ЗАКЛЮЧЕНИЕ ДИССЕРТАЦИОННОГО СОВЕТА Д 002.199.01 НА БАЗЕ ФЕДЕРАЛЬНОГО ГОСУДАРСТВЕННОГО БЮДЖЕТНОГО УЧРЕЖДЕНИЯ НАУКИ САНКТ-ПЕТЕРБУРГСКОГО ИНСТИТУТА ИНФОРМАТИКИ И АВТОМАТИЗАЦИИ РОССИЙСКОЙ АКАДЕМИИ НАУК ПО ДИССЕРТАЦИИ НА СОИСКАНИЕ УЧЕНОЙ СТЕПЕНИ КАНДИДАТА НАУК

аттестационное дело № _____

решение диссертационного совета 29.09.2016 г. № 2

О присуждении Тушкановой Ольге Николаевне, гражданке Российской Федерации, ученой степени кандидата технических наук.

Диссертация «Семантические структуры и причинные модели больших данных для принятия решений с приложением к рекомендательным системам» по специальности 05.13.01 – «Системный анализ, управление и обработка информации» принята к защите 26 июля 2016 г. (протокол № 3) диссертационным советом Д 002.199.01 на базе Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук, 199178, Россия, Санкт-Петербург, 14 линия ВО, дом 39, утвержден приказом Рособрнадзора номер 2472-618 от 8 октября 2010 года.

Соискатель Тушканова Ольга Николаевна 1988 года рождения в 2011 году получила степень магистра по направлению «Системный анализ и управление» на факультете высоких технологий в Южном Федеральном Университете, в 2015 г. окончила очную аспирантуру Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук. Удостоверение о сдаче кандидатских экзаменов №6/196 выдано в 11.05.2016 года Федеральным государственным бюджетным учреждением науки Санкт-Петербургским институтом информатики и автоматизации Российской академии наук. В настоящее время Тушканова Ольга Николаевна работает научным сотрудником лаборатории интеллектуальных систем Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук.

Диссертация выполнена в лаборатории интеллектуальных систем Федераль-

ного государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук.

Научный руководитель – доктор технических наук, профессор ГОРОДЕЦКИЙ Владимир Иванович, основное место работы: Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук, главный научный сотрудник.

Официальные оппоненты:

ГАВРИЛОВА Татьяна Альбертовна, доктор технических наук, профессор, заведующая кафедрой информационных технологий в менеджменте института «Высшая школа менеджмента Санкт-Петербургского государственного университета (ВШМ СПбГУ)»;

МУРОМЦЕВ Дмитрий Ильич, кандидат технических наук, доцент, заведующий кафедрой информатики и прикладной математики Федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»
дали положительные отзывы на диссертацию.

Ведущая организация – Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)» в своем положительном заключении, подписанном Цехановским Владиславом Владимировичем, кандидатом технических наук, заведующим кафедры информационных систем, ученым секретарём кафедры информационных систем Коробкиными Владимиром Петровичем и утвержденном Жарковским А.В., зам. проректора по научной работе, указала, что диссертационная работа О.Н. Тушкановой представляет собой законченную научно-квалификационную работу на актуальную тему, результаты которой обладают научной новизной, теоретической и практической значимостью. В работе решена важная и актуальная научная задача разработки семантических моделей и алгоритмов ассоциативно-причинной классификации для принятия решений в области обработки больших данных. Пред-

ставленная работа характеризуется полнотой изложения рассматриваемых проблем, обладает внутренним единством, содержит аргументированные выводы по каждой главе, свидетельствует о личном вкладе соискателя в соответствующую отрасль науки. Работа отвечает общепринятым критериям научного стиля изложения материала. Основное содержание работы, а также выводы и результаты диссертационного исследования, достаточно полно отражены в автореферате.

Диссертация Тушкановой О.Н. отвечает требованиям, п. 9 Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013 № 842, предъявляемым к кандидатским диссертациям, а ее автор заслуживает присуждение ученой степени кандидата технических наук по специальности - 05.13.01 – «Системный анализ, управление и обработка информации (технические системы)».

Соискатель имеет 20 опубликованных работ, в том числе, 9 работ по теме диссертации, опубликованных в рецензируемых научных изданиях, из них 3 работы опубликованы в изданиях, рекомендуемых ВАК РФ.

Основные научные результаты опубликованы в 7 научных работах общим объемом 86 с., из которых 5 работ объемом 64 с., выполнены в соавторстве, а 2 работы объемом 22 с. – лично.

Наиболее значимые работы по теме диссертации:

1. **Тушканова, О. Н.** Экспериментальное исследование численных мер оценки ассоциативных и причинных связей в больших данных [Текст] / **О. Н. Тушканова** // Информационные технологии и вычислительные системы. - 2015. - №3. - С. 16-25.

2. **Тушканова, О. Н.** Ассоциативная классификация: аналитический обзор. Часть 1 [Текст] / **О. Н. Тушканова,** В. И. Городецкий // Труды СПИИРАН. - 2015. - №1(38). -С. 183 – 203. *Личный вклад соискателя – 50%.*

3. **Тушканова, О. Н.** Ассоциативная классификация: аналитический обзор. Часть 2 [Текст] / **О. Н. Тушканова,** В. И. Городецкий // Труды СПИИРАН. - 2015. - №2 (39). - С. 212–240. *Личный вклад соискателя – 50%.*

4. **Tushkanova, O.** Agent-based customer profile learning in 3G recommending systems: ontology-driven multi-source cross-domain case / **O. Tushkanova**, V.

Gorodetsky, V. Samoylov // Proc. of the Tenth International Workshop on Agents and Data Mining Interaction (ADMI-14), May 5-9, 2014, Paris, France. Lecture Notes in Artificial Intelligence. Eds. Symeonidis A.L., Zeng Y., Cao L., Gorodetsky V., An B., Coenen F., Yu P.S. - Springer. -Vol. 9145. - 2015. - pp. 12 – 25. *Личный вклад соискателя – 40%.*

5. **Tushkanova, O.** Comparative Analysis of the Numerical Measures for Mining Associative and Causal Relationships in Big Data / **O. Tushkanova O.** // Creativity in Intelligent, Technologies and Data Science. Communications in Computer and Information Science. Eds. Kravets A., Shcherbakov M., Kultsova M., Shabalina O. - Springer. - Vol. 535. - 2015. - pp 571-582.

6. **Tushkanova, O**. Data-driven Semantic Concept Analysis for Automatic Actionable Ontology Design / **O. Tushkanova**, V. Gorodetsky // Proc. of the IEEE International Conference on Data Science and Advanced Analytics (DSAA). - 2015. - pp. 1-9. *Личный вклад соискателя – 60%.*

7. **Тушканова, О. Н.** Онтологии и персонификация профиля пользователя в рекомендующих системах третьего поколения [Текст] / **О. Н. Тушканова,** В. И. Городецкий // Онтология проектирования. - 2014. - №3(13). - С. 7-31. *Личный вклад соискателя – 50%.*

Оригинальность содержания диссертации составляет не менее 96% от общего объёма текста; цитирование оформлено корректно; заимствованного материала, использованного в диссертации без ссылки на автора либо источник заимствования, не обнаружено; научных работ, выполненных соискателем учёной степени в соавторстве без ссылок на соавторов не выявлено.

На автореферат диссертации поступило 6 отзывов, все отзывы положительные:

1) Институт автоматики и процессов управления Дальне-Восточного отделения Российской академии наук. Отзыв составила заместитель директора по научной работе, доктор технических наук, профессор Грибова В.В. Замечания: (1) целью диссертационного исследования заявлена разработка набора алгоритмов и их экспериментальная оценка по таким характеристикам как масштабируемость, вычислительная эффективность и точность. Однако ни для одного из предложенных

алгоритмов в работе не описано, за счет чего достигается масштабируемость и как экспериментальная оценка этого критерия была проведена; (2) насколько отчуждаемыми являются предложенные алгоритмы и технология? Насколько сложно настроить данную технологию для решения задач в других предметных областях, какими свойствами должны обладать данные в них, есть ли какие-то ограничения? В работе заявлено, что автором даны практические рекомендации, но в автореферате их найти не удалось.

2) Северо-Кавказский Федеральный университет. Отзыв составила заведующая кафедрой информационных систем и технологий, доктор физико-математических наук, профессор Дроздова В.И. Замечания: (1) в автореферате отсутствуют аргументы в пользу выбора в качестве источника данных облачного ресурса средства DBpedia; (2) в автореферате не приведены описания примеров применения предложенных алгоритмов.

3) Южно-Российский Государственный политехнический университет (НПИ) им. М.И. Платова. Отзыв составил зав. кафедрой «Информационные и измерительные системы и технологии», доктор технических наук, профессор Горбатенко Н.И. Замечания: (1) сложность понимания деталей описанной технологии без рассмотрения её шагов на конкретных примерах; (2) для построения модели выработки рекомендаций используется метод бинарного дерева решений (страница 13), однако критериев или причин выбора именно этого метода автор не приводит.

4) Институт компьютерных и информационных технологий Южного федерального университета. Отзыв составил зав. кафедрой дискретной математики и методов оптимизации, доктор технических наук, профессор Курейчик В.М. Замечания: (1) представляется, что некоторые второстепенные результаты работы описаны в ней слишком детально. Это относится к обзорной части диссертации по разделу ассоциативно-причинного анализа, методам автоматизации процесса построения онтологий, описанию технологии исследования метрик для оценки причинных зависимостей. Эти результаты достаточно подробно опубликованы, и поэтому здесь можно было бы ограничиться кратким резюме со ссылками на публикации; (2) в описании разработанного программного продукта не акцентируют-

ся аспекты повторного использования компонент, что затрудняет оценку программного обеспечения в этом смысле.

5) Институт компьютерных наук и технологий Санкт-Петербургского политехнического университета Петра Великого. Отзыв составил кандидат технических наук, профессор кафедры «Системный анализ и управление» Станкевич Л.А. Замечания: (1) в разделе «Краткое содержание работы», глава 4, приведена схема построения семантического профиля пользователя (рисунок 4), где предложено использовать материалы Википедии, в которой, как правило, имеется множество недостоверной информации, которая может исказить семантический профиль; (2) из материалов автореферата трудно понять, что такое коллаборативная фильтрация, и как она позволяет улучшить результаты по точности классификации.

6) Московский Государственный университет им. М.В. Ломоносова. Отзыв составил доцент кафедры математической теории интеллектуальных систем механико-математического факультета, доктор технических наук, профессор Рыжов А.П. Замечания: (1) автор использует достаточно специфическую терминологию без определения или хотя бы объяснения этих терминов – например рекомендательные системы «третьего поколения» (с. 4 и ниже); (2) в автореферате упоминается, что работоспособность алгоритмов проверена на примерах из экспериментального набора данных Amazon (с. 12, первый абзац), однако не приведено никаких результатов такого эксперимента; (3) имеются огрехи в оформлении текста – например, символ в конце предпоследней строки на с. 14; (4) один из важнейших вопросов фильтрации правил профиля – порог отсечения – описан в очень общих словах (с. 15, первый абзац) и не достаточен для понимания логики автора; (5) интересно было бы увидеть результаты сравнения работы предложенных алгоритмов с имеющимися на известных тестовых наборах данных, достаточно хорошо представленными в Интернете.

Выбор официальных оппонентов и ведущей организации обосновывается тем, что д.т.н., профессор Гаврилова Т.А. является одним из ведущих российских специалистов в области инженерии знаний, автором множества работ по тематике управления знаниями, в том числе соавтором первого учебника в данной области; к.т.н., доцент Муромцев Д.И. – известный специалист в области интеллектуаль-

ных систем и семантических технологий, является руководителем международной лаборатории «Интеллектуальные методы обработки информации и семантические технологии» Университета ИТМО, а также автором множества работ в области интеллектуальных технологий и семантических моделей данных; ведущая организация - Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)», является известной как в России, так и за рубежом организацией в области разработки и создания интеллектуальных систем, основанных на знаниях.

Диссертационный совет отмечает, что на основании выполненных соискателем исследований:

доказана (теоретически и экспериментально) вычислительная эффективность и семантическая корректность меры оценки «силы» причинной зависимости между атрибутами данных – коэффициент регрессии случайных событий. Рекомендации по ее выбору основаны на обширном экспериментальном исследовании и построены на численных оценках;

разработаны:

– алгоритм автоматического построения семантической модели больших данных, ориентированный на извлечение причинных правил и минимизацию их количества в модели ассоциативно-причинной классификации. В его основу положена новая методика семантического анализа понятий. Алгоритм отличается тем, что построен как комбинация методов автоматического извлечения иерархии понятий онтологии данных из глобального источника знаний и методов генерации двойственных формальных понятий, вероятностные свойства которых позволяют задать условия останова процесса построения семантической модели;

– алгоритм поиска множества причинных зависимостей между атрибутами данных. Алгоритм отличается тем, что использует обоснованную соискателем меру оценки силы причинных зависимостей в данных (коэффициент регрессии случайных событий) и семантическую модель данных, а именно, метаинформацию о данных, содержащуюся в ней, что обеспечивает вычислительную эффективность предложенного алгоритма;

– алгоритм минимизации размерности пространства причинных правил в модели ассоциативно-причинной классификации, который позволяет устранить избыточность модели принятия решений. Алгоритм отличается тем, что использует механизм кластеризации множества правил на основе метода корреляционных плеяд и алгоритма разрезания графа на связные компоненты;

предложена новая семантическая модель данных, которая позволяет представлять синтаксис и семантику данных и метаинформацию о них в рамках единой семантической структуры. Такая структура строится с помощью разработанного алгоритма автоматического построения семантической модели больших данных и обеспечивает эффективный поиск причинных зависимостей в данных для алгоритмов ассоциативно-причинной классификации;

введены:

– новое понятие онтологии данных как семантической надстройки над обрабатываемыми данными;

– методика семантического анализа понятий, которая представляет собой комбинацию методов извлечение иерархии понятий онтологии и методов анализа формальных понятий;

– новое понятие семантической структуры данных как средства представления семантической модели данных в виде двух полурешёток – полурешетки понятий онтологии данных и полурешётки двойственных им формальных понятий;

– свойства, которыми должна обладать мера причинной зависимости между атрибутами данных.

Теоретическая значимость исследования обоснована тем, что в нем:

доказано, что разработанные алгоритмы обработки больших данных, используемые на различных этапах построения модели принятия решений (на этапе агрегирования данных, на этапе построения семантической модели, на этапе генерации причинной модели данных и при ее оптимизации) являются устойчивыми и вычислительно эффективными, что обеспечивается специальной организацией вычислений, при которой не происходит накопления ошибок и не появляют-

ся ложные корреляции. Теоретическая значимость работы обосновывается также предложенной в работе новой методологией и алгоритмом минимизации размерности причинной модели данных для принятия решений.

Применительно к проблематике диссертации результативно (эффективно, то есть с получением результатов, обладающих новизной):

изучено текущее состояние исследований в области интеллектуальной обработки больших данных, ассоциативно-причинной классификации и автоматического построения семантических моделей данных;

проанализированы основные нерешенные методические и алгоритмические проблемы обработки больших данных. Большинство традиционных методов интеллектуального анализа данных и машинного обучения не могут быть напрямую применены к анализу больших данных из-за их вычислительной неустойчивости и/или вычислительной сложности;

использованы методы корреляционного, ассоциативного и причинного анализа, методы машинного обучения и объединения решений распределенных классификаторов, методы теории графов, теории вероятностей и математической статистики, методы и средства онтологического моделирования, методы теории частично упорядоченных множеств и решеток, методы анализа формальных понятий и кластерного анализа;

изложены особенности задач построения семантических и причинных моделей больших данных и алгоритмические основы семантического анализа понятий, который используется в качеств механизма для построения семантической модели данных;

проведена модернизация существующих алгоритмов поиска причинных связей в больших данных и предложена их модернизация за счет использования семантической модели данных и меры, определяющей «силу» причинных зависимостей между атрибутами данных.

Значение полученных соискателем результатов исследования для практики подтверждается тем, что:

разработаны и внедрены (указать степень внедрения) следующие результаты диссертационной работы:

1) алгоритм построения онтологии данных с помощью технологии семантического анализа понятий; модель больших данных, которая представляет метасвойства данных, их синтаксис и семантику в рамках единой структуры; мера оценки силы причинной связи (коэффициент регрессии случайных событий), которая положена в основу ассоциативно-причинного анализа, и алгоритм поиска причинных зависимостей между атрибутами. Эти результаты диссертационной работы внедрены в компании Самсунг на уровне исследовательского программного прототипа, который был разработан в рамках проекта «Многоагентные алгоритмы для кросс-доменных рекомендательных систем», руководитель Городецкий В.И., выполненной по контакту СПИИРАН с ООО «Исследовательский Центр Самсунг» в 2014 году;

2) мера оценки силы причинной связи (коэффициент регрессии случайных событий) и алгоритм поиска зависимостей между атрибутами реализованы на уровне рабочего кода, разработанного при выполнении контракта с компанией EMC (США) в 2013-2016 гг. (руководитель Городецкий В.И.), (1) в алгоритме автоматического инкрементного обучения для улучшения распознавания табличных данных; (2) в алгоритме оценки степени размытия мобильных изображений документов для предсказания качества работы инструмента EMC Captiva OCR;

определены возможности и перспективы практического использования полученных результатов диссертации на примере приложения из области рекомендательных систем;

создана программная библиотека, реализующая все разработанные модели и алгоритмы в виде набора Java-классов, построенных с учетом возможности повторного использования в различных прикладных задачах обработки больших данных для принятия решений;

представлены практические рекомендации по выбору алгоритмов и технологий для разработки современных рекомендательных систем и предложения по дальнейшему усовершенствованию разработанных моделей и алгоритмов.

Оценка достоверности результатов исследования выявила:

для экспериментальных работ - воспроизводимость результатов многократных экспериментов, выполненных на сертифицированном современном оборудовании;

наличие положительного опыта внедрения полученных научных результатах;

а также, что:

идея базируется на разностороннем критическом анализе современного состояния исследований в области интеллектуальной обработки и семантического моделирования больших данных;

корректно **использованы** принципы, методы и подходы теории вероятностей и математической статистики, ассоциативного и причинного анализа, машинного обучения и теории графов, теории частично упорядоченных множеств и решеток, анализа формальных понятий, кластерного анализ;

теория (основные теоретические положения) согласована с выводами по результатам экспериментальной проверки предложенных моделей и алгоритмов, выполненной при помощи разработанной программной библиотеки;

установлено: теоретические и экспериментальные результаты не противоречат результатам других исследований в данной области, представленных в независимых источниках;

основные научные результаты **апробированы** в печатных трудах и докладах на высокорейтинговых российских и международных научных конференциях.

Личный вклад соискателя состоит в:

– критическом анализе современного состояния исследований в области ассоциативно-причинной классификации;

– теоретическом и экспериментальном обосновании и выборе численной меры оценки «силы» причинной зависимости между атрибутами данных;

– разработке алгоритма поиска причинных зависимостей между атрибутами данных и его практической реализации в области рекомендательных систем;

– непосредственном участии в разработке методики извлечения понятий из текстовых данных и построении семантической модели данных;

– разработке алгоритма минимизации мощности множества причинных правил в модели ассоциативно-причинной классификации;

– разработке программной библиотеки, реализующий все предложенные модели и алгоритмы обработки больших данных;

– экспериментальном исследовании разработанных алгоритмов;

– подготовке основных публикаций и докладов по выполненной работе.

Диссертационный совет считает, что в диссертационном исследовании Тушкановой О.Н. решена актуальная научная задача разработки алгоритмов обработки больших данных для принятия решений на конечном множестве альтернатив на основе ассоциативно-причинной классификации и их реализация в форме программного прототипа. Результаты диссертационного исследования вносят вклад в развитие направления актуальной задачи вычислительно эффективной и устойчивой обработки больших данных в классе задач ассоциативно-причинной классификации и автоматического построения семантических моделей данных. В работе представлены новые научно обоснованные разработки в указанной области.

На заседании 29.09.2016 г. диссертационный совет принял решение присудить Тушкановой О.Н. ученую степень кандидата технических наук.

При проведении тайного голосования диссертационный совет в количестве 22 человека, из них 6 докторов наук по специальности рассматриваемой диссертации, участвовавших в заседании, из 26 человек, входящих в состав совета, проголосовали: за 22, против нет, недействительных бюллетеней нет.

Председатель диссертационного совета
доктор технических наук,
член-корреспондент РАН Юсупов Рафаэль Мидхатович

Ученый секретарь диссертационного совета
кандидат технических наук, доцент Фаткиева Роза Равильевна

29.09.2016 г.