

Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

На правах рукописи

Шершаков Сергей Андреевич

**МЕТОДЫ И ИНСТРУМЕНТЫ ПОВЫШЕНИЯ
ЭФФЕКТИВНОСТИ АЛГОРИТМОВ
МАЙНИНГА ПРОЦЕССОВ**

РЕЗЮМЕ

диссертации на соискание ученой степени
кандидата компьютерных наук

Москва — 2020

Работа выполнена в Федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет “Высшая школа экономики”».

Научный руководитель: **Ломазова Ирина Александровна**

профессор, доктор физико-математических наук,
Факультет компьютерных наук Национального исследовательского университета «Высшая школа экономики»

Тема диссертации

Последние два десятилетия обозначились бурным развитием информационных технологий. Возросли мощности оборудования, сократился его размер и вес. Вкупе с понижением энергопотребления современной электроникой все это дало мощный толчок к развитию мобильной потребительской цифровой техники (субкомпактные лэптопы, планшеты и смартфоны), а также носимых устройств (например, умные часы). Другим аспектом стало широкое повсеместное проникновение сетей передачи данных, в первую очередь, интернета. Развиваются новые поколения беспроводных сетей. Стоимость 1 Мб переданной информации неуклонно падает, а скорости передачи информации — повышаются.

Два этих обстоятельства существенно изменили структуру и потоки передачи информации. Интернет как глобальная коммуникационная структура все чаще рассматривается не только как технологический феномен, но и в различных проекциях: традиционная проекция интернета контента; в виде интернета сервисов для бизнес-проекции; в виде интернета людей и интернета мест в социальной проекции; наконец, даже вещи получили свой интернет вещей. Все вместе они порождают новую большую сущность — *интернет событий*.

Цифровизация общества также происходит в разных направлениях. Это проявляется не только в появлении новых цифровых инструментов, но и в перемещении традиционных оффлайновых сервисов с их очередями и длительным ожиданием в онлайн с доступом 24/7 и быстрым получением результата. Интернет-магазины, интернет-банки, заказ билетов, госуслуги, онлайн-образование — все эти сервисы нашли свое отражение в виртуальном пространстве. Даже такие классические услуги, как такси, стремительно переносят в интернет функционал, не связанный непосредственно с движением автомобиля. Бизнесы, сделавшие ставку на классические оффлайновые бизнес-процессы, в одночасье оказались неконкурентоспособными; этот феномен, по примеру тех же такси, стали называть *уберификацией*.

Разумеется, одного только скоростного интернета и быстрых телефонов недостаточно для того, чтобы перевести бизнес в цифру. Огромная задача ложится на специальный класс систем поддержки бизнес-процессов, называемых *процессно-ориентированными информационными системами* (ПОИС). Они представляют собой сложные распределенные программные комплексы, берущие на себя большое число функций по инициализации, сквозной поддержке и завер-

шению бизнес-процессов. Функционируя, такие системы накапливают большое число побочных данных, записываемых в *логах*. Логи могут рассматриваться с технической точки зрения (*системные логи*) или с точки зрения *следов поведения системы* относительно событий бизнес-процесса (*логи событий*). Последние являются ценным источником знаний о реальном поведении системы. Сравнивая реальное поведение с ожидаемым, можно обнаружить массу полезной информации, как-то: несоответствие реального и желаемого поведения системы (что особенно актуально при наличии человеческого фактора); узкие места и проблемы с производительностью; анти-паттерны построения бизнес-системы и т.д.

Изучением этих задач занимается дисциплина «извлечение и анализ моделей процессов» (англ. *process mining, PM*) [1]. Субъектами этой дисциплины являются бизнес-аналитики, владельцы процессов, исследователи данных. Ее основными объектами являются бизнес-процессы. Процессы представляются различными абстракциями в виде статистических, графических и других видов моделей, а также конкретными выполненными экземплярами процессов, записанными в логах событий.

Извлечение процессов тесно связано с *добычей данных* (data mining), *машинным обучением* (machine learning), моделированием и анализом моделей процессов. Основные задачи и цели этого направления исследований изложены в Манифесте (Process Mining Manifesto) [2] и могут быть кратко представлены тремя ключевыми проблемами: 1) извлечение модели процесса из журнала событий (process discovery), 2) проверка соответствия некоторой модели реальным данным, записанным в журнале событий (conformance checking), и 3) исправление и улучшение процесса с учетом изменяющихся данных (process enhancement).

Для описания моделей процессов используются различные средства — *языковые нотации* (workflow notations) и *программные инструменты* (workflow engines/systems). В настоящий момент известно значительное число нотаций для описания процессов, потоков работ (в основе большинства из них лежат *сети Петри*) и систем, реализующих эти нотации в программных инструментах для конечного пользователя. В качестве примера можно привести нотацию для описания бизнес-процессов BPMN [3], язык исполнения бизнес-процессов BPEL [4], язык моделирования процессов YAWL [5].

Системы переходов (transition systems, TS) являются одной из наиболее естественных нотаций для представления журналов событий. Они могут рассматриваться и как самостоятельная модель представления журнала событий, и в ка-

честве промежуточной модели, на основании которой с помощью специальных алгоритмов можно получать другие модели, учитывающие, например, явления параллелизма (сети Петри).

Сети Петри являются низкоуровневым формализмом с хорошей теоретической базой и позволяют проверять такие свойства моделируемых систем, как *безопасность*, *ограниченность*, *достижимость* и др. По причине своей низкоуровневости сети Петри не всегда используются в качестве конечной модели для анализа процессов; однако они являются базовыми моделями для построения на их основе таких высокоуровневых моделей, как BPMN, которые используются для описания бизнес-процессов, или UML Activity Diagram [6], которые используются для описания поведения информационных систем.

Несмотря на то, что извлечение процессов является относительно новым направлением, оно уже активно применяется при моделировании и анализе бизнес-процессов [7] в менеджменте, при разработке программного обеспечения [8; 9], в управлении технологическими процессами, в медицине [10; 11].

Одна из ключевых проблем, с которыми сталкивается дисциплина process mining, — большое количество данных для анализа, генерируемых ПОИС. Данная проблема известна как *проблема больших данных* (Big Data) [12] и включает несколько аспектов (четыре «V» больших данных): *размер* (volume), *скорость генерирования* (velocity), *разнообразие* (variety) и *достоверность* (veracity). К настоящему моменту уже разработано и продолжает разрабатываться большое число алгоритмов *синтеза* и *анализа* моделей процессов. Они основаны на разных концепциях, различаются *сложностью* по времени/памяти, а так же *точностью* и *достоверностью* получаемых с их помощью результатов.

В контексте *больших данных* лишь немногие алгоритмы способны выдавать какой-либо результат за приемлемое время. Еще меньше алгоритмов способны выдавать результат, удовлетворяющий критериям точности и достоверности. Например, алгоритм построения *нечетких карт* (fuzzy maps) [13] широко используется в большинстве существующих инструментов process mining, причем во многих он является основным, если не единственным. Это определяется скоростью его работы и простотой получаемых результатов. Позволяя получать первичную модель для больших входных данных, он, тем не менее, не дает возможности для представления и анализа многих важных аспектов сложных процессов, таких как *параллелизм* и *наличие дедлоков*. Противоположным примером является алгоритм на основе *теории регионов*, который позволяет получить точную и теоретически

обоснованную модель в виде сети Петри. Однако временная сложность работы этого алгоритма существенно выше.

Проблема процессных моделей и алгоритмов, которые с ними работают, заключается в *предельном размере журнала событий*, который может быть обработан такими алгоритмами. Сложность некоторых важных алгоритмов, имеющая от размера входных данных полиномиальную или экспоненциальную зависимость, не позволяет обрабатывать журналы событий реальных систем с большим количеством сохраненной информации, сохраняемой в этих журналах.

В случае с алгоритмом регионов основной проблемой, ограничивающей его широкое практическое применение, является *временная сложность* этого алгоритма. Доказано, что задача синтеза некоторых разновидностей сетей Петри алгоритмом регионов относится к классу NP-полных задач [14]. Это означает, что временная сложность этого алгоритма имеет экспоненциальную зависимость от размера *системы переходов*, являющейся *входной моделью* для этого алгоритма. Для других классов сетей Петри алгоритм регионов имеет полиномиальную временную сложность от размера входной модели с высокой степенью полинома [15]. Система переходов строится на основе журнала событий, и, в худшем случае, ее размер линейно связан с размером самого журнала, что в случае реальных журналов событий существенно ограничивает практическую применимость алгоритма.

Таким образом, *актуальной становится задача редукции систем переходов*, синтезированных по логам реальных систем, до таких размеров, при которых, с одной стороны, они становятся обозримыми и подходящими для использования со сложными алгоритмами (например, алгоритмом регионов), а с другой, — не искажают поведение реальной системы или делают это в наименьшей степени. Для оценки качественных характеристик синтезируемых моделей, способной показать, насколько модель соответствует реальному поведению процесса, используются *метрики качества модели: степень соответствия (fitness), сложность (simplicity) и точность (precision)* [16].

Несмотря на то, что теоретическую сложность алгоритма регионов невозможно «улучшить», тем не менее, возможно существенно повысить применимость алгоритма за счет его эффективной реализации, а это накладывает специальные требования и ограничения на программные инструменты. Таким образом, *эффективная реализация программных инструментов для process mining также является важной и актуальной задачей*.

Целью данной работы является разработка методов, алгоритмов и инструментов редукции систем переходов с контролем качественных характеристик (метрик), таких как *степень соответствия* (fitness), *сложность* (simplicity) и *точность* (precision).

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Исследовать существующие подходы к задаче редукции систем переходов и других похожих моделей.
2. Разработать алгоритм редукции систем переходов, основанный на частотной характеристике появления индивидуальных активностей в журнале событий.
3. Разработать алгоритм вычисления метрики «точность» (precision) для систем переходов.
4. Разработать критерии адаптивной вариации параметров алгоритма сокращения исходной системы переходов.
5. Разработать расширение метода регионов для синтеза сетей Петри по системам переходов *с петлями*.
6. Исследовать влияние параметров алгоритма сокращения системы переходов на качественные характеристики модели в виде сети Петри, построенной на основе сокращенной системы переходов с помощью алгоритма регионов.
7. Выработать архитектурные принципы проектирования программного инструментария для РМ в контексте работы с большими логами событий реальных информационных систем.
8. Спроектировать и разработать на основе выработанных архитектурных принципов программный инструмент, реализующий: а) алгоритм сокращения системы переходов, б) алгоритм расчета метрик соответствия заданной системы переходов исходному журналу событий, в) алгоритм конвертации системы переходов в сеть Петри, г) алгоритмы визуализации моделей.
9. Провести комплексные эксперименты на журналах событий реальных информационных систем, позволяющие оценить качественные характеристики полученных моделей при различных вариациях параметров алгоритмов.

10. Оценить эффективность — по времени и по памяти — реализации указанных алгоритмов в спроектированном инструменте по сравнению с реализацией тех же алгоритмов в других существующих инструментах для РМ.

Научная новизна

1. Разработан новый параметрический алгоритм редукции систем переходов, основанный на частотных характеристиках активностей журнала событий. По сравнению с существующими подходами «простой» редукции, новый алгоритм показал лучшие результаты как по уменьшению сложности модели, так и по увеличению ее точности. Эксперименты проводились на синтетических и реальных журналах событий.
2. Впервые для задач дисциплины process mining было выполнено исследование влияния размера и качественных характеристик промежуточной модели в виде системы переходов на размер и качественные характеристики целевой модели в виде сети Петри при конвертации с помощью алгоритма регионов.
3. Был разработан подход к представлению больших журналов событий с применением технологии встраиваемых СУРБД, позволяющий эффективно извлекать данные из журналов с возможностью мультиперспективного обращения без перестраивания журнала событий.

Практическая значимость. Данная работа позволит повысить эффективность обработки больших наборов данных алгоритмами process mining с целью построения моделей процессов. Это позволит улучшить применимость алгоритмов, чувствительных к размеру входных данных, к большим журналам событий реальных информационных систем. Использование таких алгоритмов с моделями, которые получены на основе реальных журналов событий, дает возможность более комплексного изучения поведения систем, для которых выполняется построение процессных моделей.

Методология и методы исследования. Методологическую основу исследования составляют теория множеств, теория графов, теория автоматов, теория алгоритмов и математическая логика, теория формальных языков и грамматик, принципы машинного обучения, моделирование, синтез и анализ.

Основные результаты диссертации

Основные результаты данной диссертации, полученные её автором и выносимые на защиту:

1. Метод параметрической редукции систем переходов с сохранением качественных характеристик моделей.
2. Алгоритм частотной редукции систем переходов.
3. Алгоритм вычисления метрики «точность» (precision) для систем переходов.
4. Метод представления больших журналов событий с помощью реляционных баз данных с мультиперспективным извлечением данных.
5. Расширение метода регионов для систем переходов с петлями.
6. Архитектурные принципы, архитектура и реализация расширяемого программного комплекса, использующего предложенные метод адаптивной параметрической редукции систем переходов и синтеза конечных моделей в виде сетей Петри.

Достоверность полученных результатов обеспечивается сравнением с результатами смежных алгоритмов и методов. Результаты находятся в соответствии с результатами, полученными другими авторами.

Публикации и апробация работы

Результаты данной диссертации были опубликованы в российских и международных рецензируемых журналах и трудах конференций.

Публикации

Основные результаты данной диссертации опубликованы в следующих статьях, которые приводятся с разбиением на группы в соответствии с правилами *Диссертационного совета по компьютерным наукам НИУ ВШЭ*: «Публикациями стандартного уровня являются статьи в изданиях, входящих в список рекомендованных журналов НИУ ВШЭ или в системы цитирования Web of Science (квартили Q3, Q4) или Scopus (квартили Q3, Q4), а также в рецензируемых трудах конференций, входящих в рейтинг CORE (ранг B)»¹.

¹https://www.hse.ru/en/science/disscoun/council_computerscience/

Публикации стандартного уровня

1. Sergey A. Shershakov, Anna A. Kalenkova, Irina A. Lomazova. *Transition Systems Reduction: Balancing Between Precision and Simplicity* // Lecture Notes in Computer Science, Vol. 10470: Transactions on Petri Nets and Other Models of Concurrency XII. Berlin, Heidelberg : Springer, 2017. С. 119–139. (Статья в научном издании, проиндексированная SCOPUS Q3; вклад автора 70 %) [17]
2. Shershakov S. *DPMine graphical language for automation of experiments in process mining* // Automatic Control and Computer Sciences. 2016. Vol. 50. No. 7. С. 477–485. (Статья в научном журнале, проиндексированная SCOPUS Q4 и WoS Q4; вклад автора 100 %) [18]
Перевод с русского статьи: Шершаков С. А. Графический язык DPMine для автоматизации экспериментов в области автоматического построения и анализа процессов // Моделирование и анализ информационных систем. 2014. Т. 21. № 5. С. 102–115.
3. Shershakov S. A., Rubin V. A. *System Runs Analysis with Process Mining* // Modeling and Analysis of Information Systems (MAIS). 2015. Vol. 22. No. 6. С. 818–833. (Статья в научном журнале из рекомендованного списка НИУ ВШЭ; вклад автора 70 %) [19]
4. С. А. Шершаков. “VTMine for Visio”: инструмент графического моделирования в области process mining // Моделирование и анализ информационных систем, том. 27 № 2, С. 210–233. (Статья в научном журнале из рекомендованного списка НИУ ВШЭ; вклад автора 100 %) [20]
5. Sergey A. Shershakov. *Multi-Perspective Process Mining with Embedding Configurations into DB-based Event Logs* // Proceedings of the International Conference on Software Testing, Machine Learning and Complex Process Analysis (TMPA-2019). Springer, 2020 (в печати) (Статья в сборнике трудов конференций, проиндексированная SCOPUS; вклад автора 100 %) [21]
6. Shershakov S. A. *DPMine/P: Modeling and process mining language and ProM plug-ins* // ACM International Conference Proceeding Series. 2013. (Статья в сборнике трудов конференций, проиндексированная SCOPUS; вклад автора 100 %) [22]

Прочие публикации

1. *Sergey Andreevich Shershakov. VTMine Framework as Applied to Process Mining Modeling // International Journal of Computer and Communication Engineering. 2015. Vol. 4. No. 3. P. 166–179. (Статья в международном научном журнале; вклад автора 100 %) [23]*
2. *Shershakov S. DPMine/C: C++ Library and Graphical Frontend for DPMine Workflow Language, in: Proceedings of the 8th Spring/Summer Young Researchers' Colloquium on Software Engineering (SYRCoSE 2014). М.: 2014. P. 96–101. (Статья в сборнике трудов конференции; вклад автора 100 %) [24]*
3. *Shershakov S. DPMine: modeling and process mining tool, in: Proceedings of the 7th Spring/Summer Young Researchers' Colloquium on Software Engineering, SYRCoSE 2013 / Отв. ред.: А. Камкин.; Ed. by A. Petrenko, A. Terekhov. Kazan, 2013. P. 19–24. (Статья в сборнике трудов конференции; вклад автора 100 %) [25]*

Другие публикации автора диссертации

Список публикаций автора диссертации не ограничивается статьями, перечисленными выше. Ниже приводятся работы автора по теме извлечения и анализа моделей процессов, результаты которых не представлены в данной диссертации.

1. *N. S. Zubkova, S. A. Shershakov. Method for Building UML Activity Diagrams from Event Logs // Proceedings of the Institute for System Programming of the RAS. 2019. Vol. 31. No. 4. P. 139–150. [26]*
2. *A.M. Rigin, S.A. Shershakov. SQLite RDBMS Extension for Data Indexing Using B-tree Modifications // Proceedings of the Institute for System Programming of the RAS. 2019. Vol. 31. No. 3. P. 203–216. [27]*
3. *Shunin T., Zubkova N., Sergey Shershakov. Neural Approach to the Discovery Problem in Process Mining, in: Analysis of Images, Social Networks and Texts. 7th International Conference AIST 2018 / Ed. by W. van der Aalst, D. I. Ignatov. Springer, 2018. P. 261–273. [28]*
4. *K. Davydova, S. Shershakov. Mining Hybrid UML Models from Event Logs of SOA Systems // Proceedings of the Institute for System Programming of the RAS. 2017. Vol. 29. No. 4. P. 155–174. [29]*

5. Davydova K. V., Shershakov S. A. *Mining Hierarchical UML Sequence Diagrams from Event Logs of SOA Systems while Balancing between Abstracted and Detailed Models* // Proceedings of the Institute for System Programming of the RAS. 2016. Vol. 28. No. 3. P. 85–102. [30]
6. Alexey Mitsyuk, Anna Kalenkova, Sergey A. Shershakov, van der Aalst W. *Using process mining for the analysis of an e-trade system: A case study* // Business Informatics. 2014. Vol. 29. No. 3. P. 15–27. [7]

Конференции и семинары

Результаты данной диссертации были представлены и обсуждены на следующих конференциях и семинарах:

- конференции Conference on Software Testing, Machine Learning and Complex process Analysis (TMPA-2019) (Грузия). Доклад: «Multi-Perspective Process Mining with Embedding Configurations into DB-based Event Logs»;
- 207-м заседании Московской секции ACM SIGMOD. Доклад: «PCУБД в Process Mining»;
- конференции Algorithms & Theories for the Analysis of Event Data (ATAED 2016) (Польша). Доклад: «Transition Systems Reduction: Balancing between Precision and Simplicity»;
- конференции 5th International Conference on Networking and Information Technology (ICNIT 2014) (Сингапур). Доклад: «VTMine Framework as Applied to Process Mining Modeling»;
- конференции Spring/Summer Young Researchers' Colloquium on Software Engineering 2014 (Санкт-Петербург). Доклады: «DPMine/C: C++ Library and Graphical Frontend for DPMine Workflow Language» и «Component-based VTMine/C Framework: Not Only Modelling»;
- семинаре «Unleashing Operational Process Mining» (2013) международной рабочей группы IEEE Task Force on Process Mining (Германия);
- конференции «Разработка ПО» CEE-SECR 2013 (Москва). Доклад: «DPMine/P: modeling and process mining language and ProM plug-ins»;
- конференции Spring/Summer Young Researchers' Colloquium on Software Engineering 2013 (Казань). Доклад: «DPMine: modeling and process mining tool».

Грантовая поддержка

Диссертационная работа была выполнена при поддержке Российского фонда фундаментальных исследований (РФФИ): проект № 18-37-00438 «мол_а»: «Разработка методов эффективной реализации алгоритмов извлечения и анализа моделей процессов на основе больших логов событий» (https://www.rfbr.ru/rffi/ru/rffi_contest_results/o_2058328).

Содержание диссертации

Во **введении** обозначается предметная область работы, дается обоснование актуальности ее темы, определяются цели и задачи, а также раскрывается теоретическая и практическая значимость работы.

В **первой главе** дается необходимый обзор работ по связанным тематикам, вводятся основные определения и понятия, используемые в настоящей работе.

Во-первых, даются определения *журнала событий*, *трассы в журнале*, (*класса*) *активности* и *события*, *системы переходов* и префиксного дерева. Далее, определяется принцип воспроизведения трассы в системе переходов и определяется частичное воспроизведение.

Затем вводятся метрики качества моделей процессов. В работе используются следующие основные метрики: *степень соответствия* модели и журнала событий, *точность* модели и *сложность* модели. Данные метрики ранее определялись для сетей Петри [16], но не определялись для систем переходов. Определения метрик для систем переходов, вводимые в данной работе, согласованы с соответствующими определениями для сетей Петри, что позволяет сравнивать эти два типа моделей. Это свойство используется в дальнейшем при синтезе сетей Петри из промежуточных систем переходов с помощью *алгоритма регионов*.

Глава вторая содержит описание разработанного метода параметрической частотной редукции систем переходов. Представляется новый параметрический алгоритм частотной редукции систем переходов и доказывается его корректность. Далее дается описание нового алгоритма расчета метрики «точность» для систем переходов. Приводится первичный анализ эффективности разработанного метода частотной редукции в сравнении с методом простой редукции. В заключении главы определяется подход к адаптивной редукции промежуточной системы пе-

переходов для построения результирующей модели (сети Петри) с заданными характеристиками.

Метод параметрической частотной редукции систем переходов содержит в своей основе разработанные *алгоритм частотной редукции систем переходов* и *алгоритм расчета метрики «точность» для систем переходов*. Для описания *алгоритма частотной редукции* определяется понятие *системы переходов с k-окном*. Сам алгоритм является трехшаговым. На *первом шаге* алгоритма выполняется построение полной системы переходов. Для этого определяется понятие *частотной характеристики* системы переходов. Показывается, что система переходов, построенная на первом шаге алгоритма, полностью соответствует журналу событий. На *втором шаге* алгоритма выполняется редукция системы переходов, построенной на первом шаге, за счет сокращения частей модели, частотная характеристика которых не превышает некоторого заданного порога *Threshold* (параметр алгоритма редукции). После этого шага *сложность* системы переходов уменьшается, но при этом она перестает полностью соответствовать журналу событий, что является исходным требованием к модели. Система переходов *восстанавливается* на *третьем шаге* алгоритма. При этом учитываются частотные свойства журнала событий, поэтому при восстановлении модели редкие поведения *обобщаются*, а частые поведения *сохраняются*. В результате восстановленная система переходов становится менее *сложной*, чем исходная система переходов, и при этом ее *точность* уменьшается незначительно.

На третьем шаге алгоритма возможно управлять степенью редукции модели с помощью дополнительного параметра V_{wsc} . Для любой системы переходов при соответствующем выборе значений обоих параметров алгоритма редукции возможно получить максимально редуцированную систему переходов, которая также известна как «цветочная модель». Такая система переходов содержит ровно одно состояние и столько переходов-петель, сколько классов активностей содержится в журнале событий. Доказывается, что после третьего шага алгоритма редукции *восстановленная* система переходов *полностью соответствует* журналу событий.

Для оценки метрики *точность* систем переходов разработан новый алгоритм, основанный на симуляции наиболее точной для заданного журнала событий системы переходов (которым полагается *префиксное дерево*) более генерализованной моделью. Последняя допускает больше поведений, чем наиболее точная

модель; для всех дополнительных поведений начисляются *штрафные очки*, которые затем нормируются и уменьшают *точность* модели.

На основе *метода редукции систем переходов* разработан *метод адаптивной редукции промежуточной системы переходов*, используемой в схеме синтеза целевой сети Петри алгоритмом регионов. В основе метода лежит принцип, согласно которому с помощью сочетания параметров алгоритма редукции можно получить систему переходов, размер которой не будет превышать некоторого заданного значения. Это значение определяется возможностью практического применения *алгоритма регионов* [31] для синтеза конечной сети Петри по промежуточной системе переходов. Известно, что *алгоритм регионов* имеет большую временную сложность от размера входных данных — системы переходов (в общем случае это NP-полная задача).

В соответствии с *методом адаптивной редукции* выполняется построение набора промежуточных систем переходов при различных значениях параметров алгоритмов редукции и выполняется расчет их метрик *точность* и *сложность*. Среди полученного набора выбирается такая система переходов, *сложность* которой не превышает порога практической применимости алгоритма регионов, а *точность* является максимально возможной. К выбранной таким образом системе переходов *алгоритм регионов* применяется однократно. Он синтезирует конечную модель — сеть Петри, метрики *точность* и *сложность* которой будут *согласованы* с соответствующими метриками исходной системы переходов. Таким образом, дорогостоящая операция построения сети Петри выполняется лишь однажды, а качественные характеристики синтезируемой модели определяются уже на этапе перед выполнением этой операции.

В **третьей главе** рассматривается расширение *метода синтеза сетей Петри с помощью алгоритма регионов*. Классический *двухшаговый метод синтеза* [32] включает генерацию системы переходов по журналу событий с последующим синтезом по ней сети Петри. Расширение метода добавляет новый шаг — редукцию сгенерированной системы переходов разработанным алгоритмом редукции с оценкой качественных характеристик редуцированной системы — *перед* выполнением синтеза сети Петри. Синтезируемая сеть Петри обладает такими же качественными характеристиками, что позволяет задавать их еще на этапе редукции. Кроме того, предлагаемое расширение включает модификацию для работы с сетями Петри специального вида — *сетями потоков работ*, а также несколь-

ко подходов для решения проблемы петель в редуцированной системе переходов, которые недопустимы в классическом определении алгоритма регионов.

В главе вводятся новые определения, необходимые для описания указанного расширения алгоритма регионов. Во-первых, дается формальное определение *сетей Петри*, сетей Петри с *разметкой*, *помеченных* сетей Петри, а также *сетей потоков работ*. Определяется понятие *графа достижимости* сети Петри. Затем приводятся необходимые определения из *теории регионов*. Дается собственно определение *региона* на системе переходов как множества состояний, для которых выполняются определенные свойства относительно активностей (пометок).

Далее рассматривается классическая реализация *алгоритма синтеза сети Петри по системе переходов методом регионов*. Этот алгоритм лежит в основе *двухшагового метода синтеза моделей процессов* по журналам событий, предложенного ранее [32]. Ограничением к применению данного метода на практике являются предельные размеры промежуточной системы переходов. Для больших журналов событий, встречающихся на практике, размер такой промежуточной системы переходов будет сопоставим с размером самого журнала событий. Описываемая в этой главе модификация *двухшагового метода синтеза* заключается во включении дополнительного шага — шага редукции с использованием разработанного *адаптивного метода редукции систем переходов*.

Одной из важных особенностей метода частотной редукции является формирование *петель* на графе системы переходов. Петли являются важным механизмом абстракции, позволяющим существенно сокращать размер моделей процессов для больших журналов событий. Одновременно с этим они позволяют таким моделям полностью соответствовать журналам событий, для которых эти модели были построены. Классический алгоритм регионов применим для систем переходов, которые *не содержат петель*. Для преодоления этого ограничения в рамках диссертационной работы было разработано расширение этого алгоритма, которое также описывается в третьей главе.

Четвертая глава посвящена вопросам эффективности программной реализации инструментария для process mining. На основе анализа типичных сценариев использования программных инструментов для решения задач синтеза и анализа моделей процессов выделяется набор требований, предъявляемых к производительному ПО для process mining. Определяются архитектурные принципы, лежащие в основе разработанной высокоутилитарной библиотеки LDOPA для process mining (доступна на сайте проекта <https://prj.xiart.ru/projects/ldopa>).

Обозначаются специальные подходы к управлению памятью и разработке специализированных структур данных, отражающих специфику моделей в данной предметной области.

Отдельное внимание уделяется вопросам применения реляционных СУБД для работы с большими журналами событий на примере исследования поведения реальных информационных систем. Представляется подход к детектированию поведения специального типа (паттернов и анти-паттернов) в ПОИС с помощью РСУБД. Наконец, описывается новый метод представления больших журналов событий с помощью СУБД с мультиперспективным извлечением данных.

Пятая глава описывает методы и инструменты повышения эффективности проведения экспериментов в области process mining. На основе подготовленного стека программного обеспечения приводится описание разработанного графического языка DPMINE моделирования экспериментов; инструмента графического моделирования VTMINER FOR VISIO; набора моделей экспериментов. Описывается подход к автоматизации комплексных экспериментов с помощью сценариев Python и компонента доступа к библиотеке LDOPA, даются примеры сценариев.

В **шестой главе** описываются результаты экспериментального исследования эффективности разработанных алгоритмов и инструментов на основе наборов журналов событий — искусственных и реальных. Предлагается и экспериментально проверяется ряд гипотез, касающихся влияния параметров алгоритма частотной редукции на метрики качества результирующих моделей и время синтеза целевых сетей Петри алгоритмом регионов. Также приводится оценка эффективности реализации алгоритмов по времени и памяти в библиотеке LDOPA в сравнении с аналогичными решениями.

Наконец, в **заключении** приводятся основные результаты работы.

Список литературы

1. *Aalst W. M. P. van der* Process Mining - Discovery, Conformance and Enhancement of Business Processes. — Springer, 2011. — С. I—XVI, 1—352.
2. *IEEE Task Force on Process Mining* Process Mining Manifesto // BPM 2011 Workshops. Т. 99 / под ред. F. Daniel, S. Dustdar, K. Barkaoui. — Springer-Verlag, Berlin, 2011. — С. 169—194. — (Lecture Notes in Business Information Processing).
3. *(OMG) O. M. G.* Business Process Model and Notation (BPMN) Version 2.0: тех. отч. — 01.2011. — URL: <http://taval.de/publications/BPMN20>.
4. Web Services Business Process Execution Language Version 2.0 (OASIS Standard) / A. Alves [и др.]. — 2007. — WS-BPEL TC OASIS, <http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.html>.
5. *Aalst W. M. P. van der, Hofstede A. H. M. ter* YAWL: Yet Another Workflow Language // Inf. Syst. — Oxford, UK, UK, 2005. — ИЮНЬ. — Т. 30, № 4. — С. 245—275. — URL: <http://dx.doi.org/10.1016/j.is.2004.02.002>.
6. *(OMG) O. M. G.* OMG Unified Modeling Language (OMG UML), Superstructure, ver. 2.4.1: тех. отч. — 2011.
7. Using process mining for the analysis of an e-trade system: A case study / A. Mitsyuk [и др.] // Business Informatics. — 2014. — Т. 29, № 3. — С. 15—27.
8. *Rubin V., Lomazova I., Aalst W. M. van der* Agile Development with Software Process Mining // In proceedings: ICSSP 2014, Nanjing, Jiangsu, China. — ACM, 2014. — С. 70—74.
9. Process Mining Can Be Applied to Software Too! / V. Rubin [и др.] // Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. — NY: ACM, 2014.

10. Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital / R. S. Mans [и др.] // Biomedical Engineering Systems and Technologies / под ред. A. Fred, J. Filipe, H. Gamboa. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2009. — С. 425—438.
11. Process Mining in Healthcare: Data Challenges When Answering Frequently Posed Questions. / R. Mans [и др.] // ProHealth/KR4HC. Т. 7738 / под ред. R. Lenz [и др.]. — Springer, 2012. — С. 140—153. — (Lecture Notes in Computer Science). — URL: <http://dblp.uni-trier.de/db/conf/bpm/kr4hc2012.html#MansAVM12>.
12. Big data: The next frontier for innovation, competition, and productivity: тех. отч. / J. Manyika [и др.]. — 2011.
13. *Günther C. W., Van Der Aalst W. M. P.* Fuzzy Mining: Adaptive Process Simplification Based on Multi-perspective Metrics // Proceedings of the 5th International Conference on Business Process Management. — Brisbane, Australia : Springer-Verlag, 2007. — С. 328—343. — (BPM'07). — URL: <http://dl.acm.org/citation.cfm?id=1793114.1793145>.
14. *Badouel E., Bernardinello L., Darondeau P.* The synthesis problem for elementary net systems is NP-complete // Theoretical Computer Science. — 1997. — Окт. — Т. 186. — С. 107—134. — URL: [https://doi.org/10.1016/S0304-3975\(96\)00219-8](https://doi.org/10.1016/S0304-3975(96)00219-8).
15. *Badouel E., Bernardinello L., Darondeau P.* Polynomial algorithms for the synthesis of bounded nets // TAPSOFT '95: Theory and Practice of Software Development: 6th International Joint Conference CAAP/FASE Aarhus, Denmark, May 22–26, 1995 Proceedings / под ред. P. D. Mosses, M. Nielsen, M. I. Schwartzbach. — Berlin, Heidelberg : Springer Berlin Heidelberg, 1995. — С. 364—378. — URL: http://dx.doi.org/10.1007/3-540-59293-8_207.
16. *Buijs J., Dongen B., Aalst W.* On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery // OTM Federated Conferences, 20th International Conference on Cooperative Information Systems (CoopIS 2012). Т. 7565 / под ред. R. Meersman [и др.]. — Springer-Verlag, Berlin, 2012. — С. 305—322. — (Lecture Notes in Computer Science).

17. *Shershakov S. A., Kalenkova A. A., Lomazova I. A.* Transition Systems Reduction: Balancing Between Precision and Simplicity // Transactions on Petri Nets and Other Models of Concurrency XII / под ред. M. Koutny, J. Kleijn, W. Penczek. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2017. — С. 119—139. — URL: https://doi.org/10.1007/978-3-662-55862-1_6.
18. *Shershakov S. A.* DPMine graphical language for automation of experiments in process mining // Automatic Control and Computer Sciences. — 2016. — Т. 50, № 7. — С. 477—485. — URL: <http://dx.doi.org/10.3103/S014641161607018X>.
19. *Shershakov S. A., Rubin V. A.* System Runs Analysis with Process Mining // Modeling and Analysis of Information Systems. — 2015. — Дек. — Т. 22, № 6. — С. 818—833.
20. *Шершаков С. А.* “VTMine for Visio”: инструмент графического моделирования в области process mining // Моделирование и анализ информационных систем (МАИС). — 2020. — Т. 27, № 2. — С. 210—2.
21. *Shershakov S.* Multi-Perspective Process Mining with Embedding Cofigurations into DB-based Event Logs // Communications in Computer and Information Science. — Springer, 2020. — In Press.
22. *Shershakov S.* DPMine/P: modeling and process mining language and ProM plugins // Proceedings of the 9th Central & Eastern European Software Engineering Conference in Russia / под ред. A. N. Terekhov, M. Tsepkov. — ACM New York, NY, USA, 2013. — URL: <http://dl.acm.org/citation.cfm?id=2556622&CFID=415147702&CFTOKEN=35395117>.
23. *Shershakov S. A.* VTMine Framework as Applied to Process Mining Modeling // International Journal of Computer and Communication Engineering. — 2015. — Май. — Т. 4, № 3. — С. 166—179.
24. *Shershakov S.* DPMine/C: C++ Library and Graphical Frontend for DPMine Workflow Language // Proceedings of the 8th Spring/Summer Young Researchers' Colloquium on Software Engineering, SYRCoSE 2014 / под ред. A. Kamkin, A. Petrenko, A. Terekhov. — ISP RAS, 2014. — С. 96—101. — URL: http://syr cose.ispras.ru/2014/files/SYRCoSE2014_Proceedings.pdf.

25. *Shershakov S.* DPMine: modeling and process mining tool // Proceedings of the 7th Spring/Summer Young Researchers' Colloquium on Software Engineering, SYRCoSE 2013. — 2013.
26. *Zubkova N. S., Shershakov S. A.* Method for Building UML Activity Diagrams from Event Logs // Proceedings of the Institute for System Programming of the RAS. — 2019. — T. 31, № 4. — C. 139—150.
27. *Rigin A., Shershakov S.* SQLite RDBMS Extension for Data Indexing Using B-tree Modifications // Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS). — 2019. — T. 31, № 3. — C. 203—216. — URL: <https://ispranproceedings.elpub.ru/jour/article/view/1188/948>.
28. *Shunin T., Zubkova N., Shershakov S.* Neural Approach to the Discovery Problem in Process Mining // Analysis of Images, Social Networks and Texts / под ред. W. M. P. van der Aalst. — Cham : Springer International Publishing, 2018. — C. 261—273.
29. *Davydova K. V., Shershakov S. A.* Mining Hybrid UML Models from Event Logs of SOA Systems // Proceedings of the Institute for System Programming of the RAS / под ред. A. Avetisyan. — 2017. — T. 29, № 4. — C. 155—174.
30. *Davydova K. V., Shershakov S. A.* Mining Hierarchical UML Sequence Diagrams from Event Logs of SOA Systems while Balancing between Abstracted and Detailed Models // Proceedings of the Institute for System Programming of the RAS / под ред. V. P. Ivannikov. — 2016. — T. 28, № 3. — C. 85—102.
31. Deriving Petri Nets from Finite Transition Systems / J. Cortadella [и др.] // IEEE Trans. Comput. — Washington, DC, USA, 1998. — Август. — Т. 47, № 8. — C. 859—882. — URL: <http://dx.doi.org/10.1109/12.707587>.
32. Process mining: a two-step approach to balance between underfitting and overfitting / W. M. P. van der Aalst [и др.] // Software & Systems Modeling. — 2008. — Т. 9, № 1. — C. 87—111. — URL: <http://dx.doi.org/10.1007/s10270-008-0106-z>.