Cls(E): E \to C$ – функция, ставящая в соответствие каждому событию название класса, к которому оно относится. В работе используются два связанных понятия: *экземпляр процесса* (ЭП) и *журнал событий* (лог). Под ЭП понимается последовательность событий: $\sigma \in E^*$. В ЭП последовательно накапливаются фиксируемые информационной системой предприятия события, происходящие в рамках некоторого процесса. Например, если в ка-

честве процесса выступает «ремонт оборудования», то экземплярами этого процесса могут быть «ремонт котла №1», «ремонт конвейерной линии» и др., каждый из которых включает свои последовательности реально произошедших событий. Под *журналом событий* в работе понимается комплект ЭП $l \in B(E^*)$, L – множество всех логов. В логах накапливается информация обо всех ЭП. *Под моделью процесса* (МП) понимается способ задания подмножества всех последовательностей событий, $M \subseteq E^*$. МП определяет множество допустимых ЭП для данного процесса. Под *сетью Петри* (СП) в работе понимается ориентированный двудольный граф $СП = (P,T,F,V,R,H)$, где: P – конечное множество *позиций*, T – конечное множество *переходов*, при этом $P \cap T = \varnothing$, $F \subseteq (P \times T) \cup (T \times P)$ – множество направленных дуг, называемое *отношением маршрутизации*, V *– множество строковых имен*, $R(T): T \to V$ *– функция именования переходов*, а $H(T): T \to \{1,0\}$ *– функция задания скрытых переходов*. Переход t называется *скрытым*, если $H(t) = 1$, в противном случае переход называется *явным*. При определении соответствия между ЭП из лога и МП в нотации СП подразумевается *соответствие* события e ЭП и перехода t СП, когда $Cls(e) = R(t)$. Скрытые переходы не соответствуют ни одному событию в логах, но выступают в роли вспомогательных конструкций, позволяющих СП более точно описывать маршрутизацию процесса. Для описания динамики процессов используется понятие *маркированной сети Петри* $МСП = (СП, M)$, где $СП = (P,T,F,V,R,H)$ и $M \in B(P)$ – комплект над множеством позиций, также называемый *маркировкой*. Каждый элемент комплекта назовем *жетоном*. Множество всех МСП обозначим Q. Назовем элементы множества $P \cup T$ *узлами*. Узел x является *входящим* для y, если $(x,y) \in F$. Множество входящих узлов для y задается как $\bullet y = \{x \mid (x,y) \in F\}$. Узел x является *исходящим* для y, если $(y,x) \in F$. Множество исходящих узлов для y задается как $y\bullet = \{x \mid (y,x) \in F\}$. Переход $t \in T$ является *разрешенным*, что обозначается как $(СП, M)\left[t\right\rangle$, тогда и только тогда, когда $\bullet t \le M$. Множество разрешенных переходов для маркировки M обозначим как $T(M)$. Для описания смены состояний в МСП вводится *правило срабатывания* $\_\left[\_\right\rangle\_ \subseteq Q \times T \times Q$ – это отношение, выполняемое для любых Q и $t \in T$ если $(СП, M)\left[t\right\rangle \Rightarrow (СП, M)\left[t\right\rangle (СП, (M \setminus \bullet t) \cup t\bullet)$. *Последовательностью срабатываний переходов* (ПСП) для МСП $(СП, M_0)$ будем называть кортеж $\sigma = t_1 t_2 .. t_n$, если $\exists n \in N$ такое, что для маркировок $M_1, M_2, ..., M_n$ и переходов $t_1, t_2, ..., t_n$ для $\forall i \in N$ и $0 \le i < n$ выполняется $(СП, M_i)\left[t_{i+1}\right\rangle$ и $M_{i+1} = (M_i \setminus \bullet t_{i+1}) \cup t_{i+1}\bullet$. ПСП M_0 для МСП будем называть *начальной маркировкой. Достижимой* из M_0 маркировкой M_{0X} назовем такую, для которой существует ПСП, приводящая из M_0 в M_{0X}. Маркировка называется *конеч-*

ной M_K тогда и только тогда, когда в ней не существует разрешенных переходов. M_{0K} будем обозначать конечную маркировку, которая достижима из начальной. Определим *отношение связности* l для позиций: $l(p_1,p_2)=1$, если $(p_1\bullet\cap p_2\bullet)\neq\varnothing$, в противном случае $l(p_1,p_2)=0$. Определим *расширенное отношение связности* L: $L(p_k,p_m)=1$, если $l(p_k,p_m)=1$ или $\exists p_1,p_2,...,p_i$, $i\in N\wedge i>0: l(p_k,p_1)=1; l(p_1,p_2)=1;...; l(p_i,p_m)=1$, в противном случае $L_i(p_k,p_m)=0$. *Конструкцией объединения* назовем такое $p\in P$, для которого $|\bullet p|>1$. *Конструкцией конфликта* назовем пару (P_L,T_L), где P_L – множество позиций, T_L – множество переходов, для которых: $\forall p_k,p_m\in P_L: L(p_k,p_m)=1\wedge T_L=P_L\bullet\wedge|T_L|>1\wedge|P_L|>1$. *Развилкой процесса* назовем пару (P_F,T_F), где P_F – множество позиций, T_F – множество переходов, для которых:

1. $\exists M$, $P_F=\{p\mid M(p)>0\}$, $T_F=T(M)$,
2. $|T_F|>1$.

Если $|P_F|>1$, то $\forall p_1,p_2\in P_F\ \ L(p_1,p_2)=1$.

Предлагаемый метод дискретно-событийного моделирования, сформулированный на основе представленной формальной понятийной базы, разбивается на пять основных этапов, каждый из которых состоит из нескольких подэтапов.

На ***этапе фильтрации*** производится подготовка данных для построения имитационной модели. Фильтрации подвергаются экземпляры процесса, *редко встречающиеся* в журналах, *устаревшие* экземпляры процесса, а также *неактивные* экземпляры процесса. *Редко встречающимися* экземплярами процесса признаются встретившиеся в логе менее трех раз. Все редкие экземпляры отсеиваются и далее не участвуют в анализе. *Устаревшими* признаются экземпляры процесса, для которых при представлении последовательности событий в виде комплектов выполняется следующее неравенство:

$$\left\lfloor \frac{Ti_G - Max(Ti_E^{\sigma_k})\big|_{k=1..n}}{Max(Ti_E^{\sigma_i} - Ti_S^{\sigma_j})\big|_{\substack{i=1..n\\ j=i+1}}} \right\rfloor > 0,$$ где Ti_G – момент получения журнала событий, $Ti_S^{\sigma_i}$ – момент начала экземпляра процесса с индексом i. $Ti_E^{\sigma_i}$ – момент окончания экземпляра процесса с индексом i, *Max* – получение максимального значения множества. Устаревшие экземпляры процесса далее не участвуют в рассмотрении. *Неактивными* экземплярами процесса считаются такие, для которых время пребывания в последнем состоянии оказывается больше времени пребывания в этом состоянии завершенных процессов и активных экземпляров процесса. Все неактивные экземпляры процесса исключаются из дальнейшего рассмотрения. В рамках первого этапа также выполняется сопоставление *весов* экземплярам процесса. Каждому экземпляру процесса в зависимости от времени его начала ставится в соответствие весовой коэффициент из промежутка [0,1), характеризующий разницу между

временем получения журнала событий и временем начала экземпляра процесса.

Этап построения модели процесса позволяет, используя фильтрованные данные, автоматизировать построение модели процесса в нотации сетей Петри. В работе показывается, что наиболее применим к реальным ситуациям алгоритм, скомбинированный из генетического и эвристического алгоритмов. Такая комбинация позволяет получать модели с хорошими показателями качества за строго фиксированное время.

Этап построения модели среды включает в себя несколько подэтапов. Вначале строится *множество внутренних признаков*, которое включает в себя признаки, связанные с историей протекания процесса, а также статические (неизменные для одного экземпляра процесса) и динамические признаки (изменяющиеся в рамках одного экземпляра процесса). *Множество внешних признаков* интерпретируется как входные данные и формируется вручную. После формирования множества внутренних признаков для каждого из них строится математическая модель зависимости от других признаков (как внешних, так и внутренних). Для внешних признаков математические модели задаются вручную. Этап заканчивается построением единого множества, включающего как внутренние, так и внешние признаки.

Этап насыщения модели зависимостями от признаков состоит из нескольких подэтапов. Вначале по модели потока работ формируется *множество элементов модели процесса*, для которых на следующем шаге рассчитываются математические модели зависимостей от признаков. Элементами модели процесса являются: множество событий, множество развилок модели процесса и временная задержка следующей инициализации экземпляра процесса. Для каждого элемента из множества событий строится математическая модель длительности протекания события. Для каждого элемента из множества развилок строится математическая модель вероятностей выбора альтернатив. Строится модель временной задержки следующей инициализации экземпляра процесса. Данные для построения математических моделей берутся из журнала событий. В качестве множества признаков, которые могут оказывать влияние на элементы процесса, используются элементы единого множества признаков. Пример элементов модели процесса представлен на рис. 2.

На этапе *объединения моделей* производится слияние модели среды и моделей элементов процесса. Полученный объект содержит всю необходимую информацию для проведения имитационных экспериментов.

Этап имитационного эксперимента включает выбор периода прогнозирования, задание множества выходных переменных, для которых требуется получить прогноз. Имитационный эксперимент предполагает многократный прогон имитационной модели для выбранного периода прогнозирования. Усредненные значения прогнозируемых переменных по результатам прогонов рассматриваются как прогнозы значений этих переменных для выбранного периода.

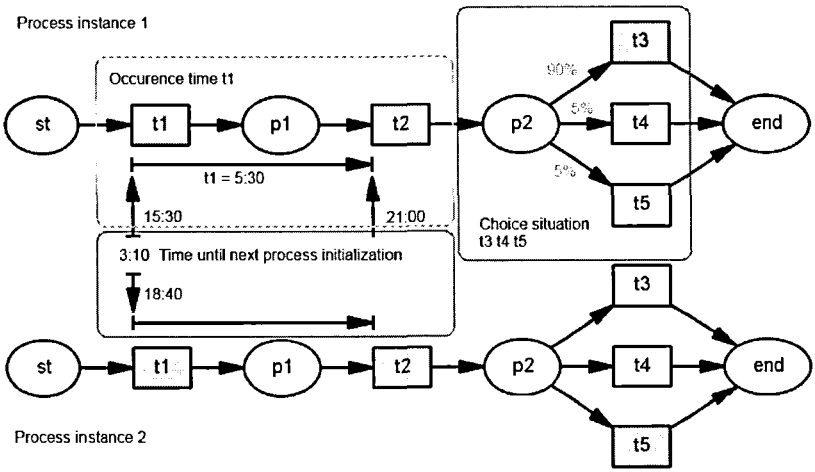


Рисунок 2 – Пример элементов модели процесса

В третьей главе подробно описаны процедуры, применяемые в рамках каждого из этапов (см. главу 2). К описываемым процедурам относятся:

- фильтрация устаревших комплектов экземпляров процессов;
- фильтрация неактивных экземпляров процессов;
- построение модели процесса на основе комбинации генетического и эвристического алгоритмов;
- сопоставление весов экземплярам процесса;
- построение множества контекстных признаков;
- поиск развилок процесса;
- формализации математических зависимостей отклика от предикторов.

Фильтрация устаревших комплектов экземпляров процессов предполагает поиск множества экземпляров процессов, которые в представлении в виде комплекта, являются устаревшими. Используется алгоритм, позволяющий избежать проблем пороговых алгоритмов. При использовании пороговых алгоритмов в выборе слишком большого порога попавшие значения могут оказаться неактуальными к настоящему моменту, а при выборе слишком малого порога некоторые значения, которые могут повториться в будущем, исключаются из рассмотрения. Основная идея алгоритма основывается на предположении, что устаревшими могут быть такие комплекты экземпляра процесса, период отсутствия которых к моменту построения модели превышает период отсутствия, встречающийся в журнале событий до этого.

Цель фильтрации неактивных экземпляров процесса – определить, какие экземпляры более не будут использоваться и удалить их из дальнейшего

рассмотрения. Алгоритм основан на том, что время протекания некоторого события в экземпляре процесса, не достигшем завершающего состояния, должно быть не больше максимального времени протекания этого события, встречающегося в журнале событий у завершенных процессов.

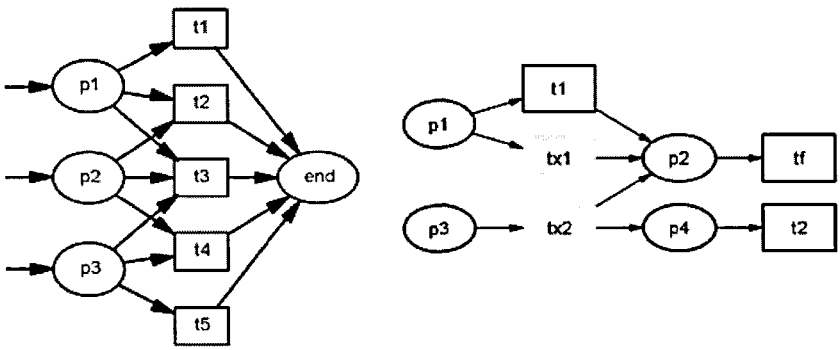
Для нахождения метода построения модели потока работ были проведены эксперименты с α-алгоритмом, эвристическим алгоритмом и генетическим алгоритмом. Для сравнения используются три критерия: время работы алгоритма (T_A измеряется в секундах); соответствие модели журналам событий: $f = \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{k} n_i m_i}{\sum_{i=1}^{k} n_i c_i}\right) + \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i w_i}\right)$, где n_i – число раз, сколько i-й ЭП встречается в логе, m_i – сумма недостающих жетонов при прогоне i-го ЭП по логу, r_i – число лишних жетонов, оставшихся после прогона ЭП по логу, c_i – число порожденных и w_i – число жетонов, поглощенных по правилу срабатывания ($0 \le f \le 1$); адекватность модели: $aaB = 1 - \frac{\sum_{i=1}^{k} n_i (x_i - 1)}{(m-1) \cdot \sum_{i=1}^{k} n_i}$, где x_i – сумма числа разрешенных переходов, m – количество явных переходов ($0 \le aaB \le 1$). В третьей главе описываются схема экспериментов, а также метод нахождения модели, разработанный с учетом результатов экспериментов, описанных в главе 4. Этот алгоритм предполагает комбинирование эвристического и генетического алгоритмов для достижения наилучшего качества моделей при ограничении времени ее построения.

Сопоставление весов экземплярам процесса используется для учета устаревания данных. Каждому экземпляру процесса присваивается вес в промежутке [0,1) в соответствии с временем его начала: чем ближе ЭП к началу лога, тем больше его значение стремится к 0, чем ближе к окончанию – тем ближе к единице. Все значения атрибутов этого ЭП, используемые для построения математических моделей зависимостей от данных, учитываются с весовым коэффициентом.

Единое множество контекстных признаков формируется из множества внутренних и внешних признаков. Внутренними называются такие признаки, значения которых порождаются и/или изменяются только внутри одного экземпляра процесса. К таким признакам относятся исторические, статические и динамические. Выделяется два типа исторических признаков: наличие события *e* в ЭП; число случаев наступления события *e* в ЭП. К статическим относятся признаки, значение которых неизменно для ЭП, в противоположность динамическим, значение которых может изменяться в рамках ЭП. Внешними являются такие признаки, значения которых могут изменяться не только в рамках конкретного ЭП. Внешние признаки разделяются на глобальные и организационные. Глобальными являются любые переменные, не относящиеся напрямую к процессу и организации. К организационным отно-

сятся внутренние признаки организации – наличие ресурсов, служебные расписания и другие. Внешние признаки задаются экспертом.

Одним из элементов модели процесса является множество развилок. Под развилкой процесса понимается ситуация, в которой экземпляр процесса может пойти по одному из нескольких исключающих друг друга путей. В методе Розинат-Аалста предполагается поиск таких ситуаций, но их интерпретация является узкой и не применима для всех случаев. Метод Розинат-Аалста не позволяет искать развилки в ситуациях конфликта, а также не полностью учитывает ситуацию скрытых переходов (рис. 3).



а) б)

Рисунок 3 – Конструкции, которые не учитывает метод Розинат-Аалства: а) конструкция конфликта, б) конструкция со скрытыми переходами

Предложен алгоритм, позволяющий корректно обрабатывать конструкции с конфликтами и со скрытыми переходами. Анализ ситуаций конфликта возможен после переопределения понятия развилки и включения в него позиций сети Петри, которые формируют ситуацию конфликта. Полнота учета скрытых переходов достигается с помощью «заглядывания» вперед при анализе журнала событий до получения информации о сработавших скрытых переходах.

Также в данной главе описана процедура построения моделей зависимостей от данных, которая применяется для элементов модели процесса и внутренних признаков. Предлагается использовать методы классификации для нахождения моделей неметрических атрибутов и методы регрессии – для нахождения моделей метрических атрибутов. Для получения значений по модели в имитационных экспериментах для метрических атрибутов предложено использовать метод сглаженных гистограмм. Сглаживание осуществляется с помощью непараметрических ядерных методов, оценка плотности распределения в которых задается формулой: $f(x)=\frac{1}{nh}\sum_{i=1}^{n}K(\frac{x_i-x}{h})$, где K – функция ядра, h – ширина окна, n – число значений в окне. Оптимальная ширина

окна задается максимизацией логарифма функции правдоподобия: $\sum_{i=1}^{n}\log\left(\frac{1}{(n-1)h}\sum_{j=1,j\neq i}^{n}K\left(\frac{x_j-x}{h}\right)\right)$. Максимизация выполняется алгоритмом оптимизации роем частиц (Particle Swarm Optimisation). Сглаженная гистограмма проходит процедуру дискретизации, что позволяет рассчитывать конкретные значения в имитационных экспериментах. Для неметрических атрибутов генерация конкретных значений производится с учетом частот соответствующих значений.

В **четвертой главе** представлены результаты экспериментальных исследований метода прогнозирования показателей с использованием журналов событий, а также его отдельных процедур. Представлены результаты исследования алгоритмов извлечения модели потока работ: α-алгоритма, эвристического и генетического алгоритмов. Задачей исследования было сравнение качества построенных этими алгоритмами моделей как в синтетических экспериментах, так и в экспериментах с реальными данными. Использовались три журнала событий реальных компаний для процесса продажи автомобилей (Car_Sell), ремонта автомобилей (Car_Rep) и ремонта оборудования на фабрике (Fac_Rep). В результате проведенных экспериментов было показано, что эвристический алгоритм показывает наилучшие результаты в синтетических тестах в большинстве случаев. На реальных данных (табл. 1) он несколько уступает генетическому алгоритму в двух экспериментах и незначительно превосходит его в третьем. При этом адекватность моделей, полученных эвристическим алгоритмом, является низкой по сравнению с генетическим алгоритмом.

Таблица 1 - Результаты экспериментов по сравнению алгоритмов извлечения моделей потоков работ

Критерий	Car_Sell			Car_Rep			Fac_Rep		
	α	G	H	α	G	H	α	G	H
f	0,62	0,99	0,94	0,68	1	0,90	0,73	0,96	0,99
aaB	1	0,63	0,41	1	0,67	0,48	1	0,61	0,76
T_A	1,18	2278,30	1,97	3,23	4719,90	5,66	2,01	2854,70	3,25

С другой стороны, преимуществом эвристического алгоритма является время работы, которое сравнимо со временем работы α-алгоритма. Качество моделей, полученных α-алгоритмом значительно хуже качества моделей, полученных другими алгоритмами в большинстве экспериментов.

Кроме того, в данной главе приведены результаты синтетического эксперимента, демонстрирующие преимущества алгоритма фильтрации устаревших комплектов экземпляров процесса над пороговыми методами фильтрации, описанные в третьей главе. Также приведен результат исследования алгоритма поиска развилок по модели процесса на синтетической модели, в

которой присутствуют сложные конструкции конфликта и скрытые переходы. Показано, что предложенный в диссертационной работе алгоритм лучше справился с задачей определения развилок, чем алгоритм Розинат–Аалста.

Приведены примеры поиска внутренних контекстных признаков на синтетических моделях, также представлены результаты экспериментов по выбору наиболее подходящего метода классификации и регрессии. В экспериментах сравнивались следующие методы формализации метрических зависимостей: среднее значение, линейная параметрическая регрессия, деревья регрессии на основе энтропии. Для задачи классификации сравнивались пять методов: среднее значение, наивный байесовский классификатор, логистический классификатор, классификатор на основе нейронных сетей и деревья классификации на основе энтропии. Эксперименты показали, что наиболее подходящими являются деревья классификации и алгоритм регрессии на основе энтропии.

Для обоснования работоспособности метода в целом выполнен эксперимент по прогнозированию показателей деятельности предприятия на основе журналов информационных систем реальных предприятий. Были использованы журналы событий бизнес процессов продажи и ремонта автомобилей автодилера и ремонта оборудования парфюмерной фабрики. Для сравнения использовались предложенный в работе метод (REP), алгоритм Розинат-Аалста (ROZ) и расчет «по среднему» (AVG). Исследовалось влияние выбранного множества признаков, периода прогнозирования и удаленности от реальных данных на результаты рассматриваемых методов. При оценивании качества использовались следующие виды ошибок: $BIAS = \frac{\sum_{t=1}^{N} E_t}{N}$, $PBIAS = \frac{\sum_{t=1}^{N} \frac{E_t}{Y_t}}{N}$, $MAE = \frac{\sum_{t=1}^{N} |E_t|}{N}$ и $MAPE = \frac{\sum_{t=1}^{N} \left|\frac{E_t}{Y_t}\right|}{N}$, где N – число экспериментов, E_t – разность реального значения и значения прогнозируемого показателя в конкретном эксперименте t. В табл. 2 приведены некоторые результаты экспериментов.

В результате проведения экспериментов определено, что предложенный метод улучшает качество прогнозов с увеличением периода прогнозирования (эксперимент 1, 2 из табл. 2). Также существенное влияние оказывает выбор признаков прогнозирования, так для экспериментов 3 и 4 на процесс влияли «биржевые» признаки и их использование позволило сократить MAPE в 3 раза (эксперимент 3), в то время как отсутствие нужных признаков существенно ухудшило качество прогнозов (эксперимент 4). В экспериментах 5 и 6 показано влияние удаленности от реальных данных на качества прогнозов. При отсутствии удаленности (эксперимент 5) качество прогнозов оказывается существенно выше, чем при значительной удаленности.

Таблица 2 – Результаты экспериментов

Эксперимент	Алгоритм	BIAS	PBIAS (%)	MAE	MAPE (%)
1 (Период:1 месяц, эксперимент 5)	REP	-238	-27	309	35
	ROZ	-351	-35	351	39
	AVG	-644	-73	644	73
2 (Период:6 месяцев, эксперимент 5)	REP	-147	-16	147	16
	ROZ	-354	-39	354	39
	AVG	-793	-88	793	88
3 (Признаки: «Дата», эксперимент 2)	REP	2917	41	7613	107
	ROZ	3059	43	6397	92
	AVG	4766	67	6134	86
4 (Признаки: «Биржевые», эксперимент 2)	REP	-1422	-20	2521	36
	ROZ	3059	43	6397	92
	AVG	4766	67	6134	86
5 (Реальные данные удалены на 0 месяцев, эксперимент 6)	REP	-1	-2	1	2
	ROZ	-8	-20	8	20
	AVG	6	14	6	14
6 (Реальные данные удалены на 9 месяцев, эксперимент 6)	REP	-14	-61	14	61
	ROZ	-22	-96	22	96
	AVG	-10	-43	10	43

Эксперименты показывают, что точность результов предложенного в работе метода в среднем превосходит точность метода Розинат-Аалста на реальных данных: от 5 % – в неблагоприятных условиях (неверное множество признаков, малый период прогнозирования, большая удаленность от реальных данных), до 15% – в благоприятных (верно подобранное множество признаков, большой период прогнозирования, малая удаленность от реальных данных).

Заключение. В ходе выполнения диссертационной работы получены следующие основные результаты:

- разработан метод, автоматизирующий построение динамической модели бизнес-процессов на основе данных, извлекаемых из журналов событий, с учетом их устаревания, наличия развилок, а также неактивных экземпляров процессов;
- разработана процедура идентификации и моделирования характерных элементов динамической модели бизнес-процессов с учетом зависимости от контекстных переменных на основе классификации внешних факторов;
- разработан метод имитационного моделирования динамики бизнес-процессов в условиях изменчивости внешней среды для прогнозирования показателей деятельности предприятия;
- разработано алгоритмическое и программное обеспечение для экспериментальных исследований перечисленных методов и моделей на языке Java 1.6 с использованием библиотек Prom Framework 5 и Weka 3.6.4;
- прогностические свойства предложенного метода экспериментально сравниваются со свойствами существующих аналогов, выполнена апробация метода на прикладных задачах моделирования деятельности предприятий, результаты исследований продемонстрировали, что в среднем точность прогнозов повышается до 15 % по сравнению с аналогом.

Научные работы по теме диссертации,

опубликованные в изданиях, определенных ВАК

1. *Ходырев И.А., Попова С.В.* Сравнение алгоритмов process mining для задачи поиска моделей процессов // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. 2011. Т. 2, № 138. С. 170–178.

2. *Ходырев И.А. Татаринов Ю.С.* Прогнозирование показателей организации с использованием журналов событий // Известия СПБГЭТУ (ЛЭТИ). 2011. № 5. С. 53-59

3. *Ходырев И.А., Попова С.В.* Обнаружение развилок в моделях процессов // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. 2012. Т. 3, № 150. С. 82–88.

4. *Ходырев И.А., Бухановский А.В.* Метод моделирования бизнес-процессов в информационных системах на основе журналов событий // Динамика сложных систем–XXI век. 2013. № 3. С. 46–50.

5. *Khodyrev I., Popova S.* Discrete modelling and simulation of business processes using event logs // 14th Intern. Conf. on Computational Science. Proc. Comp. Sci. Elsevier, 2014. Vol .29. P. 322–331.

Другие публикации

6. *Ходырев И.А.* Многоцелевой алгоритм выработки рекомендаций по журналам событий // Матер. XI Междунар. науч.-техн. конф. «Информационно-вычислительные технологии и их приложения». Пенза, 2009. С. 253–258.

7. *Ходырев И.А.* Построение социограмм на основе журналов СУБП с использованием технологии Process Mining // Матер. XXIII Междунар. науч.-техн. конф. «Математические методы и информационные технологии в экономике, социологии, образовании». Пенза, 2009. С. 66–69.

8. *Ходырев И.А.* Создание поведенческой модели искусственного интеллекта с использованием журналов событий // Матер. XXIV Междунар. науч.-техн. конф. «Математические методы и информационные технологии в экономике, социологии, образовании». Пенза, 2009. С. 78–81.

9. *Ходырев И. А., Татаринов Ю.С.* Технология Process Mining и оценка эффективности представленных в ней алгоритмов // Матер. XXXVIII Междунар. науч.-практ. конф. «Неделя науки СПбГПУ». Ч. XVIII. СПб, 2009. С. 51–53.

Формат: 60x84 1/16 Печать офсетная.
Бумага офсетная. Гарнитура Times.
Тираж: 100 экз. Заказ: 414 Отпечатано:
Учреждение «Университетские телекоммуникации»
197101, Санкт-Петербург, Саблинская ул., д.14
+7(812) 9151454, zakaz@tibir.ru, www.tibir.ru