Санкт-Петербургский политехнический университет Петра Великого

Институт компьютерных наук и технологий

На правах рукописи

Воронков Илья Александрович

Разработка и исследование методов обработки больших объемов данных платформ совместной работы (collaboration)

09.06.01 Информатика и вычислительная техника

Код и наименовани

09.06.01_06 Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Код и наименование

НАУЧНЫЙ ДОКЛАД

об основных результатах научно-квалификационной работы (диссертации)

Автор работы: Воронков Илья Александрович
Научный руководитель: доцент, кандидат технических наук, Сараджишвили Сергей Эрикович

Санкт-Петербург – 2021

Научно-квалификационная работа выполнена в ВШ/на кафедре Института ____ федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого»

Директор ВШ/зав. кафедрой:	*– Дробинцев Павел Дмитриевич, кандидат технических наук, доцент*
Научный руководитель:	– Сараджишвили *Сергей Эрикович, кандидат технических наук, доцент*
Рецензент:	*– Яковлев Валерий Петрович, кандидат технических наук, доцент, заведующий кафедрой прикладной математики и информатики Высшей Школы технологии и Энергетики СПБГУПТД*

С научным докладом можно ознакомиться в библиотеке ФГАОУ ВО «Санкт-Петербургский политехнический университет Петра Великого» и на сайте Электронной библиотеки СПбПУ по адресу: http://elib.spbstu.ru

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы

Платформы совместной работы (в дальнейшем в работе это понятие будет встречаться с использованием английского варианта Collaboration) прошли долгий путь от унифицированных программ с простейшим интерфейсом и возможностью обмениваться тестовой информации до современных программно-аппаратных комплексов, объединяющих на единой платформе тысячи людей, устройств. Возросшая сложность порождает необходимость в эффективном управлении, развитии таких комплексов. Анализ научных и инженерных работ (при том последние представлены куда более широко в связи со спецификой тематики) позволяет определить направления, по которым идут современные научно-исследовательские группы, для получения значимых результатов. В ходе работы были отмечено, что первая большая группа исследователей концентрируется на математическом аппарате, считая, что Collaboration platforms являются частными случаями любой другой информационной системы. Публикация иностранных ученых, среди которых можно отметить (Д. Ортега, С. Чой, Б. Джин) полностью абстрагированы от практической составляющей и не имеют последующей инженерной реализации в рамках тестовой эксплуатации выносимых положений. Следующая группа, которую отдельно стоит отметить, в свою очередь описывает и исследует конкретные информационные системы, для которых характерны те или иные приемы, алгоритмы, но нет достоверных сведений о том, что данные техники применимы и способны к миграции на другие комплексы. В частности, исследования геймификации обучения (Д. Сингх) или организация уровней доступа к информации (Д. Диффин) содержат в себе конкретные приемы для Microsoft SharePoint. Третья, на момент написания, наиболее молодая группа исследователей продвигает идею о том, что наиболее эффективным способом получить качественные результаты в понимания изучения Collaboration platforms – это отдать на откуп эти исследования, тесты пользователям этих систем. До недавнего

времени такой подход был ограничен техническими возможностями, уровнем навыков пользователей. Концепция Zero/Low code, начиная с 2015 года начала менять это представление. На сегодняшний день мир насчитывает множество реализаций этого стрима. В данной работе проведены исследования о применимости данного подхода в работе. Данная работа вобрала в себя все вышеперечисленные направления и привнесла новые подходы как в математический аппарат, при этом сохраняя гибкость и возможность перенесения методов с системы на которой проводились испытания на другие подобные комплексы. Таким образом тема работы является актуальной.

Цель и задачи исследования

Целью работы является разработка и применение методов обработки взаимодействий для платформ совместной работы (Collaboration). Задачи, которые были сформулированы в ходе работы:

1) Определить применима ли парадигма, при которой Collaboration выносится в отдельный класс программных комплексов
2) Разработать или адаптировать математический аппарат для выявления закономерностей и исключительных ситуациях в таких платформах.
3) Применить концепцию Zero/Low code для практических задач и сделать предположение о перспективности данного подхода в разрезе темы работы.

Научная новизна

Научная новизна заключается в том, что в ходе работы были представлены и смоделированы взаимодействия, характерные для реальных промышленных систем. При этом:

1) Разработана модель работы промышленного предприятия, которая предоставляет возможность развертывать реальные информационные системы (DMS, SAP и др.), и в то же самое время симулировать работу промышленных установок, путем интеграции через общую шину.

2) Разработана и протестирована методика экспорта/импорта данных из промышленных систем (ETL процессы) для организации тестовой эксплуатации.
3) Сформулированы и подготовлены алгоритмы реализации поиска ключевых событий в системе. Исследована возможность предсказаний появления событий в системе.
4) Проведено исследование о применимости концепции построения исследований взаимодействий в системе путем усовершенствования инструментов для пользователя системы с расширением функционала по организации flow.

Положения выносимые на защиту

1) Набор практик по организации модель предприятия, которая может быть развернута в сжатые сроки на платформе Azure.
2) Методы переноса и портирования данных между тенантами.
3) Практическая реализация и результаты применения алгоритмов детекции, фильтрации, категоризации ключевых событий в системе.
4) Описание применения концепций Zero/Low code и прогнозирование развития данного направления.

Теоретическая и практическая значимость

К теоретической значимости работы следует отнести расширение математического аппарата, который ранее применялся к моделям для поиска и фильтрации событий, а также обоснование применимости концепции Zero\Low code для промышленного применения. Моделирование и описания возможности реализации Collaboration platforms через современные инструменты относится к практической значимости работы. Отдельно стоит отметить описание возможностей расширения Power FX, которое стало одним из первых в России м привело к нескольких выступлениям на сторонних конференциях с докладом.

Апробация работы

Основные положения работы были представлены на следующих конференциях: International Scientific Conference “Telecommunications, Computing and Control” (TELECCON-2019) (г. Санкт-Петербург, 2019); СИСТЕМНЫЙ АНАЛИЗ В ПРОЕКТИРОВАНИИ И УПРАВЛЕНИИ (г. Санкт-Петербург, 2018); II INTERNATIONAL SCIENTIFIC CONFERENCE ON APPLIED PHYSICS, INFORMATION TECHNOLOGIES AND ENGINEERING (г. Красноярск, 2020); 7TH INTERNATIONAL YOUNG SCIENTISTS CONFERENCE ON INFORMATION TECHNOLOGY, TELECOMMUNICATIONS AND CONTROL SYSTEMS, ITTCS 2020 (г. Иннополис, 2020); SYRCoSE-2021 Software Engineering Colloquium (г. Москва 2021)

Публикации

По теме работы опубликованы 9 работ, из которых 3 работы представлены в журналах, индексируемых в Scopus и 3 работы опубликованы в журналы, входящие в перечень ВАК.

Представление научного доклада: основные положения

1) BART алгоритм расширением использования cognitive services.
2) Расширенный алгоритм фильтрации и поиска по ключевым элементам.
3) Реализация концепции Low/Zero code для организации работы (ETL процессы).
4) Применение средств визуализации данных, полученных в ходе тестирования.

СОДЕРЖАНИЕ РАБОТЫ

Во введение приведено обоснование актуальности темы, вычленен объект исследования. Описаны цели и задачи работы. В первой главе можно ознакомиться с подробным описанием существующих практик по организации исследований платформ совместного доступа. Приведен анализ направлений, а также инструментария по работе с объектом исследования. В главе присутствует краткая выжимка для каждого из сегодняшних векторов развития: математические методы над абстрактной системой, применение

ключевых подходов для отдельно выбранной системы, инженерный подход с передачей инструментария в руки конечного пользователя системы. Сделаны предположения о ключевых тенденциях развития направления.

Вторая глава посвящена разработки и усовершенствованию методов математического аппарата. Первый подраздел посвящён модификации BART алгоритма средствами Cognitive azure services. Автор использует CRISP-DM для реализации модели сбора и анализа генерируемой информации.

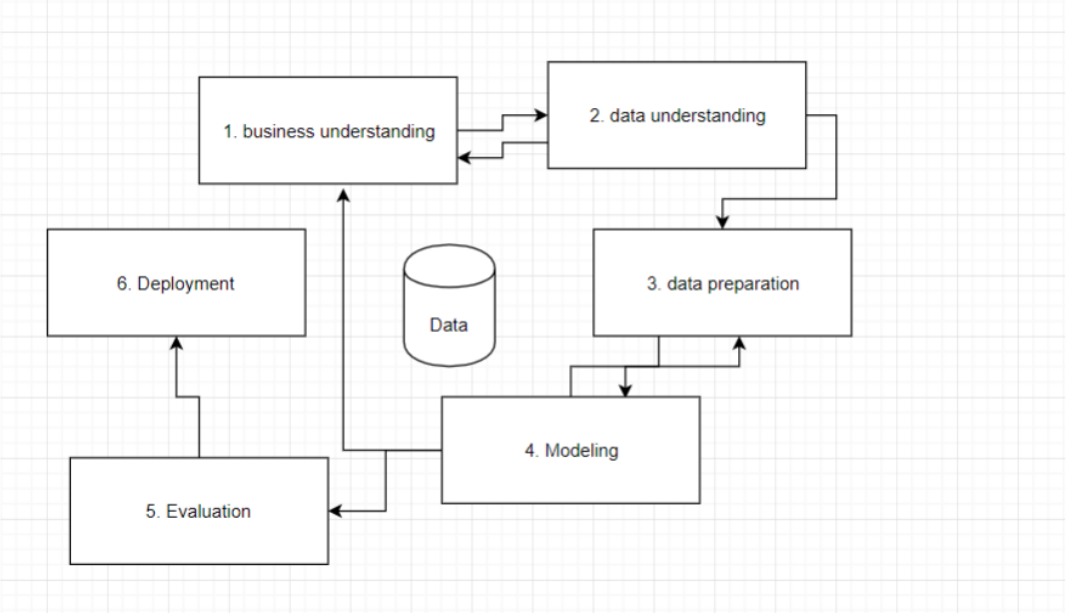

Рис. 1 Схематическая имплементация цикла CRISP-DM

Этот метод представляет собой непрерывный цикл работы с данными, в течение которого выполняется N раз, ошибка в прогнозируемых данных стремится к нулю на основе информации, полученной в предыдущем шаге. Этот метод является одним из наиболее распространенных методов моделирования процесса интеллектуального анализа данных. Далее происходит описание применения регрессионных деревьев для моделирования задачи. Дерево регрессии — это класс моделей регрессии, который позволяет разделить входное пространство фактора переменной на сегменты. Впоследствии вся цепочка регрессионной модели может быть дополнена и будет обрабатываться для каждого из узлов, представляющих функцию регрессии в интуитивно понятной визуальной форме. В работе

представлено определение того, как модель взаимодействия в среде совместной работы относится к выражению процесса в виде дерева:

1) Внутренний узел дерева — это описание правил разбиения пространства, объясняющие переменные. (Пример: элемент системы «сообщение» - неделимая единица для отбора данных для конечного сегмента пользователей. У одного автора может быть много сообщений, а каждое сообщение имеет одного автора в упрощенной модели).

2) Листья деревьев - собственная модель локальной регрессии.

3) Ветка - условия перехода между узлами. Чтобы применить дерево регрессии, в работе описаны связи в системе сотрудничества по принципу разделения и выделения набора информации, не пересекающейся с другими классами системы. Сегментация проводилась последовательно, пока не осталось возможности выделения нового класса. Рассмотрим Байесовский подход к оценке непараметрических функций с использованием деревьев регрессии. Этот алгоритм позволяет обобщить выделенное выше дерево регрессии и временные ряды для итераций с использованием метода CRISP-DM. BART — это комбинация (C&RT) алгоритма и стандартной модели авторегрессионного интегрированного скользящего среднего (ARIMA) и их компоненты (AR,MA Модели SETAR и ASTAR являются линейными моделями однородных моделей (сообщение Inst, office 365 Exchange), которые строят несколько сплайнов адаптивной регрессии (MARS) на основе временных рядов в разовой итерации обрабатывающего комплекса. BART имеет два основных отличия от SETAR и модели ASTAR:

1) Оценки ошибок для моделей BART могут отличаться друг от друга как для каждого узла, так и для каждой итерации цикла.

2) BART характеризуется разрывом между моделями авторегрессии.

Для преобразования модели по временному ряду воспользуемся методом преобразования, где результирующая переменная CDt соответствует сумме предыдущего значения $CDt - 1$ и задержанного значения $CDt - p$ с поправкой на коэффициент тональности β, который, в свою очередь, является

совокупной оценкой для каждого узел системы за время t, полученный с помощью функций текста данных Azure.

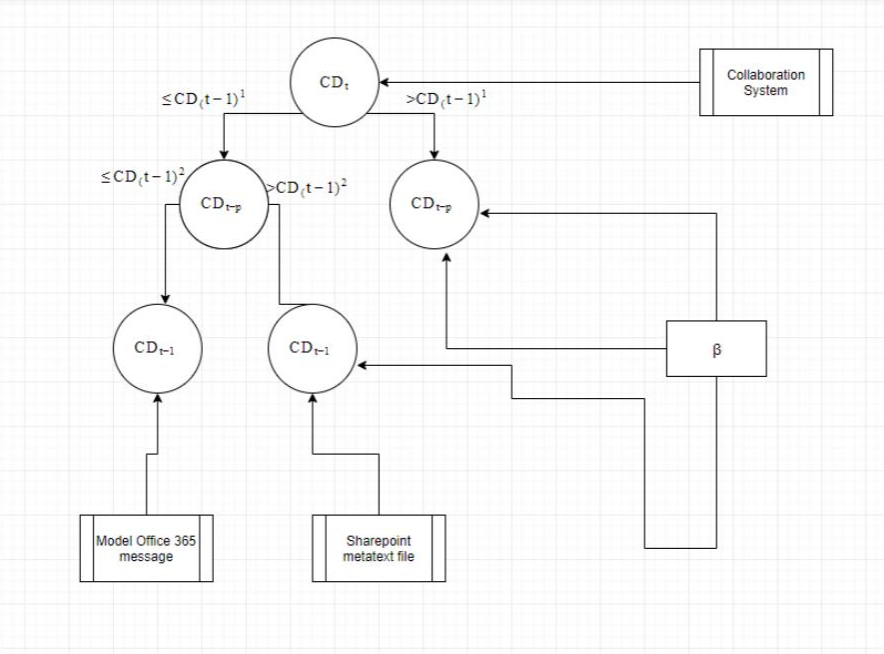

Рис. 2 Схема построения дерева авторегрессии

Большинство алгоритмов используют рекурсивное разделение данных, на которых происходит обучение. BART использует итерационный метод построения. Кроме того, в примерах работы автор добавляет коэффициент тональности β на каждом шаге расчета, который находится в диапазоне значений [0,1] и является накопительным. Этот коэффициент рассчитывается для каждого в отдельности и влияет на результат в каждом узле. 0 - соответствие абсолютного отрицательному комментарию в системе (сообщение в переписке, с выраженным возмущением или недовольство). 1 - положительное значение. Во время моделирования не было встречено ни одного ответа от 0 до 1.

Рассмотрим алгоритм обработки данных:

Шаг первый. Построение дерево регрессии для корневого значения (рассматривается вся система сотрудничества). Построение дерева регрессии начинается с одного значения (корневого узла), которое определяется как медиана (второго квартиля) всего временного ряда. В этом выражении медиана — это сумма действительного числа с вероятностью превышения

произвольного размера, равного 0,5, и неотрицательное значение β, которое стремится к 1 при идеальной системе взаимодействия в системе.

Шаг второй: для каждого необработанного узла мы находим лучший раздел. Сам раздел будет выбран исходя из метрик по заранее определенному правилу.

Механизм выбора правила разделения похож на алгоритм C&RT. Разница заключается в правиле для выбора критерия для оценки прекращения расщепления. Во время тестирования автор использовал информационный критерий для лучшего разделения на основе показателя энтропии, потому что он предпочитает варианты с меньшей сложностью дерева. Этот алгоритм будет определять прирост энтропийной информации. Когда при использовании BART любая вершина в дереве (кроме корня) имеет ровно два дочерних элемента. Финальная цепочка построена из узлов дерева сверху вниз, и происходит информационная оценка предиктора узла, разделение временного рядя в подмножестве экспериментов.

В конце раздела приведено описание эмпирических результатов 30-дневной симуляции. Выдержка из результатов:

1. Лучшее значение для отдельных узлов показало исследование системы чата. Этот узел не критичен для поддержания эффективности модели предприятия, поскольку электронную почту можно дублировать. Этот факт и тот факт, что при расчете был использован когнитивный анализ текстовых сообщений чата, которые могут быть глубоко проанализированный системой на каждой итерации, может объяснить лидерство в использовании этого подхода для текстовых сообщений.

2. Относительно высокое место SharePoint DMS обусловлено формализацией процессов, где в рабочем процессе обработки документов невозможно ввести серьезные нарушения (комментарии). В то же время полезность использования когнитивных сервисов в таких формализованных объектах могут были подвергнуты сомнению.

3. Относительный успех моделирования в модулях SAP также объясняется высокой формализацией и стандартизация процессов. Полученный результат может указывать на несовершенство метода при работе с внешними
Нарушениями и сложность прогнозирования аварийных ситуаций на конвейере.

4. Результаты моделирования конвейера могут указывать на несовершенство применяемого подхода.

Второй раздел второй главы посвящен изучению метода совместной фильтрации. В реальных системах, основанных на использовании совместной эксплуатации, оценка, которая может иметь строго положительный или строго отрицательный числовой коэффициент имеет низкий потенциал для использования. В процессе работы были сформированы Q&A отображение функций, в котором приложение логики использует данные каждого отдельного сообщения в пакете Outlook. В работах идентифицированы такие вопросы по наличию в них ключевых слов. Этот подход описан в работах по поиску ключевых фраз длиной не более трех слов. На основе расширение этой практики, мы можем использовать до 5 слов. В практических исследованиях снижение эффективности метода отмечается увеличением максимального количества слов в ключевых фразах. В начальных итерациях был получен набор обучающих документов, для которых известно, что их образец является наивной байесовской моделью. Эта модель используется в следующих итерациях для поиска ключевых фраз в новых документах. В тесте на первой итерации не было документов, помеченных как начальные для поиска и сопряжения (вопрос-ответ). Это связано с тем, что во время первого теста значение функции вопроса может принимать только два значения (ИСТИНА, ЛОЖЬ). Из этого можно извлечь наиболее часто используемые N-граммы.В последующих итерациях эти комбинации (до 5 слов) помечались как ключевые фразы. В одном и том же электронном письме, может быть, несколько наборов ключевых фраз разного размера, ограниченных только максимальной длиной фразы, указанной в первой

итерации, и общего количества слов в документе. Результирующий набор данных был использован для построения наивной байесовской модели для ключевых фраз. В письме с просьбой (возмущение, функция) вопросы обозначаются как те предложения, которые содержат ключевые фразы. Точно так же ответы определены как те предложения, которые содержат ключевые фразы. Как только вопросы и ответы были определены, необходимо было сопоставить каждый вопрос с соответствующим ответом. Любое электронное письмо может содержать несколько вопросов или несколько ответов. Для формирования такого словаря был применен метод формирования словаря стемминга Портера. Совместная фильтрация — это метод выявления и построения рекомендаций, в которых субъект анализа — это реакция пользователя, которую можно измерить с помощью рейтинговой шкалы. Оценки могут быть различными: явными - теми, которые можно точно измерить количественно, и неявными или субъективными. Количество оценок напрямую влияет на качество рекомендательной системы. С увеличением количество и качество построения системы оценки, точность рекомендаций тоже увеличивается. Сложность — это побочный эффект.

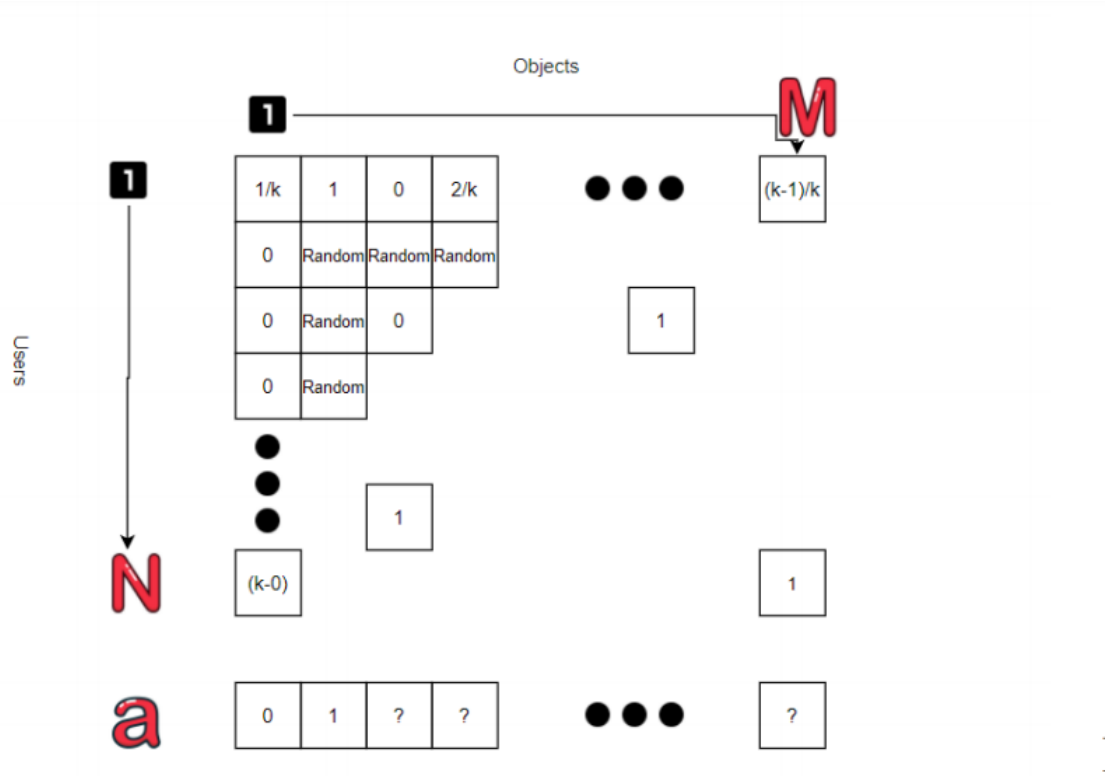

Рис. 3 Схематическая имплементация алгоритма фильтрации

Третья глава посвящена подходу Zero/Low code и его практическому применению. В главе произведено описание языка Power FX. В главе описаны методы организации практически любых взаимодействий с кодовой базой. Реализация ETL:

Рис. 4 Реализация отправки сообщения

Данная глава содержит и стремится доказать существующую тенденцию по интеграции концепции, которая вынесена в названии и средств визуализации информации. Подробно описана концепция реализации «мгновенного» создания ETL-процесса и встраивания этого потока в систему отображения Dashboards. Произведено исследование Power Platform, как наиболее приспособленной реализации для интеграции в модель предприятия через платформу Azure. Показана возможность быстрой развертки, тестирования и администрирования подобного сервиса. Предложен прогноз развития данного направления и постепенная смена фокуса для научной и преподавательской деятельности. Переход от процесса «Мы обучаем людей специализированным языкам для общения с машиной» к идее «Мы учим компьютер понимать естественный язык человека». При этом концепция low/zero code является промежуточным звеном. Преимущество данной концепции заключено в том, что наши существующие программно-аппаратные комплексы позволяют применять данный подход на практике уже в настоящее время, а количество специалистов необходимых для разработки и поддержки подобных решений строго равно существующему количеству специалистов в той или иной отрасли (пример: работник нефтяной отрасли лучше всего осведомлен о том какой набор приложений ему необходим для ежедневной деятельности. Начиная с настоящего момента в его распоряжении есть весь необходимый пакет инструментов, которым он может воспользоваться при минимальной компьютерной грамотности).

Результаты и их обсуждение

К результатам выполненной работы можно отнести появление и расширение транслируемых методов и алгоритмов для Collaboration. Данные методы уже проходят тестирование на предприятиях. Полученные экспериментально данные могут служить доказательством правильности вычленения платформ совместной работы в отдельную категорию информационных систем.

Концепция передачи основного направления формирования подходов от научного сообщества к инженерному так же считаю показательным результатом в том числе и моей деятельности статьи и выступления на конференциях. Отдельно отметим определенный интерес к двум основным парадигмам работы. Методы по категоризации и детекции ключевых событий в информационный системах могут быть применены в системах с отсутствием общей шины обмена информациями между подсистемами. Особый интерес вызывает возможность проведения исследований с применением данного подхода в сложных системах, где часть модерации отдана на откуп человека. Существует предположение, что такой полуавтоматический подход несет в себе потенциально высокие риски для формирования ошибок, которые в работе названы ключевыми событиями в системе.

Заключение

Данная работа может быть использована как основа для масштабных исследований сразу по двум направлениям: математический подход и инженерный. Следует отметить особенность того, что два эти направления могут быть объектом/субъектом исследования одновременно, выступая поочередно, сначала ядром (базисом), в то время как второй компонент является лишь инструментарием, необходимым для углубления понимания работы ядра. Необходимо выбрать путь о разделении данной работы на две составляющие или попытки равноправного применения обоих подходов. Следует расширить команду исследования путем итеративного подхода к созданию и поддержке автономных приложений на базе существующей модели.

Список работ, опубликованных по теме научно-квалификационной работы (диссертации)

Публикации в изданиях, рецензируемых ВАК

1. Voronkov I. A., Saradgishvili S. E Power Fx: Low-code Language for Collaboration Tools Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS). № 33(3). С. 101-108. doi: 10.15514/ISPRAS-2021-33(3)-8.

2. Русcков О. В., Воронков И. А., Сараджишвили С. Э. Методический подход к прогнозированию неравномерного временного ряда // Современная наука: актуальные проблемы теории и практики. 2021. №4. С. 142-147. doi: 10.37882/2223-2966.2021.04.32.

3. Воронков И. А., Сараджишвили С. Э. Использования Power FX для работы в Collaboration platforms // Современная наука: актуальные проблемы теории и практики. 2021. №7. С. 61-66. doi: 10.37882/2223–2966.2021.07.08.

Публикации в других изданиях

1. Saradgishvili S. E., Voronkov I. A Usage of a BART algorithm and cognitive services to research collaboration platforms // Smart Innovation, Systems and Technologies. 2021. №220. С. 267–276. doi: 10.1007/978-981-33-6632-9_23.

2. Voronkov I. A., Saradgishvili S.E. Usage of collaborative filtering in sharing platforms // Journal of Physics: Conference Series. 2020. №1679(3). doi: 10.1088/1742-6596/1679/3/032091.

3. Voronkov I. A., Saradgishvili S.E. Usage of a BART algorithm and cognitive services to research collaboration platforms // Journal of Physics: Conference Series. 2020. № 1694(1). doi: 10.1088/1742-6596/1694/1/012028.

4. Воронков И. А., Сараджишвили С. Э. Microsoft Power BI в цифровой обработке многомерных сигналов: учебное пособие. Санкт-Петербург: Издательство: Федеральное государственное автономное образовательное учреждение высшего образования "Санкт-Петербургский политехнический университет Петра Великого" (Санкт-Петербург), 2019. doi: 10.18720/SPBPU/2/i19-92.

5. Сараджишвили С. Э., Воронков И. А. Проблемы в обработке больших объемов данных для платформ collaboration // СИСТЕМНЫЙ АНАЛИЗ В ПРОЕКТИРОВАНИИ И УПРАВЛЕНИИ Сборник научных трудов XXII

Международной научно-практической конференции. 2018. Санкт-Петербург: Федеральное государственное автономное образовательное учреждение высшего образования "Санкт-Петербургский политехнический университет Петра Великого" (Санкт-Петербург), 2018.

6. Воронков И. А. Реализация механизмов визуализации для управления объектами электросети // АВТОМАТИЗАЦИЯ И IT В ЭНЕРГЕТИКЕ. 2017. №2 (91). С. 22-23.

Аспирант ______ФИО

(подпись)