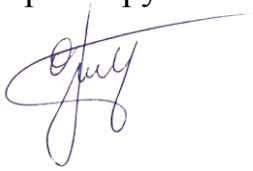
На правах рукописи



Тушканова Ольга Николаевна

Семантические структуры и причинные модели больших данных для принятия решений с приложением к рекомендательным системам

Специальность 05.13.01 – Системный анализ, управление и обработка информации (технические системы)

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2016

Работа выполнена в Федеральном государственном бюджетном учреждении науки Санкт-Петербургском институте информатики и автоматизации Российской академии наук (СПИИРАН).

Научный руководитель:	доктор технических наук, профессор Городецкий Владимир Иванович
Официальные оппоненты:	Гаврилова Татьяна Альбертовна доктор технических наук, профессор Институт «Высшая школа менеджмента Санкт-Петербургского государственного университета» (ВШМ СПбГУ) заведующая кафедрой информационных технологий в менеджменте
	Муромцев Дмитрий Ильич кандидат технических наук, доцент Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики» заведующий кафедрой информатики и прикладной математики
Ведущая организация:	Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)»

Защита состоится «29» сентября 2016 года в 14:30 на заседании диссертационного совета Д 002.199.01 при Федеральном государственном бюджетном учреждении науки Санкт-Петербургском институте информатики и автоматизации Российской академии наук (СПИИРАН) по адресу: 199178, Россия, Санкт-Петербург, 14-я линия, д. 39. Факс: (812) 328-44-50, тел: (812) 328-34-11.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации РАН www.spiiras.nw.ru.

Автореферат разослан « ___ » _____ 2016 г.

Ученый секретарь
диссертационного совета Д 002.199.01,
к.т.н., доцент — Фаткиева Роза Равильевна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. В настоящее время проблема обработки больших данных относится к числу наиболее актуальных в области информационных технологий, и она же порождает наиболее трудные проблемы алгоритмического характера, связанные с обеспечением точности, устойчивости и вычислительной эффективности процессов их обработки. Эти проблемы обусловлены тем, что большинство традиционных методов интеллектуального анализа данных напрямую не могут быть применены для анализа больших данных либо вследствие вычислительной неустойчивости, либо вследствие вычислительной сложности. Не менее трудные проблемы обусловлены гетерогенным характером больших данных: они могут содержать атрибуты разных типов и неструктурированные данные, например, тексты на естественном языке.

Одно из важных требований к методам обработки больших данных – это семантически ясная интерпретация результатов. В современных моделях знаний это обеспечивается средствами онтологии, однако ее построение для больших данных также является проблемой: ввиду огромного количества потенциальных понятий ручная разработка онтологии становится непомерно трудоемкой, а потому требует максимальной автоматизации, а в ряде классов приложений – полной автоматизации.

Анализ состояния исследований и разработок по сформулированным проблемам показывает, что существующие методы и алгоритмы обработки больших данных не отвечают ожиданиям и потребностям специалистов в этой области.

Обычно целью обработки больших данных является построение эмпирической модели целевых переменных, учитывающей некоторые атрибуты данных. При этом ключевым требованием является минимизация числа используемых атрибутов при условии обеспечения заданной точности модели. В данной работе задачи такого типа являются основным предметом исследований и разработок. Среди наиболее перспективных подходов к решению этой задачи в настоящее время выделяется подход, который базируется на обнаружении ассоциативных связей в данных и последующем использовании их в моделях принятия решений, в частности, в моделях ассоциативной классификации (далее АК). Однако 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.

В исследованиях К.Ф. Алайфериса (C.F. Aliferis) строго показано, что для принятия решений, в частности, для АК, среди ассоциативных связей наиболее информативными являются связи причинного характера и их структуры. Поиск причинных структур традиционно основан на построении и анализе байесовских сетей доверия. Этот подход требует решения задач обучения экспоненциальной сложности относительно размерности пространства атрибутов, и практически он может использоваться только тогда, когда размерность не превышает 20. Поэтому этот подход бесперспективен для причинного анализа больших данных. Важно отметить, что Я. Виттен (Y. Witten) подчеркивает: в байесов-

ской сети, построенной путем машинного обучения, далеко не все выявленные связи между переменными в реальности оказываются причинными. Эти обстоятельства требуют разработки альтернативных подходов к поиску причинных связей в больших данных. Такой альтернативой в настоящее время является ассоциативно-причинный анализ данных. Он базируется на использовании специальных мер оценки ассоциативной связи, которые позволяют из всего множества таких связей выделить те, которые носят причинный характер.

Целью работы является разработка алгоритмов обучения и принятия решений в задачах классификации на основе семантических и причинных моделей больших данных, их реализация в форме программного прототипа, а также экспериментальная оценка по таким характеристикам как масштабируемость, вычислительная эффективность и точность в задачах принятия решений, в частности, в рекомендательных системах.

В соответствии с поставленной целью в работе сформулированы и решены следующие частные **задачи исследования**:

1. Теоретически и экспериментально обоснованный выбор семантически корректной и вычислительно эффективной формальной меры, оценивающей «силу» причинной связи атрибутов данных.

2. Разработка алгоритма автоматической генерации семантической модели больших данных, понятия которой используются для представления знаний о данных и результатов обучения в форме причинных моделей для принятия решений на основе ассоциативно-причинной классификации (далее АПК).

3. Разработка единой структуры для представления синтаксиса и семантики больших данных, а также метаинформации о них. Структура должна обеспечивать доступ к данным и упрощать вычисления различных статистик.

4. Разработка математически корректного, масштабируемого и вычислительно эффективного алгоритма поиска причинных связей в больших данных.

5. Разработка масштабируемого алгоритма минимизации размерности причинной модели принятия решений и механизма АПК.

6. Экспериментальная оценка разработанных алгоритмов на стандартных наборах данных из области рекомендательных систем (далее РС) третьего поколения.

Результаты решения перечисленных задач – это компоненты многошагового алгоритма генерации семантических и причинных моделей больших данных для решения задач АПК. Их разработка, интеграция в рамках единого процесса, программное прототипирование и экспериментальное исследование с использованием стандартных наборов данных из области РС составляют содержание работы.

Степень теоретической разработанности темы исследования. Первые модели АК были предложены в работах Б. Лю, В. Ли и Х. Ен. Работы Х. Фана, Д. Ли, Л. Кобылинского, Р. Шерода по этой теме во многом способствовали более глубокому пониманию специфики задач АК и путей ее эффективной алгоритмизации. Семантическое моделирование данных с помощью онтологий описано в работах Гавриловой Т.А., Хорошевского В.Ф., Муромцева Д.И. Первые попытки использовать анализ формальных понятий при разработке онтологий

были предприняты в работах Т. Гу, П. Чимиано и Р. Година. Математические основы анализа формальных понятий подробно изложены в работе Б. Гантера. Основные результаты в области РС изложены в руководстве под редакцией Ф. Ричи, Л. Рокача, Б. Шейпира, П. Кантора.

Научной новизной обладают следующие результаты работы:

1. Теоретически и экспериментально обоснованная, 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. Алгоритм автоматической генерации семантической модели больших данных, ориентированной на построение и минимизацию множества причинных правил модели АПК. 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. Семантическая модель больших данных и структура для ее представления. Эта модель включает в себя метаинформацию о данных, их синтаксис и семантику, которые представлены в единой структуре. Эта структура состоит из иерархии понятий онтологии данных и иерархии соответствующих им двойственных формальных понятий, при этом каждое понятие обеих иерархий ссылается на множество примеров данных, составляющих его объем. Такая структура для представления семантической модели данных обеспечивает эффективный поиск причинных зависимостей в данных для алгоритмов АПК. Эта структура строится за один проход по данным, а все последующие вычисления не требуют дополнительного обращения к ним. В то ж время в ней представлены и данные сами по себе, что обеспечивает возможность инкрементного обогащения сематической модели данных при поступлении новых данных.

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

5. Алгоритм минимизации мощности множества причинных правил модели АПК путем устранения избыточности правил. Алгоритм отличается тем, что основан на кластеризации множества правил: каждый кластер содержит в себе сильно коррелированные правила, а правила из разных кластеров слабо коррелированы.

Теоретическая и практическая ценность работы заключается в разработке теоретически корректных и экспериментально проверенных алгоритмов обучения моделей АПК и принятия решений с ориентацией на задачи РС третьего поколения, реализованных в виде программной библиотеки. Программная библиотека включает множество Java-классов, построенных с учетом возможности повторного использования в широком круге задач обработки больших данных и принятия решений. Разработанные алгоритмы протестированы на данных из области персонифицированных, кросс-доменных РС. На основании результатов тестирования даны практические рекомендации по выбору алгоритмов и технологий для разработки современных РС третьего поколения.

Методы и средства исследования. В работе использовались 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, кластерного анализа. При разработке архитектуры программного комплекса использованы функционально- и объектно- ориентированные подходы.

Положения, выносимые на защиту:

1. Теоретически и экспериментально обоснованная мера оценки силы причинной связи обеспечивает вычислительную эффективность и масштабируемость алгоритмов поиска причинных связей в больших данных.

2. Вычислительно эффективный алгоритм автоматического построения онтологии позволяет построить семантическую модель больших данных, пригодную для обучения и минимизации причинных моделей в задачах АПК.

3. Семантическая модель больших данных представляет мета-свойства данных, их синтаксис и семантику в рамках единой структуры и обеспечивает доступ к большим данным в задачах обучения и принятия решений на основе АПК.

4. Алгоритм поиска причинных связей в данных является вычислительно эффективным и масштабируемым, что обеспечивается выбранной мерой оценки силы причинной связи и использованием семантической модели данных.

5. Алгоритм минимизации размерности пространства причинных правил модели АПК реализует снижение её избыточности и минимизацию размерности практически без потери точности.

Степень достоверности и апробация результатов. Достоверность основных результатов диссертации обеспечена анализом состояния исследований в данной области, теоретическим обоснованием результатов и их согласованностью с выводами по итогам экспериментальной проверки разработанных моделей и алгоритмов, а также апробацией основных теоретических и прикладных положений диссертации в печатных трудах и докладах на высокорейтинговых российских и международных научных конференциях. Основные результаты работы были представлены и получили положительную оценку на следующих конференциях: международная конференция «The 10th International Workshop on Agents and Data Mining Interaction» (г. Париж, 2014 г.), Всероссийская научно-практическая конференция «Перспективные системы и задачи управления» (п. Домбай, 2015 г.), международная конференция «Creativity in intelligent technologies and data science» (г. Волгоград, 2015 г.), международная конференция «The 2015 IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015)» (г. Париж, 2015 г.), 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 г.). Основные результаты диссертационной работы использованы в проектах «Контекстно-управляемый ассоциативный и причинный анализ данных для принятия реше-

ний» ПФИ ОНИТ РАН «Интеллектуальные информационные технологии, системный анализ и автоматизация», (2013-2015 гг.), «Алгоритм автоматического инкрементного обучения для улучшения распознавания табличных данных» с «EMC International Company» (2015 г.), а также при выполнении работ по контракту «Многоагентные алгоритмы для кросс-доменных рекомендательных систем» с Московским подразделением Samsung Electronics – Samsung Research Center (2014 г.), что подтверждено соответствующими актами внедрения.

Публикации. Основные положения диссертации опубликованы в 9 печатных работах, включая 3 публикации в рецензируемых научных изданиях из перечня ВАК («Труды СПИИРАН», «Информационные технологии и вычислительные системы»), а также 4 публикации в изданиях, индексируемых в международных базах данных Web of Science и Scopus.

Структура и объем работы. Диссертация изложена на 226 страницах машинописного текста, содержит 50 иллюстраций и 19 таблиц, состоит из введения, четырех глав, заключения, списка литературы (140 наименований) и четырех приложений.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертационной работы, определена её цель и сформулированы задачи исследований, отражены научная новизна и практическая значимость работы.

В первой главе выявлены особенности задачи построения моделей принятия решений на основе больших данных, проведен анализ состояния исследований и разработок и выделены основные проблемы в этой области. Даны краткие сведения о РС как о типовом приложении обработки больших данных, описана суть методов ассоциативного и причинного анализа больших данных в задачах принятия решений, а также обоснованы и детализированы цели и задачи исследования.

Проведенный анализ состояния исследований и разработок в области больших данных позволил выявить следующие основные проблемы:

1. Несоответствие существующих средств анализа данных потребностям в области анализа больших данных. Большинство традиционных статистических методов анализа базируется на выявлении и анализе связей между атрибутами данных, но они, как правило, не могут быть использованы для анализа больших данных из-за вычислительной сложности и неустойчивости.

2. Отсутствие эффективных методов автоматического построения онтологий больших данных. Ручное построение онтологии больших данных является крайне трудоемкой задачей для эксперта. Анализ состояния исследований показал, что известные подходы позволяют лишь частично автоматизировать технологию построения онтологий. Однако существует ряд важных классов приложений, в которых онтология данных должна строиться вообще без вмешательства экспертов. Примером является разработка РС, установленных на мобильных устройствах пользователей с соблюдением конфиденциальности данных.

3. Потребность в новых вычислительно эффективных и устойчивых методах и алгоритмах обработки больших данных, а также механизмах принятия решений на основе знаний, извлекаемых из них. Необходимость создания про-

граммных средств реализации таких алгоритмов, обеспечивающих баланс между эффективностью, вычислительной устойчивостью и точностью вычислений в задачах машинного обучения и принятия решений.

4. Необходимость разработки масштабируемых методов и алгоритмов поиска причинных зависимостей в данных, которые, как показал анализ, могут дать импульс развитию новых эффективных семантически корректных методов и алгоритмов обработки больших данных.

Обоснована возможность использования РС третьего поколения в качестве приложения, пригодного для тестирования и оценки разработанных алгоритмов обработки больших данных для принятия решений. Показано, что данные, используемые при разработке таких РС, относятся к классу больших данных, а требования к результатам их разработки типичны и для других приложений в области больших данных, что дает право переносить оценки и выводы, полученные для них, на общие случаи задач анализа больших данных.

Во второй главе выполнен детальный критический обзор основных подходов, методов и алгоритмов, предложенных в области АК и АПК, в том числе для обработки больших данных. Анализ результатов обзора позволил сделать вывод о необходимости разработки новых вычислительно эффективных алгоритмов поиска множества причинных зависимостей в больших данных. Такие алгоритмы должны использовать семантическую модель данных.

В настоящее время в качестве наиболее перспективной схемы поиска причинных связей в данных рассматривается та, в которой для целевой переменной находится множество атрибутов, связанных с нею наиболее «сильными» ассоциативными связями (для этого могут использоваться стандартные алгоритмы поиска ассоциативных связей), с последующим выделением из них подмножества причинных связей, например, с помощью статистической проверки гипотезы о том, является ли найденная ассоциативная связь с целевой переменной причинной связью (С. Бринь). В данной работе вместо статистической проверки этой гипотезы предлагается использовать численную меру оценки «силы» причинной связи.

Для того, чтобы из существующих численных мер оценки «силы» ассоциативной связи выбрать, те, которые являются также и мерами оценки «силы» причинной связи, сформулированы некоторые ключевые свойства, которыми должны обладать меры, пригодные для причинного анализа данных. В силу того, что причинная связь является направленной, мера причинной связи, должна быть некоммутативной, указывая направление связи. Обычно требуется, чтобы мера принимала значения в интервале [-1, 1] либо в интервале [0, 1]. И наконец, нулевое значение меры должно указывать на отсутствие причинной связи между ее аргументами. Сформулированные свойства использованы для предварительной фильтрации исходного множества мер, включающего все меры оценки силы ассоциативной связи, которые были найдены в литературе. Отобранные меры исследованы экспериментально с привлечением различных методологий их сравнения, в частности, для каждой меры

- на заданном тестовом наборе данных проведено сравнение найденных с помощью меры правил, с теми правилами, которые получены для этого же

набора данных другими исследователями в области причинного анализа данных;

- для заданного тестового набора данных построен простой классификатор, использующий правила, полученные с помощью исследуемой меры, и проведено сравнение по точности решения задачи классификации;

- получены внеконкурсные результаты соревнования «*Causality Challenge №1: Causation and Prediction*»[1] для классификатора, построенного с помощью исследуемой меры, и определено «место», которое этот результат занял бы в соревновании.

Анализ экспериментальных результатов показал, что с точки зрения поиска причинных связей в данных наиболее перспективными являются коэффициент регрессии, мера Клозгена, убеждение и фактор уверенности. Из них две наилучшие меры, а именно, *коэффициент регрессии*

$$R(A,B)=\frac{\left(p_{AB}-p_A\cdot p_B\right)}{p_A\cdot\left(1-p_A\right)} \tag{1}$$

и *мера Клозгена*

$$K(A,B)=\sqrt{p_{AB}}\cdot\left(p_{B|A}-p_B\right) \tag{2}$$

далее использованы в алгоритмах поиска причинных моделей больших данных при построения персонифицированного семантического профиля пользователя в РС третьего поколения. В формулах (1) и (2) A и B – это атрибуты данных булевого типа, интерпретируемые как случайные события, представленные выборкой больших данных, а p - выборочные вероятности атрибутов (случайных событий), указанных в нижнем индексе p.

Отметим, что вычисление значений обеих этих мер может быть выполнено за один проход по данным, что свидетельствует о вычислительной эффективности и является важным в контексте больших данных.

В третьей главе предложена новая методика автоматического построения семантической модели данных, названная семантическим анализом понятий, разработан алгоритм, реализующий её, а также новая семантическая модель представления данных, задающая их мета-свойства, синтаксис и семантику в рамках единой структуры.

Основная идея семантического анализа понятий (далее САП) – это совместное использование спецификации семантики больших данных с помощью иерархии понятий онтологии данных, которые извлекаются с помощью средств DBpedia и формулируются в терминах понятий естественного языка, и структуры формальных понятий, которая является основой анализа формальных понятий (далее АФП). Рассмотрим этот процесс более детально.

Известно, что иерархия понятий онтологии средствами DBpedia строится по принципу обобщения, т.е. «от примеров к понятиям (признакам) и от них – к более общим понятиям», а АФП строит формальные понятия по принципу специализации: «от примеров к формальным понятиям (признакам) и от них к бо-

[1] Causality Challenge №1: Causation and Prediction // URL: http://www.causality.inf.ethz.ch/ challenge.php

лее частным формальным понятиям». САП комбинирует оба эти процесса, выполняя их параллельно и итеративно.

Построение семантической модели данных с помощью САП рассмотрено применительно к разработке персонифицированного профиля интересов пользователя в РС третьего поколения, устанавливаемой на его мобильном устройстве. Эта задача является массовой (сколько мобильных устройств, столько и различных профилей интересов), и потому не может решаться с привлечением экспертов для каждого отдельного пользователя. Выборка данных для построения профиля содержит историю активности пользователя на мобильном устройстве, например, URL-адреса веб-сайтов, посещенных пользователем, для которых можно автоматически получить тексты соответствующих веб-страниц. Такая информация, хотя она исходно не структурирована, является хорошим источником знаний об интересах пользователя.

Алгоритм САП представлен псевдокодом на рисунке 1. На нулевом шаге (строки 3-4) множество текстов на естественном языке, представляющих интересы пользователя, обрабатывается с целью извлечения понятий онтологии данных, для чего в работе используется сервис DBpedia Spotlight. Этот процесс не требует вмешательства эксперта. Для каждого текста сервис возвращает набор URI статей Википедии, которые соответствуют понятиям онтологии, присутствующим в этом тексте. Объединение понятий, извлеченных из всех текстов, формирует множество понятий базового уровня искомой онтологии данных. В строках 5-7 каждому найденному базовому понятию ставится в соответствие подмножество текстов выборки, которые содержат это понятие, а также мощность этого подмножества. Значения мощности, деленные на общее число текстов в данных (примеров обучающей выборки), будут равны значениям эмпирических вероятностей появления соответствующего понятия в данных. После того, как построены понятия онтологии данных базового уровня, часть их отфильтровывается с помощью выбранной метрики «значимости» понятий (строка 8). Как правило, оптимальную стратегию фильтрации определяет конкретное приложение. В строках 9-19 представлены операции поочередного обобщения и специализации понятий и построение очередных уровней онтологии данных. При этом в строке 11 реализовано обобщение понятий онтологии данных, построенных на предыдущем шаге, с использованием DBpedia и иерархии категорий Википедии. Так как каждое базовое понятие соответствует URI статьи в Википедии, то для него можно получить категории Википедии, к которым принадлежит данная статья. URI этих категорий и их родительских категорий формируют верхние уровни разрабатываемой онтологии данных. Первый критерий фильтрации из тех, что представлены в строке 16, реализует следующее правило:

Правило 1. Из любой пары понятий онтологии данных, связанных отношением обобщения и имеющих одинаковый объем, в искомую онтологию данных включается только одно из них, предпочтительно меньшей степени обобщения.

Второй критерий (строка 16) соответствует второму правилу:

Правило 2. Если некоторое обобщенное понятие онтологии данных таково, что двойственный ему элемент структуры формальных понятий имеет пустой

объем, то это понятие онтологии данных, все тождественные ему по объему или обобщающие его понятия, в искомую онтологию данных не включаются.

Оба правила имеют строгое обоснование, приведенное в диссертации.

SemanticConceptAnalysis (B) :

Вход: выборка данных $\boldsymbol{B} = \{ B_j \}|_{j=1}^{m}$

Выход: двойственная структура $\mathfrak{R}$

Начало

1. $k = 1$;

2. $\boldsymbol{A_k} = \varnothing$; *// множество понятий первого уровня*

3. **для** всех экземпляров $B_j \in \boldsymbol{B}$:

4. $\boldsymbol{A_k} = \boldsymbol{A_k} + DBpediaSpotlightService\ (B_j)$; *// извлечь базовые понятия из текста*

5. **для** всех понятий $A_i^k \in \boldsymbol{A_k}$:

6. вычислить $\boldsymbol{B_i^k}$ и $\left|\boldsymbol{B_i^k}\right|$ *// множество экземпляров выборки, в которых встречается понятие A_i^k и его мощность*

7. $\hat{\boldsymbol{B}}_i^k = \boldsymbol{B_i^k}$; *// подмножества экземпляров базовых понятий и мощности «объёмов» двойственных им формальных понятий равны на 1 уровне*

8. $Filter(\boldsymbol{A_k}, \mathbf{Z})$; *// отфильтровать базовые понятия по критерию Z*

9. **пока** $\boldsymbol{A_k^{(\vee)}} \neq \varnothing$: *// критерий останова разработки структуры $\mathfrak{R}$, $\boldsymbol{A_k^{(\vee)}}$ - множество понятий онтологии данных на уровне k, связанных отношением $\vee$ («или»)*

10. **для** всех понятий $A_i^{(\vee),k} \in \boldsymbol{A_k^{(\vee)}}$

11. $\boldsymbol{A_{k+1}^{(\vee)}} = \boldsymbol{A_{k+1}^{(\vee)}} + DBpedia(A_i^{(\vee),k})$; *// добавить обобщенные понятия онтологии*

12. **для** всех понятий $A_i^{(\vee),k+1} \in \boldsymbol{A_{k+1}^{(\vee)}}$:

13. вычислить $\boldsymbol{B_i^{(\vee),k+1}}$ и $\left|\boldsymbol{B_i^{(\vee),k+1}}\right|$;

14. построить $A_i^{(\wedge),k+1}$; *// двойственные формальные понятия ($\wedge$ – «и»)*

15. вычислить $\boldsymbol{B_i^{(\wedge),k+1}}$ и $\left|\boldsymbol{B_i^{(\wedge),k+1}}\right|$; *// «объем» и мощность для двойств. понятий*

16. **если** $F_1\ (A_i^{(\vee),k+1})$ == *ложь* **или** $F_2\ (A_i^{(\vee),k+1})$ == *ложь* *//крит-и фильтрации*

17. **то** удалить понятия $A_i^{(\vee),k+1}$ из $\boldsymbol{A_{k+1}^{(\vee)}}$, и $A_i^{(\wedge),k+1}$ из $\boldsymbol{A_{k+1}^{(\wedge)}}$

18. добавить $\boldsymbol{A_{k+1}^{(\vee)}}$ и $\boldsymbol{A_{k+1}^{(\wedge)}}$ в $\mathfrak{R}$;

19. *k = k+1;*

20. **вернуть** $\mathfrak{R}$;

Конец.

Рисунок 1 – Псевдокод алгоритма семантического анализа понятий

Критерий останова (строка 9) процесса построения семантической модели данных формулируется следующим образом: остановить процесс обобщения понятий онтологии данных в том случае, если на текущей итерации с помощью двух правил фильтрации были удалены все вновь сгенерированные понятия.

Алгоритм автоматического построения семантической модели данных с помощью САП описанный в псевдокоде на рисунке 1, возвращает структуру, состоящую из двух полурешеток, одна из которых соответствует иерархии понятий онтологии данных, другая – структуре формальных понятий. Такая ком-

бинированная структура названа семантической структурой понятий. Она содержит информацию о семантике данных, информацию о структуре формальных понятий этих данных, т.е. отражает синтаксис данных (эта информация используется для построения механизма принятия решений), а также вероятностную метаинформацию, которая далее используется для построения причинной модели данных для конкретных приложений без дополнительного прохода по ним. На рисунке 2 представлена схема разработанного программного обеспечения, реализующего САП.

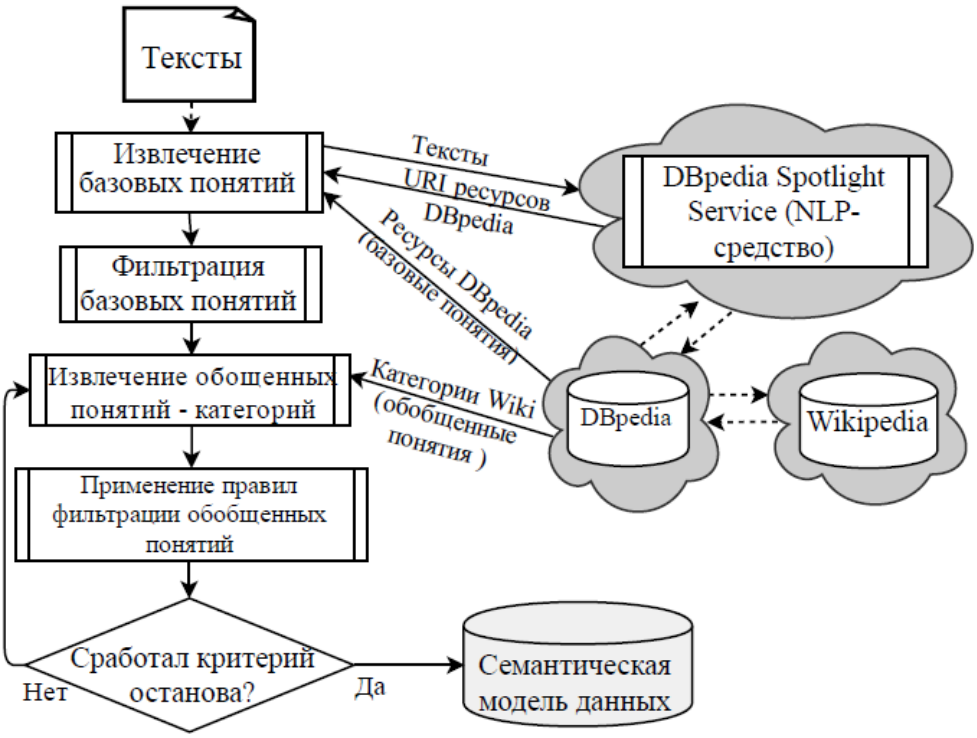


Рисунок 2 – Внешние компоненты и блок-схема алгоритма семантического анализа понятий

Работоспособность САП проверена на примерах из экспериментального набора данных Amazon. Набор содержит данные о большом количестве разнородных продуктов (фильмы, книги, музыка т.д.), которые продаются на сайте Amazon, что позволяет проверить возможности алгоритма САП при построении семантической модели, объединяющей знания из различных предметных областей. Кроме того, информация о каждом продукте может быть обогащена с помощью различных облачных ресурсов, таких как Википедия, DBpedia и др. Это дает основание говорить об этом наборе данных как о больших данных, так как возможности такого обогащения практически безграничны. Набор также содержит отзывы пользователей Amazon на купленные ими продукты. Такая информация хорошо подходит для извлечения интересов пользователя с использованием семантической модели данных, построенной с помощью САП. Эти особенности и повлияли на выбор набора данных Amazon в качестве тестового.

В четвертой главе рассматривается задача построения причинной модели данных и описываются алгоритмы построения этой модели. В качестве тестового приложения в работе используется задача построения причинной модели семантического профиля интересов пользователя РС третьего поколения для

данных Amazon. Демонстрируются разработанные алгоритмы принятия решений, а именно, алгоритмы выработки рекомендаций, а также результаты комплексного экспериментального исследования разработанных алгоритмов.

Задача построения семантического профиля пользователя РС третьего поколения формулируется следующим образом. Пусть имеется некоторое множество экземпляров продуктов (фильмы, книги, музыка т.п.), которые целевой пользователь ранее оценил в некоторой линейно упорядоченной шкале. Каждый экземпляр продукта из этого множества описывается набором атрибутов. В частности, для выбранного экспериментального набора данных Amazon такими атрибутами являются внутренние категории Amazon, к которым относится экземпляр продукта, и понятия онтологии данных, построенной с помощью САП для каждого пользователя.

В ходе построения профиля пользователя РС третьего поколения необходимо определить причины, которые побудили пользователя выставить конкретному экземпляру продукта тот или иной рейтинг. В качестве причин могут выступать различные свойства экземпляра продукта, представленные значениями атрибутов. Необходимо найти минимальный набор таких причин, который наилучшим образом характеризует интересы (мотивацию выбора) пользователя и, таким образом, описывает его профиль. В профиле пользователя РС эти причины представляются множеством правил вида

$$P_k^i\left(x_j^i \in \tilde{\boldsymbol{X}}_s^i\right) \rightarrow \omega_k,\ k = 1,\ldots,q, \qquad (3)$$

где $\tilde{\boldsymbol{X}}_s^i$ - подмножество значений некоторого атрибута X^i продукта, сформированное для целевого пользователя и отражающее его интересы; ω_k – рейтинг продукта, а предикаты вида $P_k^i\left(x_j^i \in \tilde{\boldsymbol{X}}_s^i\right)$ принимают истинное значение для заданного продукта тогда, когда значение x_j^i атрибута X^i для этого продукта принадлежат подмножеству $\tilde{\boldsymbol{X}}_s^i$.

Отметим, что главной задачей РС является предсказание отношения пользователя к некоторому новому для него продукту. В случае набора данных Amazon основная задача РС – это предсказание рейтинга $\boldsymbol{\omega} \in \{1, 2, 3, 4, 5\}$ (пятибалльная упорядоченная шкала), который целевой пользователь присвоит новому для него продукту. Решение о рейтинге должно приниматься на основе характеристик продукта и профиля пользователя, представляющего его интересы в форме правил вида (3). При этом дискретное значение рейтинга, выставляемое продуктам, выступает в роли класса $\boldsymbol{\omega}$. Тогда правила вида (3) являются ассоциативно-причинными и выступают в роли элементарных решателей, а задача предсказания рейтинга формулируется как задача ассоциативно-причинной многоклассовой классификации продуктов относительно предпочтений конкретного пользователя. В работе для построения модели выработки рекомендаций используется метод бинарного дерева решений. В этом случае для каждого узла принятия решений бинарного дерева необходимо сформировать множество ассоциативно-причинных правил вида (3). Совокупность всех правил для каждого из узлов будет являться одновременно и формальным представлением семантического профиля пользователя РС, и моделью принятия решений для

выработки рекомендаций. Таким образом разработанная модель знаний представляет знания для действий (англ. *actionable knowledge*).

Процедура формирования подмножеств значений атрибутов $\tilde{\boldsymbol{X}}_s^i$, использует идею агрегирования, предложенную В.И. Городецким и В.В. Самойловым[2]. Агрегирование (сегодня в этом смысле чаще используется термин грануляция) реализуется объединением отдельных значений атрибутов в агрегаты, которые обладают некоторым общим свойством по отношению к решаемой задаче классификации. При агрегировании область значений $\tilde{\boldsymbol{X}}^i$ каждого атрибута $\boldsymbol{X}^i$, разделяется на три подобласти $\tilde{\boldsymbol{X}}_s^i \subseteq \tilde{\boldsymbol{X}}^i$, *s = 1, 2, 3*. К первой из них относятся те значения, которые чаще встречаются в примерах одного из двух альтернативных классов задачи классификации. Ко второй области относятся те значения, которые чаще встречаются в примерах другого класса. Третья область содержит те значения, которые встречаются в обоих классах примерно с одинаковой частотой. Формально области задаются пороговыми функциями. Они не содержат общих значений и $\bigcup_s \tilde{\boldsymbol{X}}_s^i = \tilde{\boldsymbol{X}}^i$. Некоторые из этих множеств могут быть пустыми. Множества $\tilde{\boldsymbol{X}}_s^i \subseteq \tilde{\boldsymbol{X}}^i$, *s = 1, 2,* определяют области истинности предикатов $P_k^i(x_j^i \in \tilde{\boldsymbol{X}}_1^i)$ и $P_k^i(x_j^i \in \tilde{\boldsymbol{X}}_2^i)$, которые ставятся в соответствие каждому узлу принятия решений бинарного дерева и играют роль простейших классификаторов, построенных с использованием ровно одного атрибута из всего множества атрибутов. Третья область значений $\tilde{\boldsymbol{X}}_3^i$ атрибута X^i далее не рассматривается. Эта процедура позволяет преодолеть трудности, связанные с большой размерностью задачи и гетерогенностью атрибутов. Построенные предикаты далее рассматриваются как компоненты нового редуцированного пространства атрибутов. При этом типы исходных атрибутов, значения которых подвергаются агрегированию, могут быть любыми: включать в себя тексты на естественном языке, номинальные, числовые или булевы данные, а также могут быть представлены понятиями онтологии данных, построенной в ходе САП, тогда как все новые атрибуты представляют собой булевы утверждения о свойствах атрибутов в форме одноместных предикатов. В работе процедура агрегирования впервые применена к неструктурированным текстовым данным и к понятиям семантической модели данных. Построенное множество правил задает бинарные ассоциативные связи атрибутов данных и классов состояний, введенных в данных. В этой процедуре не используются обычные меры оценки силы ассоциативной связи типа поддержки и уверенности. Такое преобразование данных рассматривается как первичная фильтрация множества потенциальных интересов пользователя.

Следующим шагом построения семантического профиля пользователя является выделение из полученного множества ассоциативных правил класса тех правил, которые имеют причинный характер. Каждому правилу вида (3), полученному в результате агрегирования для каждого узла дерева принятия реше-

[2] Городецкий В.И., Самойлов В.В. Контекстно–зависимое обучение для принятия решений // Труды 2-й Междунар. конф.«Автоматизация управления и интеллектуальные системы и среды», 2011 г.

ний, ставится в соответствие значение одной из мер (1), (2) оценки «силы» причинной связи. Значение меры $\mu(P_k^i, \omega_k)$, полученное для каждого правила, рассматривается как его вес. Далее правила могут быть отсортированы в порядке убывания веса. Сортировка выполняется для каждого узла принятия решений и для правил каждого класса в этом узле отдельно. Фильтрация правил профиля выполняется следующим образом: в профиле пользователя в каждом узле принятия решений остаются только правила, для которых значение модуля меры $|\mu(P_k^i, \omega_k)|$ больше или равно некоторому порогу $\delta_{\min}^{\mu}$. Значение $\delta_{\min}^{\mu}$ выбирается экспериментально, исходя из мощности множества правил, среднего значения модуля меры причинности для правил в каждом узле и требований к точности решения задачи классификации.

В результате фильтрации в семантическом профиле пользователя РС останутся только те правила, которые соответствуют самым «сильным» его интересам - причинам, побуждающим его выставлять тот или иной рейтинг продуктам. Такая фильтрация позволяет значительно сократить размерность модели данных и тем самым улучшить вычислительную эффективность последующих процедур обучения без заметной потери точности принятия решений.

С целью дальнейшей оптимизации модели данных для принятия решений проводится анализ зависимости причинных правил и удаление избыточных правил. Для этого разработан подход, основанный на кластеризации причинных правил. Кластеры правил формируются таким образом, чтобы сильно коррелированные правила попадали в общий кластер, а слабо коррелированные – в разные кластеры. Кластеризация выполняется отдельно для подмножества правил каждого класса в узлах принятия решений. Коэффициент корреляции для пары причинных правил вычисляется по формуле:

$$Cor\left(P_k^i, P_k^j\right) = \frac{p\left(P_k^i P_k^j\right) - p\left(P_k^i\right) \cdot p\left(P_k^j\right)}{\sqrt{p\left(P_k^i\right) \cdot \left(1 - p\left(P_k^i\right)\right) \cdot p\left(P_k^j\right) \cdot \left(1 - p\left(P_k^j\right)\right)}}. \quad (4)$$

Отметим, что все оценки вероятностей, фигурирующие в формуле, либо уже были вычислены для каждого правила на предыдущих шагах процедуры построения семантического профиля пользователя, либо могут быть вычислены на основе имеющихся метаданных, и поэтому для подсчета коэффициентов корреляции не требуется дополнительных проходов по данным. На рисунке 3 представлен псевдокод алгоритма кластеризации причинных правил и минимизации семантического профиля пользователя.

В результате применения процедур фильтрации в искомом семантическом профиле пользователя в каждом узле принятия решений останутся только причинные правила, представляющие его интересы, которые слабо коррелируют друг с другом, отражая тем самым разнообразие этих интересов. На рисунке 4 представлена схема, поясняющая разработанную технологию построения семантического профиля пользователя.

Выделим достоинства предложенной модели семантического профиля пользователя в РС третьего поколения:

1. Предложенная формальная модель реализует персонализацию профиля в пользователя в терминах его интересов (мотивации выбора).

2. Модель компактно и однозначно интерпретируема с семантической точки зрения: каждое правило представляет отдельный интерес пользователя, выраженный понятием онтологии данных или подмножеством значений некоторого атрибута.

3. Профили различных пользователей представлены в терминах одних и тех же понятий онтологии данных, даже если первичная информация о пользователях была получена из различных источников, и поэтому они легко сопоставимы с помощью семантической меры сходства (это важно в случае применения алгоритмов коллаборативой фильтрации).

4. Профиль пользователя представлен в форме, хорошо применимой на практике: механизм принятия решений, а именно, дерево решений, встроен в профиль пользователя. Необходимо лишь выбрать модель слияния решений элементарных классификаторов в каждом узле дерева.

RulesClustering ($\boldsymbol{Pr}_{U_t}$):

Вход: семантический профиль пользователя $\boldsymbol{Pr}_{U_t}$ (множество правил класса);

Выход: минимизированный семантический профиль пользователя $\boldsymbol{Pr}_{U_t}^{\min}$;

Начало

1. $\boldsymbol{Pr}_{U_t}^{\min} = \varnothing$;

2. **для** всех узлов $(\omega_a, \omega_{\bar{a}})$ дерева решений $\boldsymbol{DT}$ профиля $\boldsymbol{Pr}_{U_t}$:

3. **для** $\omega_k \in \{\omega_a, \omega_{\bar{a}}\}$:

4. $\mathbf{Cor}_k = \left\| Cor_k^{i,j} \right\|_{i,j}$; *// матрица корреляции правил класса в узле дерева решений*

5. **для** всех пар правил ($\boldsymbol{P}_k^i(x_j^i \in \tilde{\boldsymbol{X}}_k^i) \to \omega_k$, $\boldsymbol{P}_k^j(x_j^i \in \tilde{\boldsymbol{X}}_k^j) \to \omega_k$) в текущем узле:

6. $\mathbf{Cor}_k[i][j] = Cor\left(P_k^i, P_k^j\right)$ *// коэффициент корреляции между правилами*

7. $\boldsymbol{T_k} = SelectThreshold\ (\mathbf{Cor}_k)$ *// выбор порога разделения кластеров*

8. **для** всех $\mathbf{Cor}_k[i][j]$

9. **если** $1 - |\mathbf{Cor}_k[i][j]| < \boldsymbol{T_k}$

10. **то** $\mathbf{Cor}_k[i][j] = 0$;

11. $\boldsymbol{\Omega_k} = DepthFirstSearch\ (\mathbf{Cor}_k)$; *// выделение связных компонент графа с помощью поиска в глубину; возвращает множество кластеров*

12. **для** всех кластеров $\Omega_k^j \in \boldsymbol{\Omega_k}$:

13. Sort(Ω_k^j); *// отсортировать правила в кластере*

14. $\boldsymbol{Pr}_{U_t}^{\min} = \boldsymbol{Pr}_{U_t}^{\min} + \Omega_k^1$; *// добавить в итоговый профиль первое по порядку правило кластера*

15. **вернуть** $\boldsymbol{Pr}_{U_t}^{\min}$;

Конец.

Рисунок 3 - Псевдокод алгоритма кластеризации причинных правил и минимизации семантического профиля пользователя

Наиболее простым и естественным способом выработки рекомендаций на основе разработанного семантического профиля интересов пользователя является метод фильтрации контента, реализованный с помощью АПК. Задача выработки рекомендаций в этом случае предполагает предсказание рейтинга, который целевой пользователь поставит некоторому новому для него продукту на

основе семантического профиля его интересов и характеристик продукта. Для каждого продукта при этом используются те правила классификации из профиля пользователя, которые «сработают» на нём. С помощью процедуры простого или взвешенного (в качестве веса правила выступает значение меры $\mu(P_k^i, \omega_k)$) голосования элементарных бинарных классификаторов, представленных правилами, в каждом узле дерева можно принять решение, к какому из двух альтернативных классов следует отнести продукт в этом узле: решение принимается в пользу класса, набравшего больше голосов. Процедура голосования применяется в каждом узле дерева принятия решений от корня дерева к его листьям. На выходе будет получен предсказанный класс продукта, соответствующий рейтингу, который, предположительно, целевой пользователь присвоит новому продукту.

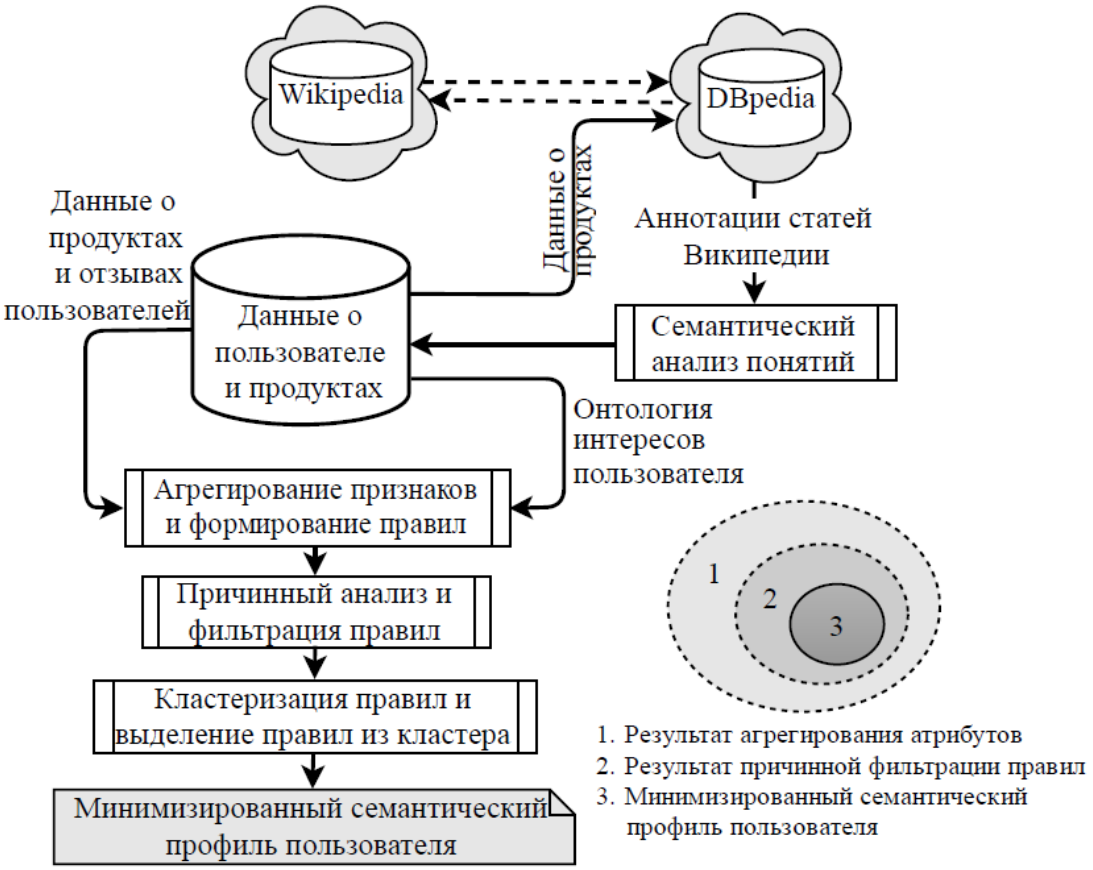


Рисунок 4 – Построение семантического профиля пользователя

Дополнительно в работе предложен алгоритм выработки рекомендаций с помощью коллаборативной фильтрации, который использует семантический профиль интересов пользователя и разработанную семантическую метрику сходства пользователей, вычисляемую на основе этого профиля. Предложенный способ можно отнести к гибридным методам коллаборативной фильтрации. Помимо этого, предложен способ выработки кросс-доменных рекомендаций на основе семантического профиля интересов пользователя с использованием метода коллаборативной фильтрации. Подчеркнем, что семантический профиль пользователя позволяет представить его интересы из разных предметных областей в рамках одной семантической модели, благодаря использованию САП.

Точность всех разработанных алгоритмов и подходов экспериментально оценивалась тестированием на наборе данных Amazon. Некоторые экспериментальные результаты приведены в таблице 1. В строках таблицы представлены предложенные методы выработки рекомендаций, в столбцах – показатели точности этих алгоритмов: RMSE (средняя квадратичная ошибка, англ. *root mean square error*), TPR (чувствительность классификатора, англ. *true positive rate, sensitivity*) и FPR (уровень значимости классификатора, англ. *false positive rate, fall-out*). Для задач выработки рекомендаций, именно эти показатели являются наиболее важными.

Таблица 1 – Результаты экспериментальных исследований алгоритмов на тестовом наборе данных Amazon

Методы / Показатели	RMSE	TPR	FPR
Фильтрация контента без применения САП	0,36	0,81	0,18
Фильтрация контента с применением САП (мера Клозгена)	0,41	0,89	0,28
Фильтрация контента с применением САП (коэфф. регрессии)	0,35	0,86	0,14
Коллаборативная фильтрация с использованием разработанной семантической метрики сходства пользователей	0,39	0,87	0,2

Отметим, что требования к точности разработанных алгоритмов определяются конечным прикладным приложением. Точность алгоритмов принятия решений может быть улучшена путём расширения обучающей выборки, либо с помощью увеличения количества правил в минимизированном семантическом профиле пользователя, т.е. путем повышения размерности итоговой модели принятия решений. Характеристики точности, приведенные в таблица 1, удовлетворяют требованиям заказчика, в интересах которого выполнялись некоторые разработки по тематике РС третьего поколения, представленные выше.

В заключении представлена итоговая оценка проделанной работы и приведены основные результаты исследования.

В приложениях представлены результаты экспериментальных исследований численных мер оценки «силы» причинных связей в данных, исходные данные и результаты экспериментов, демонстрирующие работоспособность алгоритма САП, псевдокоды всех разработанных алгоритмов и акты внедрения результатов исследования.

ЗАКЛЮЧЕНИЕ

В диссертационной работе решена актуальная научная задача - разработка алгоритмов обработки больших данных для построения модели АПК и её реализация в форме программного прототипа, а также выполнено экспериментальное исследование этих алгоритмов для конкретного приложения - РС третьего поколения, - в том числе, получены следующие результаты:

1. Теоретически и экспериментально обоснован выбор семантически корректной и вычислительно эффективной меры оценки «силы» причинной связи атрибутов данных. На основании анализа экспериментальных результатов сделан вывод о том, что наиболее перспективными с точки зрения способности выявлять правила, представляющие причинные связи в данных, являются ко-

эффициент регрессии и мера Клозгена. Однако, дальнейшие исследования показали, что алгоритмы классификации, использующие коэффициент регрессии, показывают большую точность при обучении на выбранном тестовом наборе данных.

2. Разработан масштабируемый алгоритм автоматического построения семантической модели данных в задачах принятия решений. В его основу положена методика семантического анализа понятий. Новизна алгоритма и методики состоит в совместном использовании (в рамках одной структуры данных) спецификации семантики данных с помощью иерархии понятий онтологии данных, которые извлекаются с помощью средств DBpedia и формулируются в терминах понятий естественного языка, и структуры формальных понятий модели данных, рассматриваемой в АФП.

3. Предложена новая модель больших данных, названная семантической моделью данных, основу которой составляет иерархия понятий онтологии данных и двойственная ей иерархия формальных понятий. Для её представления использована единая структура, которая наряду с семантической компонентой (иерархией понятий онтологии) и синтаксической компонентой (иерархией формальных понятий модели данных) содержит также метаинформацию, которая используется для формирования причинной модели данных без дополнительного прохода по ним. Структура также обеспечивает быстрый доступ к данным и эффективность вычислений.

4. Разработан алгоритм поиска причинных связей в больших данных, например, атрибутов данных и целевой переменной (в случае РС - причинных связей интересов пользователя и дискретных значений рейтинга некоторого продукта). Алгоритм включает в себя агрегирование данных, в ходе которого данные преобразуются к виду утверждений о свойствах атрибутов в форме предикатов, т.е. к бинарной шкале измерения. Такие предикаты рассматриваются как посылки правил классификации. Далее все полученные правила подвергаются пороговой фильтрации с использованием значения меры «силы» причинной связи.

5. Разработан алгоритм минимизации пространства атрибутов для решения задач АПК. Механизм основан на методах кластерного анализа. Он позволяет устранять избыточные правила, упрощая алгоритм принятия решений.

6. Выполнено экспериментальное исследование разработанных алгоритмов. Алгоритмы продемонстрировали точность, удовлетворяющую требованиям, выдвинутым в рамках одного из проектов, на выбранном тестовом наборе. Точность алгоритмов может быть улучшена за счёт увеличения обучающей выборки и привлечения дополнительной информации, например, из DBpedia.

Полученные результаты соответствуют п. 4 «Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации», п. 5 «Разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений и обработки информации», паспорта специальности 05.13.01 - «Системный анализ, управление и обработка информации (технические системы)».

СПИСОК РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

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.

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

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

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.

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.

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.

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.

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.

9. Тушканова О.Н. 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.