На правах рукописи

Нгуен Туан Ань

МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДАННЫХ В ПРОЦЕССАХ ПРЕДОСТАВЛЕНИЯ ПЕРСОНИФИЦИРОВАННЫХ УСЛУГ ТЕЛЕКОММУНИКАЦИОННЫМИ ПРЕДПРИЯТИЯМИ

05.13.01 — Системный анализ, управление и обработка информации (информационные технологии и промышленность)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Волгоград – 2017

Работа выполнена на кафедре «Системы автоматизированного проектирования и поискового конструирования» в федеральном государственном бюджетном образовательном учреждении высшего образования «Волгоградский государственный технический университет» (ФГБОУ ВО «ВолгГТУ»).

Научный руководитель	доктор технических наук, **Щербаков Максим Владимирович.**
Официальные оппоненты:	**Финогеев Алексей Германович,** доктор технических наук, профессор, ГБОУ ВО «Пензенский государственный университет», кафедра «Системы автоматизированного проектирования», профессор; **Барабанова Елизавета Александровна**, кандидат технических наук, доцент, ФГБОУ ВО «Астраханский государственный технический университет», кафедра «Связь», доцент.
Ведущая организация	ФГБОУ ВО «Воронежский государственный технический университет», г. Воронеж.

Защита диссертации состоится 21 декабря 2017 года в 16 часов 00 минут на заседании диссертационного совета Д 212.028.08, созданного на базе Волгоградского государственного технического университета, по адресу: 400005, г. Волгоград, проспект им. В.И. Ленина, 28, ауд. 237.

С диссертацией можно ознакомиться в библиотеке и на сайте Волгоградского государственного технического университета и на сайте http://www.vstu.ru

Автореферат разослан « »_____ 2017 г.

Ученый секретарь
диссертационного совета Орлова Юлия Александровна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. Предоставление качественных телекоммуникационных услуг (ТКУ) в условиях (1) возрастающей интенсивности пользования услугами, особенно мобильным интернетом (2), возрастающих рисков, связанных с информационной безопасностью (3), возникающей потребности персонификации услуг для нужд конкретных пользователей, является актуальной стратегической задачей современного телекоммуникационного предприятия.

Для обеспечения бесперебойности и высокого качества оказания ТКУ в телекоммуникационных компаниях (ТКП) разрабатываются корпоративные стандарты, регламентирующие создание, реализацию и управление основными и вспомогательными бизнес–процессами ТКП. Международной практикой является формализация бизнес–процессов в формате расширенной карты процессов деятельности телекоммуникационной компании eTOM (enhanced Telecom Operations Map), которая является частью программы международного консорциума TM Forum NGOSS (Next Generation Operations Systems and Software). В рамках этого аппарата обозначены процессы и системы поддержки управления предоставлением услуг, для которых предлагаются и внедряются корпоративные информационные системы и системы автоматизации. Задачи, связанные с анализом данных, как правило, решаются высококвалифицированными специалистами с использованием инструментов бизнес–аналитики (Business Intelligence). На основании результатов анализа принимаются управленческие решения, в том числе и по модификациям бизнес–процессов предоставления услуг.

В основных и вспомогательных процессах следует выделить инвариантные операции в процессах, связанных с персонификацией предоставления услуг, т.е. мероприятий, направленных на максимальное удовлетворение потребностей в коммуникации конкретного абонента. Это новый подход для ТКП, направленный на поиск оптимального плана предоставления услуг для пользователя на основании активностей и предпочтений последнего (стратегия разработки персонифицированных услуг для повышения лояльности). Для перехода к оказанию персонифицированных услуг следует изучить поведение абонента, сформировать предложение по предоставлению услуг в соответствии со стратегией развития компании и ожиданиями по достижению поставленных целей, реализовать услугу. Это требует поиска новых методов обработки имеющихся данных. В этой связи возникает актуальная научная задача, связанная с совершенствованием методов обработки данных в процессах предоставления и управления персонифицированными услугами для повышения эффективности обслуживания абонентов телекоммуникационного предприятия.

Степень разработанности темы. Изучением вопросов совершенствования методов управления и обработки данных в телекоммуникационных предприятиях занимались отечественные и зарубежные ученые: Салютина Т. Ю., Ромашин А. А, Вейнберг Р. Р.,

Корольков В. Ф., Вагель Е. В., Леонтьев Е. Д., Квятковская И. Ю., Фам Куанг Хиеп, Хилас С., (S. Hilas), Сенака В. (Senaka W.), Батпития Д. А. (Buthpitiya, D.), Масторокостас П. (P. Mastorocostas), Олайвиола В. (Olayiwola W.), Луо Юу (Luo Ye), Киуру К. (Cai Qiuru). Следует отметить, что в исследованиях проблема автоматизации и применения методов обработки информации при реализации персонифицированных услуг остается актуальной.

Объект исследования: процессы предоставления персонифицированных услуг в ТКП.

Предмет исследования: модели и методы обработки данных в процессах предоставления персонифицированных услуг в ТКП.

Цель работы заключается в разработке моделей и методов обработки данных в ТКП для повышения качества предоставления персонифицированных ТКУ. Качество оценивается критериями, включающими показатели удовлетворенности абонентов и технические показатели.

Для достижения поставленной цели были сформулированы следующие задачи:

1. Выполнить системный анализ процессов предоставления ТКУ на телекоммуникационном предприятии, выделить операции предоставления персонифицированных ТКУ.
2. Разработать модель абонента телекоммуникационного предприятия, учитывающую поведение абонента при пользовании услугами.
3. Разработать методы интеллектуальной обработки информации в процессах предоставления персонифицированных услуг ТКП.
4. Выполнить проектирование и разработать программный продукт, реализующий предложенные методы интеллектуальной обработки информации.
5. Провести испытания предлагаемых методов, программного продукта и обосновать эффективность предлагаемых положений.

Методология и методы диссертационного исследования: системный анализ, методы поддержки принятия решений, теория вероятности и математическая статистика, методы машинного обучения и интеллектуальной обработки данных.

Научная новизна заключается во впервые предложенной совокупности моделей и методов обработки информации при управлении предоставлением персонифицированных услуг телекоммуникационным предприятием, включающей в себя:

1. новую модель абонента телекоммуникационного предприятия, отличающуюся встроенными показателями анализа динамики пользования абонентом различными услугами ТКП;
2. новый метод кластеризации абонентов ТКП, отличающийся обработкой статических и динамических (информации о поведении абонентов) данных алгоритмами кластеризации и позволяющий определять группы схожих абонентов по их поведению;

3. новый метод идентификации изменений в поведении абонентов, который отличается тем, что построен на формальных методах кластеризации поведения абонентов и позволяет выявлять изменения в поведении без предварительной разметки выборки данных;

4. новый метод идентификации нетипичной активности абонентов, отличающийся проактивным способом обнаружения активностей, характеризующих проявление мошеннических действий.

Теоретическая значимость работы состоит в разработке моделей и методов обработки информации, позволяющих повысить качество предоставления персонифицированных ТКУ за счет поддержки принятия решений в бизнес–процессах ТКП. Содержащиеся в диссертационной работе анализ, выводы и предложения могут быть также использованы для управления и обработки данных в телекоммуникационных предприятиях.

Практическая значимость работы состоит в разработанном программном обеспечении, реализующем предложенные методы. Разработаны «Программное обеспечение обнаружения мошенничества в телекоммуникационных предприятиях» (свидетельство о государственной регистрации программы для ЭВМ № 2017611602 от 7 февраля 2017 г.) и «Распределенная система слияния и предобработки разнородных данных с разных источников» (свидетельство о государственной регистрации программы для ЭВМ № 2017660307 от 20 сентября 2017 г.).

Программы прошли апробацию в компании VNPT–Media (Вьетнам) (имеется акт внедрения). Результаты диссертационной работы использованы при реализации гранта Президента МД–6964.2016.9 (руководитель Щербаков М. В.).

Положения, выносимые на защиту.

1. Модель абонента телекоммуникационного предприятия, позволяющая формировать и реализовывать персонифицированные услуги для пользователей ТКП.
2. Метод кластеризации абонентов ТКП, позволяющий выделять группы абонентов по схожему поведению при потреблении телекоммуникационных услуг.
3. Метод идентификации изменений в поведении абонентов на основе алгоритмов кластеризации данных без предварительной разметки, позволяющий в автоматическом режиме реагировать на изменения при потреблении телекоммуникационных услуг.
4. Метод проактивной идентификации нетипичной активности абонентов при проявлении мошеннических действий.
5. Программное обеспечение, реализующее предложенные подходы в распределенной среде вычислений.

Степень достоверности и обоснованности полученных результатов исследования основывается на корректном применении методов системного анализа, методов поддержки принятия решений, методов машинного обучения и интеллектуальной обработки данных.

Достоверность полученных результатов подтверждается проведенными экспериментальными исследованиями на открытых источниках данных, а также внедрением и использованием рекомендаций, содержащихся в диссертационном исследовании, в телекоммуникационной компании, что подтверждается соответствующим актом.

Апробация результатов работы. Основные положения исследования докладывались и обсуждались на следующих научных конференциях: Distributed Computer and Communication Networks : 20th International Conference (DCCN 2017) (Moscow, Russia, September 25–29, 2017), 7th International Conference on Information, Intelligence, Systems & Applications (IISA) (Greece, 13–15 July 2016), XLV междунар. науч.–практ. конф. (г. Новосибирск, 26 мая, 14 июня 2016 г.), XX Региональная конференция молодых исследователей Волгоградской области (г. Волгоград, 8–11 дек. 2015 г.), 6th International Conference on Information, Intelligence, Systems and Applications. IISA2015 (Corfu, Greece, July 6–8, 2015), Интеллектуальный потенциал XXI века ‘2015 : матер. Междунар. науч.–практ. молодёжной Интернет–конф. (Украина, 10–22 нояб. 2015 г.), Юность и Знания – Гарантия Успеха – 2015 : сб. науч. тр. 2–й Междунар. науч.–практ. конф. (1–2 окт. 2015 г.), Мир науки и инноваций. – 2015.

Личный вклад автора. В диссертации представлены результаты исследований, выполненных самим автором. Личный вклад автора состоит в постановке задач исследования, разработке теоретических и прикладных методов их решения, в обработке, анализе, обобщении полученных результатов и формулировке выводов. В публикациях с соавторами авторский вклад распределяется пропорционально.

По теме диссертации издано 14 печатных работ, в том числе 4 статьи в изданиях, рекомендованных ВАК, 3 работы в зарубежных изданиях, индексируемых в базах научного цитирования Scopus. По результатам работы созданы 2 программных продукта, которые получили Свидетельства о государственной регистрации.

Структура и объем диссертации. Диссертационная работа состоит из введения, четырех глав, заключения, а также библиографического списка из 114 наименований и 2 приложений. Общий объем работы – 141 страница, в том числе 49 рисунков и 11 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы, сформулирована цель и задачи исследования, выбраны методы исследования, отмечены научная новизна и практическая значимость работы, приведены основные положения, выносимые на защиту, сведения об апробации работы.

В первой главе дана характеристика функционированию современного телекоммуникационного предприятия. Телекоммуникационное предприятие (ТКП) – это предприятие, которое осуществляет свою деятельность по

оказанию телекоммуникационных услуг (ТКУ) в соответствии с выданной лицензией.

Определены ключевые показатели эффективности функционирования предприятия, в частности показатели, связанные с лояльностью клиентов: число жалоб от пользователей по техническим и организационным моментам, степень удовлетворенности абонентов качеством обслуживания и техническими характеристиками услуг, потерь от несанкционированного доступа к ТКУ. Использована иерархическая структура для представления комплексного показателя качества ТКУ, разделенного на 3 уровня.

Проведен анализ функционирования телекоммуникационного предприятия (на примере Вьетнамской компании VNPT-Media), определены группы основных и вспомогательных бизнес-процессов ТКП в соответствии с картой процессов деятельности телекоммуникационной компании.

Выделены и выполнен системный анализ бизнес-процессов телекоммуникационного предприятия, связанные с формированием и предоставлением персонифицированных услуг, а также процессы для которых важно понимание поведения абонентов. Анализ осуществлялся на основе расширенной карты деятельности ТКП в сегменте B2C (business to customer), при этом рассмотрены сквозные (горизонтальные) группировки бизнес-процессов, объединенные по функциональному принципу. Были выделены следующие процессы:

1. C1.2. Планирование портфеля продуктов и предложений (группа C1. Управление маркетингом и предложением, группировка C. «стратегия, инфраструктура и продукт»).
2. O1.9. Удержание клиентов и повышение лояльности (группа O1. Управление взаимоотношениями с клиентами, группировка O. Операции)
3. O2.5. Тарификация услуг и отдельных событий (группа O2. Сетевая эксплуатация и управление услугами, группировка O. Операции)
4. Y4.3. Управление мошенничеством (группа Y4. Управление рисками организации, группировка Y. Управление организацией)

В рамках последней группы бизнес процессов проанализирована проблема телекоммуникационного мошенничества (ТКМ).

Выделены операции, направленные на реализацию персонифицированных услуг, которые могут быть рассмотрены как инвариантные к бизнес-процессам: формирование профиля абонента, кластеризации абонентов (в том числе и по их поведению), идентификация (классификация) поведения и прогнозирование поведения (потребления услуг) абонентов.

Во второй главе выполнен обзор современного состояния в области автоматизации телекоммуникационных предприятий: обзор систем автоматизации, подходов к поддержке принятия решений в частности при реализации персонифицированных услуг и методов обработки информации, в том числе методов интеллектуального анализа данных, машинного обучения и искусственного интеллекта.

Рассмотрены методы оценки качества предоставления услуг ТКП: Метод SERVQUAL, Метод SERVPERF, Метод ранжирования и Метод парных сравнений. Выполнен анализ методов обработки информации в системах автоматизации процессов ТКП и информационного обеспечения систем автоматизации ТКП, структуры данных о пользовании услугами.

Рассмотрена базовая информационная структура: подробные записи о вызовах абонентов (ПЗВ) (CallDetailRecords, CDR), характеризуемая следующим набором атрибутов:

$$\mathrm{CDR} = \langle \mathrm{User\_id}, \mathrm{CallNum}, \mathrm{RcNum}, \mathrm{Time\_stamp}, \mathrm{Duration}, \mathrm{Code}, \mathrm{ChCos}, \mathrm{Call\_PR} \rangle, \quad (2)$$

где: User_id - идентификатор абонента; CallNum- номер телефона вызывающего абонента; RcNum - номер телефона адресата (вызываемого абонента); Time_stamp - время начала разговора; Duration - продолжительность соединения при разговоре (в минутах); Code - код страны, в которую абонент звонил; ChCos - цена звонка (рублей/минуту); Call_PR – стоимость звонка, заносимая в биллинговую систему ТКК. Таким образом, каждая ПЗВ – это данные о событии, генерируемые в момент совершения пользователем вызова или использования другой услуги.

Пример фрагмента сгенерированных ПЗВ представлен в таблице 1.

Таблица 1 - Пример ПЗВ (фрагмент базы данных)

	id	**Start_time**	**duration**	**area_code**	**charges_cost**	**call_price**
0	655	5/30/2016 17:11	3.09	KZ	20.0	61.800
1	173	5/30/2016 17:16	0.61	UK	34.5	21.045
	…	…	…	…	…	…

Названия столбцов: «id» –идентификатор абонента, «start_time» -- время начала оказания услуги, «duration» -- продолжительность (в мин.), «area_code» -- код области, «charges_cost» -- стоимости услуги, «call_price» -- общая стоимость.

Следует отметить, что для реализации персонифицированных услуг необходимо учитывать как статические данные так и динамические данные. Для этого была предложена следующая модель абонента, объединяющая формализованное представление $a^{(s)}$, формализованное представление $a^{(d)}$ и операции над признаками.

$$a = \langle id, [t_s, t_k], \{x_j\}_{i=1}^{n_x}, cdr, D_x, D_y, \{\alpha_k\}_{k=1}^{n_\alpha} \rangle , \quad (1)$$

где id – идентификатор абонента, $[t_s, t_k]$ – временной интервал наблюдения, $\{x_j\}_{i=1}^{n_x}$ – множество признаков абонентов, cdr – схема записей о вызовах абонента (2.2), D_x, D_y -- матрицы значений признаков и $\{\alpha_k\}_{k=1}^{n_\alpha}$ – множество операций (алгоритмов) преобразования значений D_x, D_y.

Схема cdr представлена следующей формализацией

$$cdr = \langle id, ts, type, dr, \{y_j\}_{i=1}^{n_y} \rangle \quad (2)$$

где id –идентификатор абонента, ts - время начала вызова, $type$ – признак, характеризующий тип вызова (голосовой вызов, смс, передача данных и др.), $type \in \{y_j\}_{i=1}^{n_y}$, dr - продолжительность вызова при разговоре, $\{y_j\}_{i=1}^{n_y}$ – множество атрибутов вызова, которые могут включать различные признаки.

Так как модель (1) включает как статические признаки (не меняющиеся во времени) $\{x_j\}_{i=1}^{n_x}$ так и динамические признаки (меняющиеся во времени) cdr, то результатом выполнения алгоритмов из $\{\alpha_k\}_{k=1}^{n_\alpha}$ является матрица D_z, включающая или исходные признаки из матриц или новые признаки, являющиеся преобразованием исходных.

Следовательно, можно определить следующие операции преобразования признаков (представлены выборочно):

– $\alpha_\sigma(\{z\}, [t_s, t_k])$ – операция выбора значений подмножества признаков на интервале $[t_s, t_k]$ из всей выборки данных $D_x \times D_y$ в соответствии с условием σ;

– $\alpha_\pi(\{z\}, [t_s, t_k])$ – операция формирования подмножества признаков из всего множества признаков $\{x_j\}_{i=1}^{n_x} \cup \{\langle cdr \rangle\}$ в соответствии с условием π;

– $\alpha_{cnt}(z, [t_s, t_k])$ – операция вычисления числа событий изменения значеия признака z на интервале $[t_s, t_k]$;

– $\alpha_{delay}(z, t_1, t_2)$ – операция вычисления значения признака z как значение признака в предыдущий момент времени т.е. $z_{t_1} = z_{t_2}$;

– $\alpha_{sum}(z, [t_s, t_k])$ – операция вычисления суммы значений признака z на интервале $[t_s, t_k]$, т.е. $z_{sum} = \sum_{t \in [t_s, t_k]} z_t$;

Рассмотрены методы интеллектуальной обработки данных, использованные для автоматизации процессов ТКП. Особое внимание уделяется методам определения нетипичного поведения абонентов, методам выявления мошеннического поведения злоумышленников.

Проведен анализ современных технологий распределенной обработки большого объема информации в ТКП и оценены возможности применения технологий при реализации задач предоставления персонифицированных услуг. Рассмотрена концепция λ (лямбда) архитектуры распределенных систем. λ-архитектура обеспечивает доступ к «большим и быстрым» данным и представляет собой универсальную, масштабируемую и отказоустойчивую архитектуру систему обработки данных. Эта архитектура появилась и развивалась в распределенных системах обработки данных на основе Backtype и Twitter.

Третья глава содержит основные теоретические положения, обладающие научной новизной и выносимые на защиту. Перед построением методов, были сформированы два типа профиля пользователей на основе формализации (1).

Метод кластеризации абонентов ТКП по динамике их поведения. Целью данного метода является разбиение множества абонентов $\{a\}$, на несколько кластеров, число которых заранее задано. Решение задачи осуществлялась с

использованием формализации (1) и в соответствии с постановкой задачи кластеризации.

Рассмотрим предлагаемый метод.

Шаг 1. Задать значения гиперпараметров моделей кластеризации: $[t_s, t_k]$ –интервал наблюдения, где t_s – временная метка начала наблюдения, t_k – временная метка окончания наблюдения, h – длина короткого интервала наблюдения на которые равномерно разбивается интервал $[t_s, t_k]$, b – число разбиений на коротком интервале наблюдения, k - число кластеров.

Таким образом, на интервале наблюдения формируются короткие интервалы наблюдения: $[t_s, t_s + h)$, $[t_s + h, t_s + 2h)$, … $[t_s + lh, t_k]$, где l – целое число, которые в свою очередь разбиваются на b равномерных интервалов. Значения гиперпараметров могут быть заданы в виде одного значения, а также в виде сетки значений с шагом.

Шаг 2. Сформировать профили пользователей применяя алгоритмы $\alpha_\sigma(\{z\}, [t_s, t_k])$, $\alpha_\pi(\{z\}, [t_s, t_k])$, $\alpha_{sum}(z, [t_s, t_k])$ из (2) в соответствии со значением гиперпараметров $[t_s, t_k], h, b$. Отметим, что полученная матрица D_z , содержит строки, каждая из которых представляет собой входной вектор для алгоритмов кластеризации для каждого короткого интервала наблюдения.

Шаг 3. Построить ансамбль моделей кластеризации: *k*-средних, MiniBatch *k*-средних, MeanShift (для последнего задание числа *k* не требуется).

Шаг 4. Выполнить поиск наилучших вариантов моделей кластеризации используя несколько запусков и оценку качества кластеризации. Оценка качества каждой модели осуществляется в соответствии с коэффициентом:$s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))}$, который можно представить в виде:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & if \;\; a(i) < b(i) \\ 0, & if \;\; a(i) = b(i) \\ b(i)/a(i) - 1, & if \;\; a(i) > b(i) \end{cases} \qquad (3)$$

Среднее значение $s(i)$ по всем данным кластера является мерой того, насколько «сильно» сгруппированы данные в кластере.

Шаг 5. Определить центроиды кластеров и сопоставить идентификатору абонента метку кластера.

Следует обратить внимание на отличительные признаки предлагаемого метода от имеющихся подходов. На шаге 2 кроме информации о динамики звонков в определенный период использовались статистики для включения информации об изменении динамики потребления услуг в определённый периодом времени. На шаге 4 исходя их выполненного анализа принято решение об использовании ансамбля моделей. Поиск оптимальной конфигурации алгоритма кластеризации осуществляется по двухконтурной схеме: на первом контуре изменяется значение числа кластеров, на втором варьируются параметры τ, h.

Для нового абонента (данные о котором не участвовали в настройке алгоритма кластеризации) предлагается следующий алгоритм определения метки кластера.

Шаг 1. Для каждого кластера загрузить значения центроидов кластера (результат работы алгоритма кластеризации абонентов ТКП)

Шаг 2. Преобразовать данные о классифицируемом абоненте к структуре (п аналогии с шагом 2 метода кластеризации.

Шаг 3. Рассчитать расстояния от профиля классифицируемого абонента до центроидов каждого кластера: $\left\{d_1^{(*)}, d_2^{(*)}, \dots d_k^{(*)}\right\}$.

Шаг 4. Определить метку кластера как $l' = agrmin\left\{d_1^{(*)}, d_2^{(*)}, \dots d_k^{(*)}\right\}$, $l' \in [1, \dots, k]$.

Метод идентификации изменений в поведении абонентов на основе кластеризации. Целью данного метода является идентификация изменения в поведении абонентов $\{a\}$ на заданном заранее горизонте рассмотрения. В основу метода положена гипотеза о том, что если профиль абонента отличается от типового профиля абонентов аналогичного кластера, то считаем, что возникла ситуация, при которой можно говорить о факте изменения поведения.

Метод включает две стадии: стадия кластеризации и определения меток кластеров; стадия идентификации отклонения.

Стадия кластеризации включает следующие шаги.

Шаг 1. Задать значения гиперпараметров моделей кластеризации: $[t_s, t_k]$ –интервал наблюдения, где t_s – временная метка начала наблюдения, t_k – временная метка окончания наблюдения; h – длина короткого интервала наблюдения на которые равномерно разбивается интервал $[t_s, t_k]$; b – число разбиений на коротком интервале наблюдения; k – задать число кластеров; $cx = 0$.

Следует обратить внимание на задание параметра числа кластеров k. Считаем, что число кластеров определяется исходя из формулы $k = int\left(\frac{n}{q}\right)$, где $int(\cdot)$ – операция округления числа до целого, q – эмпирически подбираемый коэффициент, при этом считаем, что $q < n$. Малое число кластеров может привести к ситуации, когда практически все профили оказываются в одном кластере. Это объясняется тем, что время пользования услугами (вызовы) много меньше времени, при котором абонент услугами не пользуется.

Шаг 2. Сформировать профили пользователей, применяя алгоритмы $\alpha_\sigma(\{z\}, [t_s, t_k])$, $\alpha_\pi(\{z\}, [t_s, t_k])$, $\alpha_{sum}(z, [t_s, t_k])$ из (2.5), в соответствии со значением гиперпараметров $[t_s, t_k], h, b$.

Шаг 3. Выделить профили абонентов, для которых значения рассматриваемых признаков равно нулю, и присвоить идентификаторам пользователя признак кластера 0.

Шаг 4. Исключить из выборки данных абонентов с идентификатором 0.

Шаг 5. Выполнить шаги 3–5 метода кластеризации, описанного в предыдущем пункте.

Стадия идентификации отклонений от ожидаемого поведения.

Шаг 1. Получить фактические значения для анализируемого абонента и преобразовать их в соответствии с параметрами h, b, для которых рассчитаны модели.

Шаг 2. Получить метку кластера для анализируемого абонента id_c.

Шаг 3. Для всех абонентов с указанной меткой кластера id_c выбрать все профили, входящие в кластер id_c.

Шаг 4. Для кластера id_c на основе профилей вычислить нижнюю границу в рассматриваемые моменты времени, верхнюю границу и среднее значение профилей, входящих в кластер.

Шаг 5. Выполнить проверку: если хотя бы одно текущее значение в профиле меньше минимальной границы, т. е. $\forall p = 1, \dots b,\ \exists x_p^{(id_c)}[p] < x_{low}^{(id_c)}[p]$, или больше максимальной границы, т. е. $\forall p = 1, \dots b,\ \exists x_p^{(id_c)}[p] > x_{sup}^{(id_c)}[p]$,

то считать, что поведение пользователя изменилось, и положить $n_{cx} = n_{cx} + 1$.

Для оценки качества идентификации может быть использована мера, рассчитываемая по формуле:

$$ratio^{(id_c)} = \frac{n_{cx}}{h \cdot b}, \quad (4)$$

т.е. если $ratio^{(id_c)} = 0$, то не наблюдалось событий изменения поведения на коротком интервале наблюдения.

Метод выявления изменения поведения пользователей услуг на основе супервизорного подхода для выявления телекоммуникационного мошенничества. Рассмотрим предлагаемую методику, включающую в себя два основных этапа: построение модели и ее применение.

На этапе построения модели выполняются следующие действия.

Шаг 1. Выполнить сбор данных и их структурирование в формате, описанном выше.

Шаг 2. Выполнить предобработку исходных данных с целью обнаружения пропущенных значений для атрибутов ‘duration’, ‘call_price’и замены их средними значениями.

Шаг 3. Сформировать выборки данных для обучения модели.

(i) Формирование первой выборки данных (маркировка profile1_ds), включающей данные в соответствии с «профилем 1», которые размечены для каждого пользователя: 0 – обычный пользователь, 1 – мошенник.

(ii) Формирование второй выборки данных (маркировка profile2_ds), включающей данные в соответствии с «профилем 2» и размеченные для каждого пользователя: 0 – обычный пользователь, 1 – мошенник.

Шаг 4. Выполнить формирование обучающей, кроссвалидационной и тестовой выборок. В этом случае осуществляется разбиение выборки в

соответствии с пропорцией: 80% для обучения, а 20% оставшихся данных – для тестирования.

Шаг 5. Осуществить настройку и оценку моделей. Настройка моделей осуществляется на выделенной обучающей выборке исходя из предопределенного числа моделей. Для поиска наилучшей модели используется приём поиска гиперпараметров по сетке (GridSearch) с проверкой на кроссвалидационных выборках.

Шаг 6. Выбрать наилучшую модель (с наилучшими значениями гиперпараметров) на основе значения точности для выборок profile1_ds и profile2_ds. Затем производится дообучение этих моделей на всём (100%) наборе данных.

Шаг 7. Выполнить сохранение параметров лучшей модели для последующего обнаружения ТКМ.

Этап применения модели:

Шаг 1. Загрузить параметры модели для соответствующей задачи обнаружения ТКМ.

Шаг 2. Выполнить предобработку исходных данных в режиме реального времени.

Шаг 3. Подготовить входных векторов для представления их в качестве входов в моделях.

Шаг 4. Выполнить расчет выходных значений для полученных моделей.

Шаг 5. Генерация события-оповещения, если спрогнозированное поведение пользователя относится к МП.

В четвертой главе рассмотрены вопросы обоснования эффективности предлагаемых положений.

Испытание и оценка эффективности методов и методик (на общедоступных выборках данных). Перед внедрением в существующие бизнес-процессы целесообразно выполнить испытания на открытых данных для обоснования реализуемости и эффективности подходов. Внедрение новых решений в бизнес-процессы является сложной процедурой в первую очередь из-за рисков, которые могут возникнуть при некорректном решении. Для апробации предлагаемых подходов были использованы выборки общедоступные выборки данных: Нодобо (набор CDR1) и RealityCommons (calls dataset) (набор CDR2). Таким образом анализировались два набора данных. Первый имеющий следующую формализацию.

Для испытания методов было разработано программное обеспечение на языке Python в среде Jupyter Notebook. Основной довод в пользу такого решения – набор библиотек машинного обучения, которыми пользуются исследователи по всему миру, например, библиотека scikit-learn.

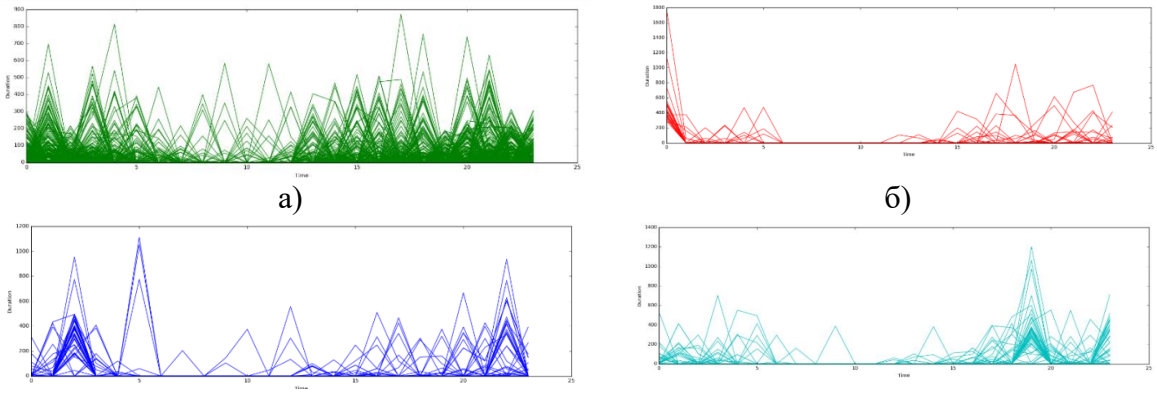
Выбор и предварительная обработка данных для проведения испытаний. При загрузке выявлено, что максимальное и минимальное значения выглядят неправдоподобно и, скорее всего, имеет место ошибка загрузки данных. Это же подтверждает и большая разница между значением среднего и медианы. В данной работе используется следующий подход:

выбросами считаются все неположительные значения признака duration, а положительные входящие в интервал [μ-aσ;μ+aσ], где μ – среднее, а σ – среднеквадратическое отклонение, а параметр a задается экспертом или определяется опытным путем (в работе значение *a* варьировалось от 0,1 до 3).

Также следует обратить внимание на отсутствующие значения признака duration. Для заполнения пропущенных значений, а также для замены значений выбросов применяется распространённый и простой прием в машинном обучении: замена пропущенных значений и выбросов средними значениями.

Апробация метода кластеризации профилей абонентов. В соответствии с настройками гиперпараметров были получены различные выборки данных, используемых для дальнейшей кластеризации. Так были получены «дневные», «недельные» и «ежемесячные» профили поведения абонентов.

Визуализация распределения профилей пользователя по кластерам представлена на рисунке 1. Следует отметить, что при небольшом числе кластеров качество кластеризации высокое, однако возникает ситуация, при которой большинство профилей пользователей формируют один кластер.



а) б) в) г)

Рисунок 1 - Визуализация результатов выполнения кластеризации дневных профилей алгоритмом MiniBatch-Kmean по каждому кластеру: а) кластер №0, б) кластер №1, в) кластер №2, г) кластер №3.

k (количество кластеров)	Значение $s(i)$ для алгоритма – средних	Значение $s(i)$ для алгоритма MiniBatch k–средних
3	0.75014	0.647782
4	0.69713	0.623702
5	0.66505	0.722554
6	0.63086	0.648688
7	0.657	0.568248

а)

k (количество кластеров)	$s(i)$ k–средних	$s(i)$ MiniBatch k–средних
4	0.65896	0.593202
6	0.53271	0.416784
8	0.46985	0.43942
10	0.5948	0.511404
12	0.39952	0.455305

б)

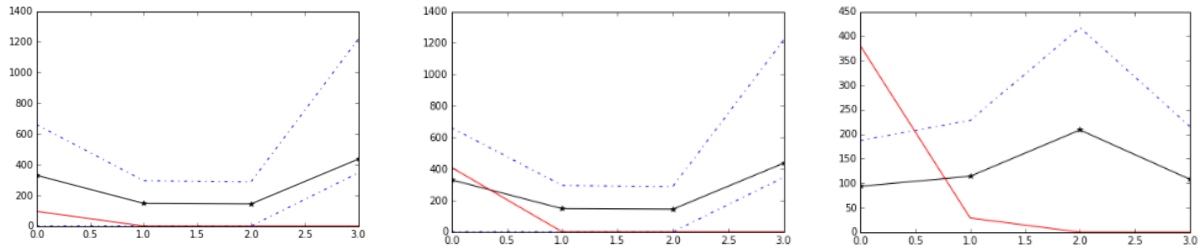
Рисунок 2. – результаты оценки качества кластеризации для а) дневных профилей, б) недельных профилей

На рисунке 2 представлены результаты оценки качества кластеризации. Были проведены различные эксперименты с различными настройками

гиперпараметров с использованием методов кластеризации на основе k-средних и MinibatchKmean на выборке данных из исходного набора данных.

Испытание метода идентификации изменений в поведении абонентов на основе кластеризации. В соответствии с разработанным в главе 3 методом кластеризации абонентов были проведены испытания на приведенном выше наборе данных. Были проведены эксперименты на дневных профилях данных. Для тестирования преобразованная выборка данных была разделена на обучающую (по которой формировались кластеры) и тестовую: профили абонентов, которые не вошли в выборку для кластеризации. Оценка качества метода производилась следующим образом: выбирался кластер, оценивалось качество идентификации изменений на профилях абонентов кластера в тестовой выборке и качество идентификации изменений на профилях абонентов других кластеров тестовой выборки.

Эксперимент на дневных кластерах с 4-мя интервалами в течение дня.



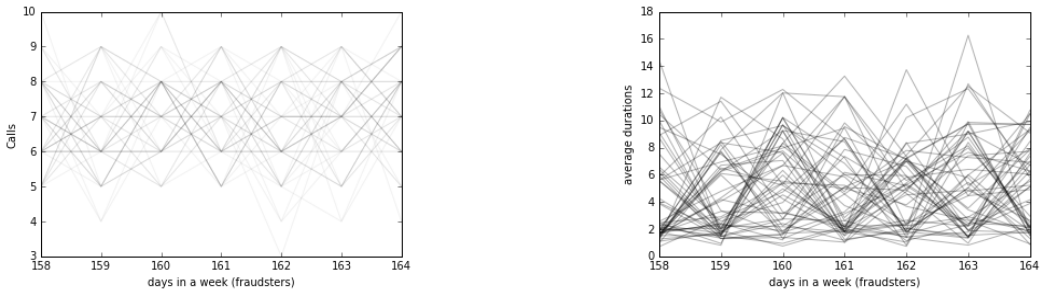
а) б) в)

Рисунок 3 - Результаты работы метода идентификации изменения поведения для абоненто в кластера а) №2, б) №1, в) №3.

На рисунке 3а) показаны результаты идентификации изменения поведения для выборочных абонентов кластера №2 ($id = 37$). Для абонентов тестовой выборки значение меры $ratio = 0.083$, т.е. 91.7% абонентов этого кластера были отнесены к абонентам для которых изменения в поведении не идентифицировано. На рисунке 3б) показаны результаты идентификации изменения поведения для выборочных абонентов кластера №1 ($id = 33$). Таким образом ожидалось, что изменение в поведении абонентов будет зафиксировано. Для абонентов тестовой выборки значение меры $ratio = 1$, т.е. все абоненты этого кластера были отнесены к абонентам для которых были идентифицированы изменения. На рисунке 3в) показаны результаты идентификации изменения поведения для выборочных абонентов кластера №3 ($id = 60$). Аналогично предыдущему $ratio = 1$, т.е. 100% абонентов этого кластера были отнесены к абонентам для которых были идентифицированы изменения.

Испытание метода выявления изменения поведения пользователей услуг на основе проактивного подхода для выявления телекоммуникационного мошенничества

На рисунке 4 представлена визуализация искусственно сгенерированных данных о звонках пользователей по дням.



а) б)

Рисунок 4 - Визуализация частоты звонков и средней продолжительности разговоров для «нормальных пользователей» (а) и «мошенников» (б) по «недельным данным». Обозначения на рисунке: Calls – число звонков в течение суток, Average Duration – средняя продолжительность звонков (мин.). По оси абсцисс – номера дней (суток), отсчитанные с начала года.

После обучения моделей была проведена оценка их достоверности (отношение суммы истинно-положительной и истинно-отрицательной оценок к сумме истинно-положительной, истинно-отрицательной и ложно-отрицательной оценками). Результаты работы для тестовой выборки представлены в таблице 2.

Таблица 2 –Результаты работы алгоритмов на обучающей выборке.

Обозначения алгоритмов	**Средняя достоверность**
LR	0.924
k-NN	**0.926**
CART	0.841
NB	0.919
SVM	**0.927**

Для определения наилучшего алгоритма был выполнен поиск лучшей комбинации гиперпараметров. Для алгоритма *k*-NN гиперпараметр '*n_neighbors*' – число ближайших соседей, по умолчанию равен 7. Были получены результаты для всех нечётных значений этого параметра от 1 до 21 на кроссвалидационной выборке данных. Для *k*-NN алгоритма лучшие результаты классификации составили 0.926 со значением параметра *n_neighbors* = 5. Алгоритм SVM со значением *C*=0.9 и сигмоидальным видом ядра показал наилучший результат с точки зрения точности классификации. Точность алгоритма SVM была выше точности алгоритма *k*-NN, хотя и незначительно. Поэтому алгоритм SVM был выбран как основной для генерации модели обнаружения ТКМ. Средние значения мер точности по этому алгоритму для тестовых выборок оказались следующие: точность (precision) = 0.78, полнота (recall) = 0.83.

Для реализации предлагаемых подходов для внедрения в корпоративную систему разработана архитектура и построена многоуровневая система поддержки реализации персонифицированных услуг в ТКП на основе лямбда-архитектуры, комбинирующий в себе обработку данных в потоковом режиме (режиме реального времени) и пакетную обработку (см. рисунок 5).

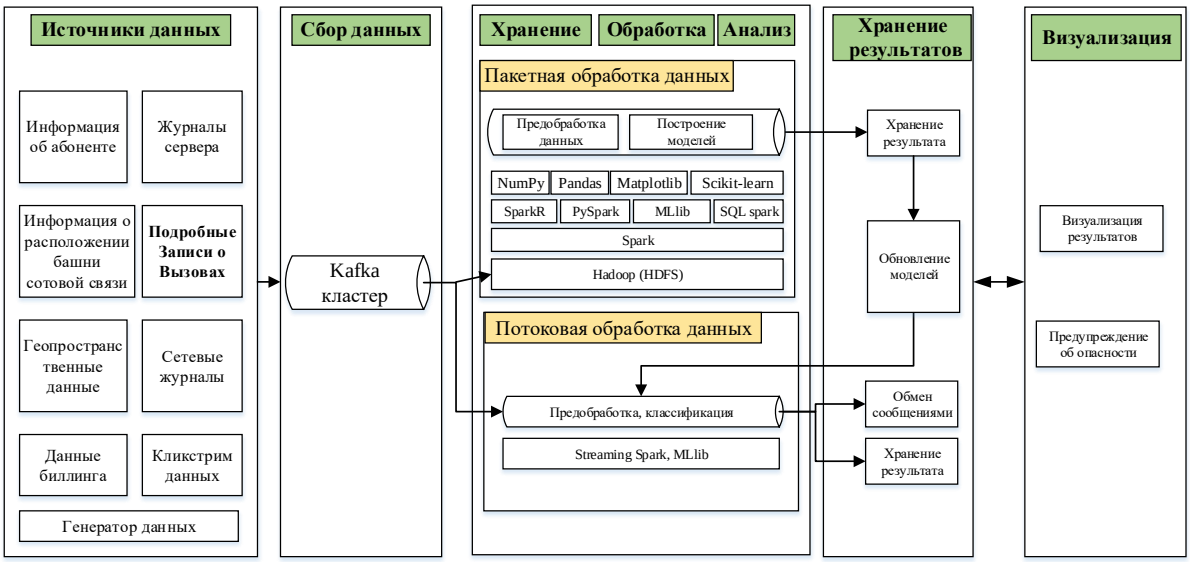


Рисунок 5 - Архитектура системы обнаружения мошенничества в телекоммуникации

Архитектура программной системы обнаружения ТКМ состоит из пяти подсистем: S1 – подсистема источников данных; S2 – подсистема сбора данных; S3 – подсистема обработки и анализа данных; S4 – подсистема хранения результатов обработки данных; S5 – подсистема визуализации результатов.

Результаты работы были апробированы во Вьетнамской государственной телекоммуникационной корпорации (ВьетГТКК) «VNPT_Media». Проведена оценка качества разработанных методов обработки данных при реализации персонифицированных услуг, работа которой основана на выборке данных из 150000 абонентов телефонной связи «VNPT_Media.

Таблица 3 - Результат оценки эффективности применения предложенного решения в ВьетГТКК «VNPT_Media» «Группа показателей 1»

БП	Оценка качества		
	Критерии	Q1	Q2
C1.2	эффективность персонифицированной рекламы	12 %	16 %
O2.5	процент новых абонентов–пользователей персонализированными услугами	23 %	35 %
Y4.3.	количество обнаруженных случаев мошенничества	22	43

Таблица 4 - Результат оценки эффективности применения предложенного решения в ВьетГТКК «VNPT_Media» «Группа показателей 2»

БП	Оценка качества		
	Критерии	Q1	Q2
O1.9	отток абонентов	5%	2,2%
O1.9	количество жалоб о качестве обслуживания	121	57
Y4.3.	количество жалоб о мошенничестве	82	47
Y4.3.	уровень ущерба, причиненный мошенничеством	2%	1.12 %

Применение предложенного решения прошло выборочно с мая 2016 г. по январь 2017 г. Сравнительные оценки качества предоставляемых персонифицированных услуг приведены в таблицах 3,4. (Q1 -- Квартал 1 (не применено предложенное решение), Q2 Квартал 2 (применено предложенное решение)).

В результате эффективность персонифицированной рекламы повысилась на 4%, число новых абонентов–пользователей персонализированными услугами увеличилось на 12%, отток абонентов уменьшился на 2.8 %, количество жалоб о качестве обслуживания снизилось на 47%; количество жалоб о мошенничестве снизилось на 57% и уровень ущерба, причиненный мошенничеством, снизился на 56%.

В заключении сформулированы основные результаты и выводы по работе.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ РАБОТЫ

Главным результатом исследования является повышение качества предоставления услуг ТКП за счет разработанных новых методов интеллектуальной обработки информации. Получены следующие основные результаты.

1. В результате системного анализа функционирования телекоммуникационного предприятия выделены и формализованы операции, направленные на реализацию персонифицированных услуг, которые могут быть рассмотрены как инвариантные к бизнес-процессам и для которых имеется проблема совершенствования методов интеллектуальной обработки данных.
2. На основе анализа информации в ТКП предложена новая модель абонента телекоммуникационного предприятия, отличающаяся встроенными показателями анализа динамики пользования абонентом различными персонифицированными ТКУ. В рамках модели определены операции преобразования признаков. Формальное представление абонента может быть использовано для определения персонифицированной услуги как для абонента, так и для группы схожих абонентов.
3. Предложены новые методы интеллектуальной обработки данных, включающие: метод кластеризации абонентов ТКП, позволяющий определять группы схожих абонентов по их поведению; метод идентификации изменений в поведении абонентов, позволяющий выявлять изменения в поведении без предварительной разметки выборки данных и метод выявления изменения поведения пользователей услуг, позволяющий проактивно определять проявления мошеннических действий.
4. Выполнены испытания и показана эффективность предлагаемых методов на открытых источниках данных.
5. Разработана архитектура многоуровневой системы управления реализацией персонифицированных услуг в ТКП, включающая 5 слоев для распределенной обработки данных.

6. Результаты работы были апробированы во Вьетнамской государственной телекоммуникационной корпорации (ВьетГТКК) "VNPT_Media". Получены следующие значимые изменения показателей качества: число новых абонентов - пользователей персонализированными услугами увеличилось на 12%, количество жалоб о качестве обслуживания снизилось на 47%; количество жалоб о мошенничестве снизилось на 57% и уровень ущерба, причиненный мошенничеством, снизился на 56%.

На основе полученных результатов можно сделать выводы об эффективности предлагаемых методов интеллектуальной обработки данных в ТКП, а следовательно, о достижении цели исследования.

Перспективы использования работы. Дальнейшее развитие данного исследования возможно по следующим направлениям. Применение разработанных методов обработки информации в других бизнес-процессах ТКП. Рассмотрение и использование других алгоритмов машинного обучения и искусственного интеллекта, например, рекуррентных нейронных сетей и сверточных нейронных сетей.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, включённых в Перечень ВАК

1. Туан Ань Нгуен Разработка метода проактивного обнаружения мошенничества потребителей услуг телекоммуникационной компании / Туан Ань Нгуен, М.В. Щербаков, Ван Фу Чан, А.Г. Кравец // Прикаспийский журнал: управление и высокие технологии. - 2016. - № 4. - С. 43-52.
2. Метод сбора и слияния разнотипных данных в проактивных системах интеллектуальной поддержки принятия решений / Ван Фу Чан, М.В. Щербаков, Туан Ань Нгуен, Д.А. Скоробогатченко // Нейрокомпьютеры: разработка, применение. - 2016. - № 11. - С. 40-44.
3. Нгуен, Туан Ань Обзор систем обнаружения мошенничества в телекоммуникационном предприятии / Туан Ань Нгуен // Современная наука: актуальные проблемы теории и практики. Серия «Естественные и технические науки». - 2016. - № 11. - С. 49-54.
4. Чан, Ван Фу Обзор архитектур систем поддержки принятия решений, использующих аналитику данных в режиме реального времени / Ван Фу Чан, М.В. Щербаков, Туан Ань Нгуен // Известия ВолгГТУ. Сер. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. - Волгоград, 2016. - № 3 (182). - С. 95-100.

Публикации в изданиях, индексируемых в международных базах научного цитирования (Scopus, Web of Science)

5. Чан, Ван Фу Yet Another Method for Heterogeneous Data Fusion and Preprocessing in Proactive Decision Support Systems: Distributed Architecture Approach / Чан, Ван Фу, М.В. Щербаков, Нгуен, Туан Ань // Distributed Computer and Communication Networks : 20th International Conference (DCCN 2017) (Moscow, Russia, September 25–29, 2017) : Proceedings / editors: V.M. Vishnevskiy [et al.]. – [Springer International Publishing AG], 2017. – P. 319-330. – (Ser. Communications in Computer and Information Science).
6. Чан, Ван Фу EVGEN: A framework for event generator in proactive system design [Электронный ресурс] / Ван Фу Чан, М.В. Щербаков, Туан Ань Нгуен // 7th International Conference on Information, Intelligence, Systems & Applications (IISA) (Greece, 13-15 July 2016) / Institute of Electrical and Electronics Engineers (IEEE). – [Publisher: IEEE]. – DOI:

10.1109/IISA.2016.7785402. – URL : http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7774711.
7. Tracking Events in Mobile Device Management System / Ngoc Duong Bui, A. G.Kravets, Tuan Anh Nguyen and Le Thanh Tung Nguyen // Proceedings of 6th International Conference on Information, Intelligence, Systems and Applications (IISA2015) July 6-8 2015. − Ionian University, Corfu, Greece, IEEE, 2015. − pp. 01-06. DOI: 10.1109/IISA.2015.7388127.

Прочие публикации

8. Нгуен, Туан Ань Архитектура обнаружения мошенничества в телекоммуникационном предприятии с Hadoop / Туан Ань Нгуен, В.А. Камаев, М.В. Щербаков // Мир науки и инноваций. - 2015. - Вып. 2, т. 2 «Технические науки». - С. 75-78.
9. Нгуен, Туан Ань Архитектор системы анализа данных в телекоммуникациальном предприятии с Hadoop / Туан Ань Нгуен // XX региональная конференция молодых исследователей Волгоградской области (г. Волгоград, 8-11 дек. 2015 г.) : тез. докл. / редкол.: А.В. Навроцкий (отв. ред.) [и др.] ; Комитет молодёжной политики Волгогр. обл., Совет ректоров вузов Волгогр. обл., ВолгГТУ. - Волгоград, 2016. - С. 210-211.
10. Нгуен, Туан Ань Архитектура системы обнаружения мошенничества в телекоммуникационном предприятии с Hadoop / Туан Ань Нгуен, В.А. Камаев // Юность и Знания – Гарантия Успеха – 2015 : сб. науч. тр. 2-й междунар. науч.-практ. конф. (1-2 окт. 2015 г.). В 2 т. Т. 2 / редкол.: А.А. Горохов (отв. ред.) / Юго-Западный гос. ун-т, ЗАО «Университетская книга» [и др.]. - Курск, 2015. - С. 70-72.
11. Нгуен, Туан Ань Архитектура обнаружения мошенничества в телекоммуникационном предприятии с Hadoop [Электронный ресурс] : доклад / Туан Ань Нгуен, В.А. Камаев, М.В. Щербаков // Интеллектуальный потенциал XXI века ‘2015 : матер. междунар. науч.-практ. молодёжной Интернет-конф. (Украина, 10-22 нояб. 2015 г.). Секция «Технические науки», подсекция «Информатика, вычислительная техника и автоматизация» / Проект SWorld. – 5 с. – Режим доступа : http://www.sworld.education/conference/molodej-conference-sw/the-content-of-conferences/archives-of-individual-conferences/november-2015.
12. Нгуен, Туан Ань Архитектура проактивной системы сбора и обработки геопространственных данных / Туан Ань Нгуен, Ван Фу Чан // Наука и современность – 2016 : сб. матер. XLV междунар. науч.-практ. конф. (г. Новосибирск, 26 мая, 14 июня 2016 г.) / под общ. ред. С.С. Чернова ; Центр развития научного сотрудничества (ЦРНС). - Новосибирск, 2016. - С. 122-127.

Свидетельства о регистрации программ для ЭВМ

13. Свид. о гос. регистрации программы для ЭВМ № 2017611602 от 7 февраля 2017 г. Российская Федерация, МПК (нет). Программное обеспечение обнаружения мошенничества в телекоммуникационных предприятиях / М.В. Щербаков, Туан Ань Нгуен, Ван Фу Чан; ВолгГТУ. - 2017.
14. Свид. о гос. регистрации программы для ЭВМ № 2017660307 от 20 сентября 2017 г. Российская Федерация, МПК (нет). Распределенная система слияния и предобработки разнородных данных с разных источников / М.В. Щербаков, Ван Фу Чан, Туан Ань Нгуен; ВолгГТУ. - 2017.

Подписано в печать _____2017 г. Заказ № _____. Тираж 100 экз. Печ. л. 1,1
Формат 60 × 84 1/16. Бумага офсетная. Печать офсетная.
Типография ИУНЛ Волгоградского государственного технического университета. 400005, г. Волгоград, просп. им. В.И. Ленина, 28, корп. № 7