

ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертацию

Шачнева Дмитрия Алексеевича

«Семантические модели классификации и анализа данных в больших информационно-аналитических системах»,
представленную на соискание учёной степени кандидата физико-математических наук по специальности 05.13.17 — теоретические основы информатики

Диссертация Д.А. Шачнева посвящена разработке и изучению математических моделей и алгоритмов решения одного специального класса задач распознавания образов – поиска экспертов в заданной предметной области. Предметная область описывается множеством понятий (ключевых слов), характерных для деятельности в этой области, на котором определена система вложенных друг в друга разделов (рубрик). Описание эксперта (портрет эксперта) представлено набором ключевых слов и рубрик, снабженных числовыми весами, указывающими на значимость элементов в описании. Наборами такого же вида определяются и поисковые запросы. Наиболее существенными особенностями рассматриваемой задачи являются следующие дополнительные факторы, которые принимаются во внимание при поиске экспертов:

- 1) при оценке соответствия произвольного эксперта предъявленным требованиям учитываются не только ключевые слова и рубрики в тематическом портрете эксперта и описании требования, но также и отношение схожести (подобия) между элементами этих описаний; числовая мера схожести определяется на основании некоторого заданного множества тематических портретов, образующих обучающую выборку;
- 2) построение тематического портрета эксперта должно проводиться автоматически на основании данных об экспертах, содержащихся в информационной системе, требований запроса, в котором указана существенность тех или иных типов данных, а также иерархии, которую образуют элементы описаний (ключевые слова и рубрики) в информационной системе;
- 3) степень значимости тех или иных элементов в информационной системе может изменяться в зависимости от того, насколько существенны эти элементы при поиске экспертов, удовлетворяющих заданным требованиям;
- 4) значимость элементов информационной системы (ключевых слов и рубрик) должна быть взаимно согласована в зависимости расположения этих элементов в иерархии.

Задача такого вида является чрезвычайно актуальной в области информатики; она возникает всякий раз, когда возникает необходимость проводить классификацию наборов разнородных структурированных взаимосвязанных данных. Эта задача является нетривиальной; ее решение требует 1) создания адекватной математической модели, в рамках которой возможна формальная постановка задачи, 2) разработки математических методов, позволяющих определить и вычислить необходимые параметры для проведения классификации (оценка степени сходства отдельных понятий и их значимости для проведения классификации, ранжирование наборов данных по степени их соответствия

предъявленному запросу, и др.). Диссертация Д.А. Шачнева, посвященная исследованию и решению этой задачи, соответствует паспорту специальности 05.13.17 — Теоретические основы информатики, прежде всего, в части п. 5: разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений, и п. 9: разработка новых интернет-технологий, включая средства поиска, анализа и фильтрации информации, средства приобретения знаний и создания онтологии, средства интеллектуализации бизнес-процессов.

Диссертация Д.А. Шачнева состоит из введения, пяти глав и заключения. Во введении автор обосновывает актуальность задачи, приводит краткий обзор ранее известных подходов к ее решению, анализирует их достоинства и недостатки, описывает цель исследований диссертационной работы, объясняет, какими методами предлагается решать данную задачу и в чем состоит новизна предложенного подхода к ее решению, и представляет выносимые на защиту научные положения. Первая глава диссертации более всего напоминает техническое задание к научно-исследовательской работе; в ней приведен обширный перечень требований к использованию данных, к функциональным возможностям проектируемой системы классификации данных, производительности работы, безопасности и пр. Эта глава играет важную роль: в дальнейшем автор, объясняя выбор того или иного варианта решения задач, возникающих при создании математической модели информационно-аналитической системы и сопутствующих этой модели алгоритмов, ссылается на те требования технического задания, которые обуславливают этот выбор. Последующие главы посвящены решению четырех основных задач, возникающих при разработке указанной системы.

Во второй главе введено понятие тематического портрета, описаны принципы оценки степени сходства отдельных элементов портрета (ключевых слов и рубрик), предложены формулы, численно выражающие степень сходства отдельных элементов тематических портретов и самих портретов, и проведена оценка вычислительной сложности оценки показателей сходства элементов тематических портретов на основе заданной обучающей выборки. Также в этой главе обсуждаются способы повышения эффективности вычислений показателей сходства, приведена программа решения этой задачи на языке SQL и результаты экспериментального исследования предложенной методики вычисления показателей сходства элементов тематических портретов.

В третьей главе рассматривается вопрос построения тематического портрета объекта в информационно-аналитической системе, приводится формализация понятия информационно-аналитической системы на основе графовой модели данных, описан метод построения тематических портретов и предложен алгоритм решения этой задачи. В этой главе также приведен пример применения предложенного алгоритма.

В четвертой главе разработана модель для составления правил вычисления показателей, учитывающих значимость элементов тематических портретов при решении задач поиска экспертов. Представлена математическая формализация

модели для вычисления коэффициентов значимости, проведена оценка вычислительной сложности применения правил подсчета показателей, описан способ использования графовой модели данных в системах управления базами данных (СУБД), работающих на основе реляционной модели и языка запросов SQL.

В пятой главе описан алгоритм поиска результатов деятельности и экспертов, его программная реализация, проведено экспериментальное исследование производительности построенной программной системы.

В заключении кратко описаны основные результаты диссертации.

Достоинства диссертации.

1. В работе предложено комплексное, завершенное и методологически цельное решение задачи поиска экспертов в заданной предметной области по их результатам деятельности и с учётом показателей этих результатов, включающее описание

- математических моделей тематического портрета (предметной области, результатов активности, эксперта) и информационно-аналитической системы,
- метрик схожести тематических портретов и отдельных составляющих их элементов,
- алгоритмов анализа данных, работающих в рамках этих моделей и с использованием указанных метрик,
- принципов и особенностей реализации предложенных алгоритмов в современных информационных системах управления базами данных,
- результатов экспериментальных исследований применения реализованных алгоритмов для решения рассматриваемой задачи.

2. Автору диссертации удалось разработать новую систему метрик для оценки степени схожести различных компонентов предложенных моделей. Предложенные метрики схожести согласуются с предъявляемыми к ним требованиями, эффективно вычислимы и подкреплены содержательными объяснениями.

3. Большой интерес представляет новый алгоритм поиска экспертов на основе данных об отдельных результатах их деятельности с использованием метрик схожести понятий, при помощи которых описываются требования, предъявляемые к экспертам и тематические портреты кандидатов.

4. Описания моделей, метрик и алгоритмов снабжены подходящими примерами, хорошо иллюстрирующими введенные математические конструкции.

Наиболее существенные недостатки диссертации.

1. Обзор ранее известных методов и подходов к решению задачи анализа данных с целью выявления экспертов, приведенный в вводной главе (стр. 9-10), недостаточно систематичен. Автор диссертации ограничивается преимущественно простым перечислением работ, посвященных изучению указанной задачи, с очень кратким описанием полученных результатов. Было бы правильнее более подробно описать основные принципы наиболее распространенных и развитых подходов к решению

рассматриваемой задачи и проанализировать их сильные и слабые стороны. Автор диссертации также недостаточно полно объяснил преимущества предложенный им концепции тематического портрета, основанной на графовом способе представления данных, по сравнению с использованием онтологий (стр. 29).

2. Введенная в разделе 2.2.1 система обозначений весов и коэффициентов имеет недостатки. Было бы правильнее провести нормировку весовых коэффициентов в компонентах тематических портретов и сразу ввести нужные обозначение функции схожести в формуле (2.4) вместо того, чтобы разделять одни и те же обозначения на «старые» и «новые».

3. Формулы (2.5) и (2.6) в разделе 2.2.2 были введены довольно произвольно, опираясь частично на содержательные (эвристические) соображения и частично на стремление сделать возможно более эффективным вычисления с использованием этих формул. Но вместо того, чтобы стремиться избегать естественных рекурсивных определений, было бы правильнее изучить поведение функций схожести, которые задаются рекурсивно, и оценить алгоритмические возможности их вычисления.

4. Формулировка Утверждения 2 в разделе 2.3 не является строгой и допускает двоякое толкование.

5. В оценка порядка возрастания функции, определяющей вероятность того, что два токена являются соседними, (стр. 42), не указано при каких зависимостях между параметрами k , m и n этой функции справедлива указанная оценка.

6. В Утверждении 3 (стр. 43) не указаны явно условия, которым должны удовлетворять числовые параметры k , m и n , для того чтобы были верны заявленные оценки сложности.

7. На стр. 60 автор диссертации выдвигает некоторые предположения о характере перераспределения веса токена при его передаче между вершинами информационно-аналитической системы («Будем исходить из того...исходному весу токена.»), но не дает никаких обоснований или пояснений для этих предположений.

8. В определениях 12-14 в разделе 4.1 автор некорректно обращается с математическими терминами, смешивая тип выражения и тип значений выражения, выражение и функцию, реализуемую этим выражением.

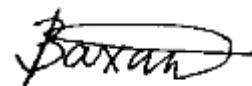
Эти замечания, впрочем, не являются критическими, поскольку отмеченные недостатки не нарушают целостность работы и не сказываются заметным образом на корректности полученных результатов.

Нужно отметить также что диссертация хорошо структурирована, изложение материала проведено ясно и подробно. Многочисленные примеры, которые сопровождают почти все определения и описания правил преобразования

программ, способствуют хорошему пониманию текста и в значительной мере проясняют смысл всех конструкций. Немногочисленные математические утверждения сопровождаются корректными доказательствами. Все основные результаты, полученные в диссертации, являются новыми.

В автореферате строго и полно описаны цели и задачи исследования, содержание работы и основные полученные результаты. Автореферат соответствует требованиям, предъявляемым к кандидатским диссертациям. Считаю, что диссертация Шачнева Дмитрия Алексеевича «Семантические модели классификации и анализа данных в больших информационно-аналитических системах», является законченным научно-квалификационным трудом и соответствует критериям, установленным в «Положении о присуждении ученых степеней в Московском государственном университете имени М.В. Ломоносова» по специальности 05.13.17 — теоретические основы информатики (физико-математические науки).

Профессор факультета вычислительной
математики и кибернетики МГУ имени
М. В. Ломоносова,
доктор физико-математических наук
(специальность 01.01.09 — дискретная
математика и математическая кибернетика)



Захаров
Владимир
Анатольевич

Сведения об организации:

Федеральное государственное бюджетное
образовательное учреждение высшего
образования «Московский государственный
университет имени М.В. Ломоносова»
119991, Москва, Ленинские горы, д. 1
Телефон: (495) 939-17-71
Факс: (495) 939-25-96
E-mail: zakh@cs.msu.su

19.10.2021