Федеральное государственное бюджетное учреждение науки

Санкт-Петербургский институт информатики и автоматизации Российской академии наук
(СПИИРАН)

199178, Санкт-Петербург, 14 линия, 39
Телефон: (812)328-33-11
Факс: (812)328-44-50
E-mail: spiiran@iias.spb.su
http://www.spiiras.nw.ru
ОКНО 04683303, ОГРН 1027800514411
ИНН/КПП 7801003920/780101001

«24» мая 2016 г. № 073-09/65/303

УТВЕРЖДАЮ
ректор СПИИРАН
корреспондент РАН

Р.М. Юсупов

«24» июня 2016 г.

ЗАКЛЮЧЕНИЕ
Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН)

Диссертация «Семантические структуры и причинные модели больших данных для принятия решений с приложением к рекомендательным системам» выполнена в лаборатории интеллектуальных систем Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук.

Тушканова Ольга Николаевна получила степень магистра техники и технологии по направлению «Системный анализ и управление» в Южном Федеральном университете, г. Ростов-на-Дону, в 2011 г.

В период подготовки диссертации Тушканова Ольга Николаевна обучалась в очной аспирантуре Федерального государственного бюджетного учреждения науки Санкт-Петербургском институте информатики и автоматизации Российской академии наук по специальности 05.13.01 – «Системный анализ, управление и обработка информации (технические системы)», которую с отличием окончила в декабре 2015 г.

Тушканова О.Н. является специалистом в области интеллектуального анализа данных, машинного обучения и технологий автоматизации разработки онтологий данных. Работает с современными языками и системами программирования, такими как Java, Python, C#.

В настоящее время Тушканова О.Н. работает в лаборатории интеллектуальных систем СПИИРАН в должности научного сотрудника.

Удостоверение о сдаче кандидатских экзаменов №6/196 выдано в 2016 году Федеральным государственным бюджетным учреждением науки Санкт-Петербургским институтом информатики и автоматизации Российской академии наук.

Научный руководитель — Городецкий Владимир Иванович, доктор технических наук, профессор, и.о. заведующего лабораторией интеллектуальных систем СПИИРАН.

По результатам рассмотрения диссертации «Семантические структуры и причинные модели больших данных для принятия решений с приложением к рекомендательным системам» принято следующее заключение:

Оценка выполненной соискателем работы

В диссертационной работе Тушкановой Ольги Николаевны сформулированы основные проблемы в области анализа больших данных. Проведенный анализ состояния исследований в области технологий семантических моделей больших данных с использованием онтологий выявил необходимость разработки новых автоматизированных или автоматических (в зависимости от типа приложения) методов и технологий построения онтологии больших данных. На основании проведенных исследований предложен новый масштабируемый алгоритм автоматического построения семантической модели больших данных и сама модель данных, представляющая метасвойства данных, их синтаксис и семантику в рамках единой структуры. В работе исследованы современные методы ассоциативной и ассоциативно-причинной классификации и их возможности применительно к большим данным. На основании этого исследования сделан вывод о необходимости разработки новых алгоритмов ассоциативно-причинной классификации для задач, решаемых с использованием больших данных. В работе обоснован выбор численной меры оценки «силы» причинной связи между атрибутами данных, в частности, показано, что в качестве такой меры целесообразно использовать коэффициент регрессии и/или меру Клозгена. Разработан алгоритм поиска причинных зависимостей между атрибутами данных, а также алгоритм минимизации размерности пространства причинных правил в модели ассоциативно-причинной классификации.

Проведена апробация предложенных в диссертационной работе моделей и алгоритмов в ряде научно-исследовательских работ, выполняемых СПИИРАН, а также работ по контрактам с зарубежными компаниями.

Актуальность и востребованность данной тематики обусловлена тем, что в настоящее время проблема обработки больших данных, в частности, извлечение знаний из них относится к числу наиболее актуальных проблем в области информационных технологий. Однако существующие методы и алгоритмы не отвечают потребностям специалистов в области обработки больших данных, поскольку большинство традиционных методов интеллектуального анализа данных не могут быть применены напрямую к анализу больших данных из-за вычислительной неустойчивости и/или вычислительной сложности.

Личное участие соискателя в получении результатов, изложенных в диссертации.

Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в 9 опубликованных работах. Подготовка к публикации полученных результатов проводилась совместно с соавторами, причем вклад диссертанта был основным.

Следующие результаты, представленные к защите, получены лично автором:

- теоретическое и экспериментальное обоснование семантически корректной и вычислительно эффективной численной меры оценки «силы» причинной связи между атрибутами данных (личный вклад 100%);
- алгоритм поиска причинных зависимостей между атрибутами данных, в частности, применительно к рекомендательным системам третьего поколения (личный вклад 100%);
- алгоритм автоматической генерации семантической модели больших данных (личный вклад 60%);
- семантическая модель данных, которая позволяет представлять мета-свойства данных, их синтаксис и семантику в рамках единой структуры (личный вклад 50%);
- алгоритм минимизации размерности пространства причинных правил в модели ассоциативно-причинной классификации (личный вклад 100%).

Достоверность результатов проведенных исследований.

Достоверность основных научных положений и результатов диссертационной работы обеспечены анализом состояния исследований в области ассоциативно-причинной классификации, интеллектуальной обработки больших данных и построения онтологий, теоретическим обоснованием основных результатов и их согласованностью с результатами экспериментальных исследований, а также апробацией в печатных трудах и докладах на высокорейтинговых российских и международных научных конференциях.

Основные научные положения диссертационного исследования были представлены и получили положительную оценку на следующих конференциях:

- международная конференция «The Tenth International Workshop on Agents and Data Mining Interaction» (г. Париж, 2014 г.);
- всероссийская научно-практическая конференция «Перспективные системы и задачи управления» (п. Домбай, 2015);
- международная конференция «Creativity in intelligent technologies and data science» (г. Волгоград, 2015);
- международная конференция «The 2015 IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015)» (г. Париж, 2015);
- объединенная международная конференция «The 2015 IEEE/WIC/ ACM International Conference on Web Intelligence (WI'15) and the 2015 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'15)» (г. Сингапур, 2015).

Научная новизна полученных результатов.

В ходе диссертационного исследования были получены новые результаты в области ассоциативно-причинного анализа и методов автоматического построения семантической модели данных, а именно:

1. Теоретически и экспериментально обоснована семантически корректная и вычислительно эффективная мера оценки «силы» причинной связи между атрибутами данных. Рекомендации по ее выбору основаны на обширном экспериментальном исследовании и построены на численных оценках.

2. Разработан алгоритм автоматической генерации семантической модели больших данных, ориентированной на построение и минимизацию множества причинных правил модели ассоциативно-причинной классификации. В его основу положена новая методика семантического анализа понятий. Алгоритм отличается тем, что построен как комбинация методов и средств автоматизированной генерации иерархии понятий онтологии данных и генерации двойственных формальных понятий, определяющих условия останова процесса генерации понятий.

3. Предложена новая семантическая модель данных, которая позволяет представлять мета-свойства данных, их синтаксис и семантику в рамках единой структуры. Эта структура строится с помощью разработанного алгоритма автоматической генерации семантической модели больших данных и обеспечивает эффективный поиск причинных зависимостей в данных для алгоритмов ассоциативно-причинной классификации.

4. Разработан масштабируемый и вычислительно эффективный алгоритм поиска множества причинных зависимостей между атрибутами данных, использующий семантическую модель данных.

5. Разработан алгоритм минимизации размерности пространства причинных правил в модели ассоциативно-причинной классификации, который позволяет устранить избыточность модели принятия решений, возможно, с некоторой потерей точности. Он использует механизм кластеризации множества правил на основании метода корреляционных плеяд и алгоритма разрезания графа на клики.

Практическая значимость результатов исследования.

Практическая значимость результатов работы определяется разработанным программным прототипом, реализующим базовые алгоритмы ассоциативно-причинной классификации на основе обнаружения причинных связей. Программный прототип реализован в форме библиотеки классов на языке Java и набора вспомогательных программ. Компоненты библиотеки с некоторыми модификациями могут быть использованы в программах, поддерживающих принятие решений на основе больших данных. Разработанные алгоритмы и программный прототип протестированы на наборе данных Amazon в приложении к обучению персонифицированных рекомендательных систем третьего поколения.

Основные результаты диссертационной работы использованы при выполнении следующих проектов:

- проект «Контекстно-управляемый ассоциативный и причинный анализ данных для принятия решений» ПФИ ОНИТ РАН «Интеллектуальные информационные технологии, системный анализ и автоматизация», (2013 - 2015 гг.);

- контракт «Многоагентные алгоритмы для кросс-доменных рекомендательных систем» с Московским подразделением Samsung Electronics – Samsung Research Center (2014 г.);

- проект «Алгоритм автоматического инкрементного обучения для улучшения распознавания табличных данных» с «EMC International Company» (2015 г.).

Специальность, которой соответствует диссертация.

Диссертационная работа Тушкановой О.Н. полностью соответствует требованиям, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук по специальности 05.13.01 – «Системный анализ, управление и обработка информации (технические системы)».

Полнота изложения материалов диссертации в работах, опубликованных соискателем.

Соискатель имеет 20 научных трудов. Основные положения диссертации опубликованы в 9 печатных работах, включая 3 публикации в рецензируемых научных изданиях из перечня ВАК Минобрнауки РФ: «Труды СПИИРАН», «Информационные технологии и вычислительные системы», также 4 работы в изданиях, индексируемых в международных базах данных Web of Science и Scopus.

Основные результаты диссертационного исследования изложены в следующих работах в необходимой полноте:

В изданиях, рекомендованных ВАК Минобрнауки РФ:

1. Тушканова О.Н. Экспериментальное исследование численных мер оценки ассоциативных и причинных связей в больших данных // Информационные технологии и вычислительные системы. 2015. №3. С. 16-25. (Личный вклад 100%)

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

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

4. Gorodetsky V., Samoylov, V., Tushkanova O. Agent-based customer profile learning in 3G recommending systems: ontology-driven multi-source cross-domain case // 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.,Zeng Y. Springer. 2015. Vol. 9145. pp. 12 – 25. (Личный вклад 40%)

5. Tushkanova O. Comparative Analysis of the Numerical Measures for Mining Associative and Causal Relationships in Big Data // Creativity in Intelligent, Technologies and Data Science. Communications in Computer and Information Science. Eds. Kravets A., Shcherbakov M., Kultsova M., Shabalina O. Springer. 2015. Vol. 535. pp 571-582. (Личный вклад 100%)

6. Tushkanova O., Gorodetsky V. Data-driven Semantic Concept Analysis for Automatic Actionable Ontology Design // Proc. of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2015, pp. 1-9. (Личный вклад 65%)

7. Gorodetsky V., Tushkanova O. Data-driven Semantic Concept Analysis for User Profile Learning in 3G Recommender Systems // Proc. of the IEEE WI-IAT'2015, Singapore. 2015. pp. 92 - 97. (Личный вклад 50%)

В других изданиях:

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

9. Тушканова О. Сравнительный анализ численных мер оценки ассоциативных и причинных связей в больших данных // Материалы 10-й Всероссийской научно-практической конференции «Перспективные системы и задачи управления» 6–10 апреля 2015 г., Домбай, т. 2. - Ростов-на-Дону: ЮФУ, 2015. С. 54-65. (Личный вклад 100%)

Диссертация «Семантические структуры и причинные модели больших данных для принятия решений с приложением к рекомендательным системам» Тушкановой Ольги Николаевны рекомендуется к защите на соискание ученой степени кандидата технических наук по специальности 05.13.01 — «Системный анализ, управление и обработка информации (технические системы)».

В диссертационном исследовании Тушкановой О.Н. решена актуальная научная задача разработки алгоритмов обработки больших данных для ассоциативно-причинной классификации и их реализация в форме программного прототипа. Результаты диссертационного исследования вносят вклад в развитие направления ассоциативно-причинной классификации и автоматического построения семантических моделей данных. В работе изложены новые научно обоснованные разработки в области анализа больших данных и принятия решений на больших данных.