ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

кандидата технических наук, доцента, заведующего кафедрой информатики и прикладной математики Федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

Муромцева Дмитрия Ильича

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

1. Актуальность темы

Работа посвящена разработке алгоритмов обучения и принятия решений в задачах классификации на основе семантических и причинных моделей больших данных. Актуальность работы обусловлена несколькими факторами. Во-первых, большинство традиционных методов интеллектуального анализа данных напрямую не могут быть применены для анализа больших данных либо вследствие вычислительной неустойчивости, либо вследствие вычислительной сложности. Во-вторых, имеет место гетерогенный характер больших данных: они могут содержать атрибуты разных типов и неструктурированные данные, например, тексты на естественном языке. Таким образом, задачи разработки вычислительно эффективных алгоритмов поиска ассоциативных правил классификации и построения механизмов принятия решений на основе этих правил пока не имеют удовлетворительных решений и остаются актуальными.

2. Степень обоснованности научных положений, выводов и рекомендаций

Обоснованность и достоверность полученных автором результатов подтверждается корректным применением методов корреляционного, ассоциативного и причинного анализа, машинного обучения, объединения решений распределенных классификаторов, теории графов, теории вероятностей, математической статистики, методов и средств онтологического моделирования, теории частично упорядоченных множеств и решеток, методов анализа формальных понятий, кластерного анализа, теории объектно-ориентированного программирования, а также тем, что в основе теоретической части диссертации лежат наработки, корректность которых была показана в предшествующих публикациях.

Кроме того, свидетельствами достоверности и обоснованности полученных результатов являются результаты вычислительных экспериментов, проведенных на тестовых примерах популярной онтологической базе знаний DBpedia.org и апробация положений работы на международных и всероссийских конференциях.

3. Оценка новизны и достоверности

Научную новизну работы обеспечивают следующие результаты:

- мера оценки силы причинной связи между атрибутами данных, выбранная и обоснованная в ходе теоретического и экспериментального исследования, которое подтвердило её точность и вычислительную эффективность;

- алгоритм автоматической генерации семантической модели данных. Алгоритм включает этап извлечения иерархии понятий из глобальной онтологии (DBpedia) и этап генерации двойственных им формальных понятий;

- семантическая модель больших данных, которая представляет собой структуру, описывающую синтаксис и семантику данных и содержит метаинформацию о них;

- алгоритм поиска причинных правил для ассоциативно-причинной классификации, использующий семантическую модель данных, продемонстрированный на примере построения профиля пользователя рекомендательной системы;

- алгоритм минимизации количества причинных правил в модели ассоциативно-причинной классификации, основанный на методах кластеризации.

Помимо этого в работе получены следующие дополнительные результаты в области рекомендательных систем также обладающие новизной:

- разработан новый алгоритм гибридной коллаборативной фильтрации, использующий семантический причинный профиль пользователя и новую метрику семантического сходства пользователей, вычисляемую на основе этих профилей;

- предложен алгоритм выработки кросс-доменных рекомендаций с помощью гибридной коллаборативной фильтрации.

Достоверность основных результатов диссертационного исследования подтверждается:

- развернутым критическим анализом состояния в области ассоциативно-причинного анализа и средств обработки больших данных;

- экспериментальной проверкой предложенных моделей и алгоритмов, подтвердившей их работоспособность;

- апробацией в печатных трудах и докладах на научных конференциях.

- положительными результатами внедрения.

4. Теоретическая и практическая значимость

Теоретическая значимость работы заключается в разработке алгоритма, названного семантическим анализом понятий и предназначенного для автоматического построения семантической модели данных; в разработке алгоритма поиска причинных связей между атрибутами данных, а также анализе и обосновании меры оценки причинных связей, которая используется в этом алгоритме. Соискателем обоснована возможность и эффективность применения предложенных алгоритмов при обработке больших данных.

Значимость результатов диссертационного исследования для практики определяется реализацией разработанных моделей и алгоритмов обработки больших данных в форме библиотеки взаимосвязанных Java-классов, работоспособность которых проверена экспериментально.

В работе также продемонстрировано, каким образом предложенные алгоритмы и модели могут быть использованы при разработке рекомендательных систем.

5. Апробация и внедрение полученных результатов

Основные научные положения диссертационной работы были доложены на пяти научных конференциях разного уровня и получили положительные отзывы. Соискателем было опубликовано 9 печатных работ по теме диссертации: 3 статьи в журналах «Труды СПИИРАН», «Информационные технологии и вычислительные системы», которые входят в перечень ВАК, и 4 статьи в изданиях, индексируемых в Web of Science и Scopus.

Основные научные и практические результаты диссертационного исследования использованы в работе по контрактам, что подтверждается актами внедрения.

6. Замечания по диссертации

1. Несколько разделов работы посвящены исследованию мер оценки причинных связей в данных, однако четкое определение термина «причинная связь» отсутствует.

2. Несмотря на то, что в диссертационной работе подчеркивается высокая вычислительная эффективность предложенных алгоритмов, результаты исследования вычислительной сложности в работе не представлены.

3. В работе отсутствует анализ лингвистических аспектов анализируемых данных. Средства NLP не являются универсальными инструментами, а работают лишь с определенными языками, для которых они разработаны. Из текста диссертации не ясно, для каких языков применимы разработанные алгоритмы автоматического построения онтологий.

4. Не достаточно полно проанализированы существующие онтологии и глобальные базы знаний. Например, нет упоминания таких популярных онтологий для исследуемой в работе предметной области, как MusicBrainz и YAGO, а также таких баз знаний, как Last.fm и WikiData. Использование в экспериментальной части работы лишь одной базы знаний DBpedia и соответствующей онтологии возможно, однако требуется обоснование, что разработанные методы и алгоритмы смогут работать и с другими источниками данных.

5. Диссертационное исследование фактически относится к проблематике жизненного цикла связных данных и предлагает новые методы и алгоритмы для отдельных его этапов. Однако в работе нет упоминания данной методологии, что несколько вырывает полученные диссертантом результаты из контекста других исследований, проводимых в данной области.

Однако перечисленные замечания не снижают общей положительно оценки диссертационной работы и не влияют на качество полученных в ходе исследования теоретических и практических результатов.

7. Заключение

Диссертационная работа Тушкановой Ольги Николаевны представляет собой завершенную научно-квалификационную работу, в которой решена актуальная научная задача разработки семантических и причинных моделей и

алгоритмов обработки больших данных на основе ассоциативно-причинной классификации. Разработанные модели и алгоритмы экспериментально проверены в приложении из области рекомендательных систем.

Новые научные результаты, полученные соискателем, имеют существенное значение для науки и практики и достаточно полно опубликованы в научных изданиях. Содержание автореферата соответствует основным результатам диссертационного исследования.

Диссертационная работа Тушкановой Ольги Николаевны соответствует паспорту специальности 05.13.01 – «Системный анализ, управление и обработка информации (технические системы)» и отвечает предъявляемым к кандидатским диссертациям критериям, установленным п. 9 Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013 № 842, , а ее автор заслуживает присуждения ученой степени кандидата технических наук.

Сведения о составителе отзыва:

официальный оппонент
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»

197101, г. Санкт-Петербург,
пр. Кронверкский, 49
+7 (812) 233-52-56
d.muromtsev@gmail.com