



Московский государственный университет
имени М.В. Ломоносова

Интеллектуальная система
тематического исследования
научно-технической информации
(«ИСТИНА»)

Под редакцией академика РАН В.А. Садовниченко

Издательство Московского университета
2014

УДК 002
ББК 73

Авторский коллектив под научным руководством академика В.А. Садовниченко: С.А. Афонин, А.В. Бахтин, В.Ю. Бухонов, В.А. Васенин, Г.М. Ганкин, А.Э. Гаспарянц, Д.Д. Голомазов, А.А. Иткес, А.С. Козицын, И.Н. Тумайкин, К.А. Шапченко.

Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА) / С.А. Афонин и др. Под ред. академика В.А. Садовниченко. – М.: Издательство Московского университета, 2014. – 262 с.

ISBN 978-5-19-011015-9

Настоящая книга посвящена описанию Интеллектуальной Системы Тематического Исследования Научно-технической информации («ИСТИНА»), целью которой является сбор, систематизация, хранение, анализ и выдача по запросу информации, характеризующей результаты деятельности научных и образовательных организаций. Представлено описание функциональных возможностей системы, включая автоматизированный ввод данных, отображение их как в открытом, так и в ограниченном доступе, автоматическое формирование отчетных материалов, статистический и тематический анализ. Рассмотрены возможные направления развития системы, в том числе в области использования онтологий для учета специфики предметных областей различных подразделений университета при анализе данных. Приведена краткая характеристика результатов внедрения системы в МГУ имени М.В. Ломоносова с привлечением более 10000 сотрудников из 880 подразделений.

ISBN 978-5-19-011015-9
Оформление обложки: Кс.П. Семёнова
© Коллектив авторов, 2014

Оглавление

От редактора	7
В.А. Садовничий	
Введение	11
В.А. Васенин	
1 Назначение и архитектура	15
С.А. Афонин, В.А. Васенин, Д.Д. Голомазов	
1.1 Назначение и функциональные требования	15
1.2 Аналогичные системы	21
1.3 Архитектура системы ИСТИНА	33
1.4 Отличительные особенности	42
2 Модули, реализующие основные функции	47
С.А. Афонин, В.А. Васенин, Г.М. Ганкин, Д.Д. Голомазов, А.С. Козицын, И.Н. Тумайкин	
2.1 Персонализация	47
2.2 Ввод данных	51
2.3 Отображение данных	62
2.4 Формирование отчетов	68
2.5 Статистика	73

2.5.1	Статистические данные публикационной активности	75
2.5.2	Статистика подразделения	81
2.6	Методы повышения качества данных	93
2.7	Взаимодействие с другими системами	100
2.7.1	Возможности импорта и экспорта данных	100
2.7.2	Сопоставление информационных объектов	101
2.8	Техническая реализация	103
2.8.1	Ядро системы	104
2.8.2	Кэширование	111
2.8.3	Профилирование	113
2.8.4	Поиск похожих объектов	115
2.8.5	Хранение данных	117
2.8.6	Автоматическое тестирование кода	124
3	Алгоритмы	135
	С.А. Афонин, В.А. Васенин, А.Э. Гаспарянц, Д.Д. Голомазов	
3.1	Использование онтологий	135
3.2	Разрешение неоднозначности авторства	140
3.2.1	Описание модели	142
3.2.2	Формализация критериев оценки	143
3.2.3	Определение функции оценки качества набора	146
3.2.4	Реализация и тестирование	147
3.3	Критерии результативности исследований	148
3.4	Анализ графа соавторов	156
3.4.1	Выделение ключевых узлов графа соавторов	157
3.4.2	Поиск устойчивых авторских коллективов	160
4	Внедрение в МГУ	165
	В.А. Васенин, Д.Д. Голомазов	

4.1	Поощрение авторов публикаций	166
4.2	Официальный сайт диссертационных советов	171
4.3	Механизм обратной связи	172
4.4	Общая статистика внедрения в МГУ	176
5	Информационная безопасность	179
	В.Ю. Бухонов, В.А. Васенин, А.А. Иткес, К.А. Шапченко	
5.1	Модели логического разграничения доступа	179
5.1.1	Модели, действующие в системе в настоящее время	180
5.1.2	Реляционная модель	182
5.1.3	Основные множества и сущности реляционной модели	183
5.1.4	Требования к подсистеме логического разграничения доступа к данным на основе реляционной модели и её основные атрибуты	216
5.1.5	Выразительные свойства реляционной модели	222
5.1.6	Алгоритмы, необходимые для работы макета .	227
5.1.7	Результаты тестовых испытаний	231
5.2	Механизмы защиты, встроенные в Django	233
5.2.1	Защита от межсайтового скриптинга	234
5.2.2	Защита от подделки межсайтового запроса . .	236
5.2.3	Защита от внедрения SQL	236
5.2.4	Защита от скликивания	237
5.2.5	Криптографическая подпись	237
6	Перспективы развития	239
	С.А. Афонин, А.В. Бахтин, В.А. Васенин, Д.Д. Голомазов	
6.1	Персонафицированный поиск	240
6.2	Онтологии	247

6.3	Актуальность тематических направлений	250
	Заключение	257
	В.А. Васенин	

От редактора

Настоящая коллективная монография представляет результаты поисковых исследований и выполненных на их основе работ прикладного характера, которые направлены на создание многоцелевого инструментария в области наукометрии.

Потребность общества и государства, отдельных организаций науки и образования в таком инструментарии созрела давно. Еще в начале 1990-х годов мы понимали необходимость иметь средства, которые помогали бы Московскому университету, его административно-управленческому персоналу оперативно получать сведения о текущем состоянии и трендах (направлениях эволюции) научных исследований, инновационной деятельности в МГУ, на основании анализа которых можно было бы принимать эффективные решения в области научной деятельности и образовательного процесса, планово-финансового характера, в сфере кадровой политики и на других направлениях. Однако отсутствие в те годы необходимых для решения этой задачи информационных технологий, средств телекоммуникаций не позволили нам такую систему создать.

В начале 2000-х годов, уже имея развитую на основе технологий Интернет сетевую инфраструктуру в Московском университете, мы вновь возвращались к этой задаче, анализировали подходы к ее решению. С тех пор интерес к этому направлению не только не исчезал, а, наоборот, от года к году возрастал. Появились новые потребности в подобной системе, которые были обусловлены нашим стремлением к адекватному, широкому и открытому освещению в Интернет

состояния научно-инновационной деятельности и образовательного процесса МГУ. Стало понятно, что такая информация, подробная и основная на сложившихся в мире наукометрических характеристиках и способах её интеллектуального анализа, способна активно влиять на мировые рейтинговые показатели университета, которые формируются различными зарубежными агентствами. Спокойно и взвешенно относясь к этим рейтингам, тем не менее, мы не считали возможным игнорировать их. Активный со стороны Министерства образования и науки РФ интерес к вопросам адекватной оценки деятельности российских вузов и институтов РАН, в том числе – научно-инновационной, также повлиял на повышение интереса к системам такого назначения.

Исследования и практические работы по построению моделей, созданию методов и программных механизмов целевой системы сначала выполнялись в инициативном порядке, затем они были включены в план научно-исследовательских работ. Полученные к 2013 г. результаты их выполнения показали нам положительными и имеющими хорошие инновационные перспективы. Как результат осмысления перечисленных выше мотивов и сложившихся факторов в Московском университете в 2013-2014 г.г. начал финансироваться проект «Развитие и сопровождение информационно-аналитической системы подготовки принятия решений на основе анализа информации о результатах научно-исследовательской, педагогической и инновационной деятельности «Наука МГУ» («ИСТИНА»).

В монографии представлены результаты выполнения работ по этому проекту. На настоящее время ИАС «Наука МГУ» («ИСТИНА») активно используется отдельными учеными, профессорско-преподавательским составом для представления вовне и строгого учета своих достижений. Система является эффективным средством для административно-управленческого аппарата, для анализа состояния дел с наукой как в университете в целом, так и в каждом его структурном подразделении. Такой анализ является основой для принятия эффективных управленческих решений на всех уровнях. Система конкурентоспособна и в

России, и в мире, что подтверждает уже проявленный интерес к ней со стороны российских вузов и НИИ, а также контакты с нашими зарубежными партнерами и разработчиками аналогичных систем. Мы уже задумываемся о более активном использовании методологии и инструментальных средств системы применительно к учебному процессу в Московском университете. Однако это ближайшее будущее.

Сегодня мы готовы предоставить все функциональные возможности и услуги ИАС «Наука МГУ» («ИСТИНА») внешним по отношению к МГУ организациям – вузам и НИИ России. Надеюсь, что представленные в монографии материалы будут интересны и полезны для широкого круга читателей – от отдельных, желающих широко представить свои результаты ученых, а также специалистов в области информатики и создания больших систем до административно-научного персонала российских вузов и институтов РАН.

Ректор МГУ, академик РАН
В.А. Садовничий



Введение

В последние 2-3 года потребность в создании информационно-аналитической системы подготовки принятия решений на основе анализа данных научно-инновационного и образовательного содержания настоятельно обозначается на всех уровнях национальной инфраструктуры управления образованием и наукой. Систему подобного назначения (далее для краткости изложения – Систему) хотели бы иметь как отдельные организации (вузы, НИИ, научно-производственные центры и т.п.), особенно крупные, претендующие на высокий национальный и международный рейтинги, так и субъекты управления наукой и образованием в регионах, Министерство образования и науки России. Эта потребность выражается в грантах и проектах различных фондов на эту тему, в нормативно-законодательных документах, выступлениях и комментариях в СМИ представителей руководящих структур разных уровней государственной власти. Основная посылка (мотив) такой потребности и интереса: создать Систему, инструментальные средства которой позволяли бы **оперативно** анализировать (по различным тематическим срезам и, соответственно, запросам) **надежно верифицируемые** данные как **персональные** и **обобщенные по отдельным научным коллективам и организациям** (вузам, НИИ и т.п.), так и данные, обобщенные **по регионам и по России в целом**, с целью получения с высокой степенью объективности **оценки показателей результативности и/или тенденции (тренды) научно-педагогической и инновационно-внедренческой деятельности**. В условиях социально-ориентированной эконо-

мики, в которой значительную роль играет государственный научно-образовательный сектор, такой инструментарий создает дополнительные механизмы стимулирующие её переход на инновационный путь развития.

Полученные с помощью Системы **оценки и результаты анализа трендов** научно-технической, образовательной и инновационной деятельности могут быть положены **в основу**: различного рода **отчетной документации; принятия управленческих** (финансово-экономических, кадровых, структурных и т.п.) **решений; определения рейтинговых показателей** субъектов подобной деятельности на всех уровнях. С учетом перечисленных требований критериями готовности целевой Системы к решению поставленных перед ней задач должны быть следующие:

1. **максимально точные** («очищенные» и верифицируемые) **в должном объеме** (по отношению к численности подлежащей анализу группы - организация, регион, страна в целом) данные о научно-педагогической и инновационно-внедренческой деятельности отдельных, участвующих в ней персон;
2. эти **данные**, в первую очередь, **должны включать**, как **сведения библиографического характера**, так и **сведения** (лучше на русском и английском языках) **аннотационного характера, включая ссылки** на другие источники;
3. для реализации требований к Системе в полном объеме **должны быть по возможности полно (в рамках авторских прав) представлены результаты деятельности персоналий в полнотекстовом виде.**

Критерий 2 позволяет проводить по базе данных Системы как аналитические исследования (выполнять запросы) по различным тематическим срезам, так и получать индексы цитирования. Эта информация является исходной для принятия управленческих решений.

Критерий 3 в представленном перечне:

- позволяет повысить рейтинг организации, т.к. объем полнотекстовых данных входит в такие показатели;
- обеспечивает рекламу и пропаганду деятельности организации вовне;
- повышает индекс цитирования как персоналий, так и организации, которую они представляют.

Решение всех перечисленных выше задач требует:

- больших ресурсоемких научных исследований, как фундаментального характера, так и прикладных;
- практических работ по развитию и сопровождению Системы;
- мер организационно-административного характера, способствующих (направленных на) вовлечению в работу максимально возможного (как можно большего) числа персоналий – ученых и преподавателей.

Представленные далее в настоящей работе материалы преследуют следующие цели:

- во-первых, они излагают позицию авторского коллектива по вопросам создания и развития систем подобного назначения в России, как в плане выполнения необходимых для этого научно-поисковых исследований, так и прикладных работ;
- во-вторых – иллюстрируют востребованность такой Системы и ее функциональные возможности, состояние и перспективы развития на примере конкретного программного комплекса – информационно-аналитической системы (ИАС) «Наука МГУ» («ИСТИНА»). В целях сокращения далее по тексту используется более широко известная аббревиатура названия этой системы «ИСТИНА» – Интеллектуальная Система Тематического Исследования НАучно-технической информации. Ее методологические, архитектурные и научно-технические решения положены в основу и реализованы в ИАС «Наука МГУ» («ИСТИНА»).

Схема и способы представления материалов данной публикации по замыслу авторов должны способствовать заинтересованному и понятному их восприятию различными категориями читателей и потенциальных пользователей систем такого назначения. Книга рассчитана не только на отдельных ученых и преподавателей, руководителей структурных подразделений (кафедр, лабораторий и т.п.) и организаций в целом (вузов, НИИ), на руководителей органов государственной власти в сфере управления наукой и образованием. В ней содержатся материалы, представляющие интерес для специалистов по информатике и информационным технологиям, которые могут включиться в работы по развитию этой или другой системы аналогичного назначения. Каждый из них может почерпнуть информацию, которая в большей степени касается его профессиональных интересов.

Прежде чем переходить к основному тексту монографии, хотелось бы отметить, что в ней в качестве иллюстрирующих материалов (в том числе – сведения, касающиеся отдельных персоналий) представлены только те данные, которые находятся в открытом доступе в Интернет, могут быть получены ответственными от подразделений или с помощью открытых для любого пользователя сервисов Системы.

Глава 1

Назначение и архитектура

1.1. Назначение и функциональные требования

Для улучшения работы вузов, научных организаций и, как следствие, для повышения эффективности развития науки и образования в масштабах государства необходимо постоянно анализировать информацию о результатах деятельности отдельных ученых и коллективов исследователей. Основными результатами деятельности организаций, входящих в научно-образовательное сообщество, как правило, считаются публикации сотрудников, результаты патентных исследований, участие в конференциях, руководство курсовыми, дипломными и диссертационными работами, чтение лекций и ряд других. При этом, как показывает практика, далеко не все результаты такой деятельности представлены в открытом доступе в Интернет. Зачастую единственным источником подобной информации могут служить лишь годовые отчеты сотрудников организаций науки и образования, представленные с той или иной степенью подробности. Естественным образом возникает необходимость автоматизированной (с участием человека) обработки данных из подобных научных отчетов в целях количественного и качественного анализа эффективности научной деятельности отдельного коллектива, вклада каждого его участника и возможной корректировки планов, мер и меропри-

ятий на основе такого анализа. Целями проведения такого анализа могут быть следующие.

- Сравнение данных о работе подразделения с данными других подразделений, в том числе – зарубежных, работающих на рассматриваемом направлении.
- Интеграция данных о работе отдельного подразделения и организации в целом с информацией других подразделений и организаций в целях анализа развития науки в рамках структур корпоративного, регионального масштаба и по стране в целом.
- Корректировка процессов финансирования подразделений, поощрения отдельных их членов на основе объективных результатов научных достижений.
- Публикация обзорных аналитических статей, посвященных научным достижениям организации с целью широкого и объективного освещения результатов ее деятельности.
- Получение интегрированной информации о направлении исследований в отдельной области знания, например, список основных публикаций, задач, методов, уровень активности ученых, ключевые персоны и конференции на этом направлении. Такая информация может представлять интерес для исследователя, которому необходимо получить первое, общее представление о еще недостаточно знакомом научном направлении.

Необходимо отметить, что перечисленные цели не исчерпывают всех функциональных возможностей, которыми может обладать такая система в будущем. Они должны и будут пополняться по мере развития системы, расширения используемых в ее составе методов, инструментальных средств и предоставляемых сервисов.

Инструментом аналитика, целью которого является получение адекватного представления о деятельности организации или объединения нескольких организаций, могут служить результаты выполнения аналитических запросов к системе, занимающейся загрузкой, обработкой и хранением информации о научной

1.1. НАЗНАЧЕНИЕ И ФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ

деятельности работающих в них сотрудников. Примером такого запроса может служить: *«Найти все публикации сотрудников интересующего учреждения за последний год, посвященные вопросам выделения информации из неструктурированных текстов и включенные в материалы международных конференций»*.

Побудительным мотивом и конечной целью исследований, результаты которых представлены далее, является создание интеллектуальной программной системы для анализа научной информации. Эти результаты были получены в 2012 г. на этапе предпроектных исследований, которые легли в основу созданной в ходе работ 2013/2014 гг. первой версии системы.

Исходными для решения рассматриваемой далее задачи являются представленные в виде электронной коллекции данные, характеризующие результаты деятельности научных организаций и вузов, в частности - сведения о публикациях, патентах, научных отчетах, докладах на конференциях, об участии в редколлегиях журналов, программных комитетах конференций, членстве в диссертационных советах, о руководстве диссертациями, авторстве учебных курсов и других видах научной, преподавательской и научно-организационной деятельности. Такая информация может храниться в различном виде, например, в форме годового отчета или списка публикаций. Управление этой информацией включает ее сбор, учет и систематизацию, хранение, анализ (статистический и семантический) и выдачу по запросу. Основными задачами системы ИСТИНА являются, во-первых, предоставление возможности сотрудникам структурных подразделений вуза, научного учреждения вести учет результатов своей научной деятельности и в автоматизированном режиме формировать научные отчеты, а во-вторых, предоставление руководителям отдельных структурных подразделений и организации в целом автоматизированного средства проведения количественного и тематического анализа научно-инновационной и образовательной деятельности каждого из сотрудников, отдельных подразделений и всей организации.

Конечной целью исследований, результаты которых представлены в настоящем отчете, является создание версии Системы, способ-

ной на основе анализа данных из соответствующей коллекции (базы данных) отвечать на различные аналитические запросы. Результатами выполнения таких запросов могут быть: перечень направлений, которые активно исследуются в рамках интересующей области знания (в запросе направление исследования может задаваться набором ключевых слов); динамика интереса исследователей к отдельному направлению исследований во времени; список направлений, методов и средств, которые появились относительно недавно; перечень задач, в которых используются методы интересующего направления; список публикаций по конкретному направлению за определенный период времени; список публикаций, похожих на заданную; научный вклад отдельного ученого в решение задач на интересующем направлении; перечень направлений научных интересов отдельного ученого; список исследователей, работающих на интересующем направлении; вклад отдельного подразделения в результаты деятельности научной организации за прошедший период и другие, аналогичные им.

Разрабатываемая система получила название ИСТИНА (Интеллектуальная Система Тематического Исследования НАучно-технической информации). К Системе предъявляются следующие далее **требования по функциям**, которые она должна выполнять.

- Автоматизированный ввод данных из слабо структурированных и неструктурированных текстовых файлов, содержащих результаты научной и некоторых аспектов учебной деятельности сотрудников.
- Вычисление типовых запросов, примеры которых приведены выше. Такие запросы являются аналитическими, они требуют интеллектуальной обработки исходных данных. Этот факт означает, например, что на них нельзя ответить, используя только относительно простые алгоритмы, такие как поиск по ключевым словам.
- Реализация логического вывода новых данных. Например, если пользователь вводит запрос *бигармонические уравнения*, а в онтологии содержится информация о том, что метод конеч-

1.1. НАЗНАЧЕНИЕ И ФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ

ных элементов может использоваться для решения бигармонических уравнений, то Система может выдать среди результатов поиска научные статьи, посвященные методу конечных элементов. При этом Система должна в удобном для восприятия пользователем виде указать причину включения этих статей в результаты поиска.

- Автоматизированная настройка Системы на конкретное научное направление в рамках более широкой области научного знания. Настоящее требование связано с возможной областью применения разрабатываемой Системы, а именно, с анализом научной информации в рамках крупной организации или корпорации, например, Московского государственного университета имени М.В. Ломоносова¹. Последнее обстоятельство указывает на необходимость анализа большого объема данных, имеющих отношение к различным научным направлениям.
- Автоматизированное обновление данных для их поддержания в актуальном состоянии.
- Возможность интеграции информации, содержащейся в Системе, с другими хранилищами.
- Отсутствие необходимости в использовании большого объема ручного труда экспертов.
- Возможность использования Системы для контроля и анализа научной деятельности сотрудников как внутри одной организации, так и на уровне большого числа организаций (например, в отдельном регионе или во всей стране).

Перечисленные требования определяют **специфику** разрабатываемого программного комплекса. Одна из главных его особенностей состоит в том, что он предназначен для обработки

¹Здесь и далее в качестве примера типовой крупной организации (учреждения) или корпоративного объединения, имеющих свою специфику и адекватно отображающих связанные с этим трудности реализации рассматриваемой Системы, приводится МГУ имени М.В. Ломоносова.

специальных классов документов (годовых научных отчетов сотрудников, BibTeX-файлов, списков публикаций и других). Отсюда, в частности, следует, что о конкретной статье сотрудника может быть доступна только ограниченная информация, например, название, авторы, место публикации и год публикации. В большинстве систем анализа научных материалов широко используются списки библиографических ссылок между работами. В рамках задачи, которая решена на настоящее время, такая информация в общем случае недоступна, что, однако, не исключает создания таких функциональных возможностей в будущем. Другой особенностью Системы является необходимость использования интеллектуальных алгоритмов выделения, обработки и поиска данных. Она обусловлена требованием выполнения сложных аналитических запросов, адекватно описывающих содержательную сторону документов, на неструктурированных данных в отсутствие обучающих выборок и ручного труда экспертов. Указанные особенности, в частности, подчеркивают отличие задачи, которая решается в рамках настоящей работы, от широко известного класса задач, связанных с построением электронных библиотек, более ориентированных на индексацию и предоставление доступа к электронным информационным ресурсам. В этих системах меньшее внимание уделяется тематическому анализу содержания работ и привязке содержащихся в них данных к модели области знания.

Основные функции, которые на настоящее время реализованы в Системе, можно разделить на четыре категории:

1. ввод данных;
2. отображение данных в открытом доступе;
3. формирование отчетов и представление данных для сотрудников в различных форматах;
4. анализ статистических показателей по подразделениям и по тематикам.

Система предназначена для использования как сотрудниками вузов и научных учреждений, так и их руководителями на различных уровнях. Для отдельных сотрудников Система предоставляет возможность вести учет результатов своей научной деятельности и в автоматизированном режиме формировать различные научные отчеты. Для руководителей структурных подразделений, организации в целом и для межведомственных структур Система позволяет проводить автоматизированный количественный и тематический анализ научной деятельности каждого из сотрудников, отдельных подразделений, всего учреждения и, в перспективе (при ее широком использовании), групп организаций регионального и национального масштаба.

1.2. Аналогичные системы

В настоящее время существуют представленные далее методы и средства, позволяющие решать задачу анализа научно-инновационной и тесно связанной с ней, педагогической деятельности подразделений организации, а также отдельных ее членов.

Методы анализа результатов научной деятельности ученых по информации из их отчетов

Рассматриваемый подход заключается в оценке числа публикаций, участия в конференциях, в подготовке научных кадров, а также ряда других позиций на основе результатов ручного анализа отчетов ученых экспертами. Отметим, что в настоящее время в научных организациях Российской Федерации, как правило, проводится лишь количественный анализ эффективности их деятельности. Предметом анализа служат такие характеристики, как число научных публикаций сотрудников, число докладов на конференциях и другие параметры.

Вместе с тем, тематический анализ научной деятельности

подразделений организации является очень важной задачей. Он позволяет выяснить: направления исследований, которые развиваются в организациях более активно, а также те, которые менее динамичны; понять, каким исследованиям уделяется недостаточное внимание по сравнению с общемировыми тенденциями; провести общее, и, что более важно, реализовать тематическое сравнение деятельности российских научных организаций с зарубежными. Главными недостатками подхода являются необходимость ручного просмотра всех отчетов, а также отсутствие возможности анализировать научную деятельность подразделения качественно, а не количественно. К сожалению, после проведения подобного анализа сложно ответить даже на такие казалось бы естественные и важные вопросы, как *«Какие направления исследований преобладают в работах интересующего подразделения (организации) за последний год?»*, *«Как результаты этих исследований соотносятся с общемировыми?»*. Еще одним важным недостатком традиционного подхода является отсутствие возможности оперативно проверить достоверность полученных данных.

Экспертный анализ материалов конференций и журналов

Анализ той или иной области научного знания можно провести вручную, обработав для этого большое число статей, патентов, свидетельств и других публикаций в научных журналах, в материалах конференций и других источниках по соответствующей тематике. Большой объем необходимой для такого анализа ручной работы является основным недостатком этого метода. Осложняют ситуацию многоязычность публикаций, а также необходимость интеграции материалов из различных источников данных (конференций и журналов).

Анализ обзорных статей

Для уменьшения объема ручной работы, которая необходима для анализа публикаций в рамках области знания, часто прибегают к прочтению обзорных статей, описывающих основные результаты

по интересующему направлению. Основными недостатками этого метода являются: субъективный взгляд авторов обзорных статей; как правило, недостаточно широкий охват области знания; отсутствие наиболее актуальной информации (из источников, опубликованных после обзорной статьи); сложность поиска обзорных статей, а также их фиксированный (относительно низкий) уровень детализации. Последнее обстоятельство означает, что невозможен анализ в автоматизированном режиме на более высоком или более низком уровне детализации, например, на уровне нескольких более узких направлений в рамках рассматриваемой области знания или, наоборот, на уровне нескольких областей знания.

Поиск по ключевым словам

С развитием Всемирной паутины (World Wide Web) и появлением эффективных поисковых средств представилась возможность фильтровать большие объемы текстовой информации в электронном виде, используя для этого поиск по ключевым словам. Этот способ поиска информации приобрел огромную популярность, однако он обладает существенными недостатками. Во-первых, неэффективна работа с многозначными словами и источниками на разных языках. Вопросы, связанные с уровнем детализации ответов на запрос, также являются актуальными. И, наконец, главным недостатком поиска по ключевым словам является отсутствие возможности получать ответы на сложные запросы, требующие трудоемкой предварительной обработки данных. Примером такого запроса является: *«Какие ученые из МГУ за последние два года участвовали в зарубежных конференциях по интеллектуальному анализу данных в больших коллекциях?»*.

Системы управления научной информацией

Для эффективного анализа научной, инновационной и образовательной деятельности, лишенного отмеченных выше недостатков, было создано большое число систем автоматизированного учета информа-

ции, которые получили широкое распространение с развитием Интернет и веб-технологий обработки данных. Одним из классов подобных систем являются **крупные веб-сервисы**, которые разрабатываются в рамках таких проектов, как Web of Science (часть системы Web of Knowledge¹), Google Scholar², Scopus³, российский портал eLibrary.ru⁴, PubMed⁵, AGRICOLA⁶ и других, аналогичных им. Несмотря на небольшие различия, в основе своей все эти системы являются базами научных статей и в той или иной степени совмещают в себе такие функции, как индексация и поиск научных работ. Среди недостатков этих систем следует отметить отсутствие эффективных средств анализа результатов поиска в большинстве предоставляемых ими сервисов, отсутствие семантических методов обработки данных, а также тот факт, что новые публикации, как правило, добавляются со значительным опозданием, хотя и по причинам объективного характера.

Ко второму классу систем анализа научной информации относятся **зарубежные проекты, использующие семантические технологии**. Здесь следует отметить такие системы, как Bibster⁷, JeromeDL⁸, Flink [1], AIR [2], а также семантические базы знаний общего назначения, такие как Cyc⁹, YAGO¹⁰, True Knowledge¹¹, Freebase¹², DBpedia¹³ и Wikitology [3]. Они основаны на онтологиях, содержат факты, и некоторые из них поддерживаются средствами логического вывода. Отметим, что эти хранилища информации

¹http://thomsonreuters.com/products_services/science/science_products/a-z/isi_web_of_knowledge

²<http://scholar.google.com>

³<http://www.scopus.com>

⁴<http://elibrary.ru>

⁵<http://www.ncbi.nlm.nih.gov/pubmed>

⁶<http://agricola.nal.usda.gov>

⁷<http://bibster.semanticweb.org>

⁸<http://www.jeromedl.org>

⁹<http://www.cyc.com>

¹⁰<http://www.mpi-inf.mpg.de/yago-naga/yago>

¹¹<http://www.trueknowledge.com>

¹²<http://www.freebase.com>

¹³<http://dbpedia.org>

используются как источники структурированных данных для более специализированных систем. Часть баз данных связана между собой ссылками в рамках проекта Linking Open Data¹.

К третьей группе проектов относятся две **русские разработки** – Интегрированная система информационных ресурсов (ИСИР), созданная в Российской академии наук, и комплекс, разработанный в Институте систем информатики Сибирского отделения РАН (ИСИ СО РАН), а также система Pure от Elsevier. Причина их выделения в отдельный класс состоит в том, что идеи и подходы, использовавшиеся при создании этих систем, более адекватны для применения к решению задачи, которая рассматривается в настоящей работе.

Целью **проекта ИСИР** является организация Единого научно-информационного пространства РАН путем интеграции разнородных научно-информационных и программных ресурсов отдельных научных учреждений [4]. Создание такого пространства позволит предоставить пользователям более эффективные средства интеграции и поиска информации, научных коммуникаций и сотрудничества. Несомненным достоинством системы ИСИР является возможность получать доступ к разнородным данным, которые хранятся в различном виде, например, в реляционных базах данных, объектно-ориентированных базах данных, каталогах LDAP, а также в виде RDF-утверждений, с помощью так называемых оберток (wrappers). Система ИСИР способна осуществлять работу с распределенной информацией, обеспечивая интеграцию схем данных с помощью так называемых посредников, которые выполняют такие задачи, как репликация, маршрутизация и кэширование запросов. Гибкая настраиваемость системы ИСИР и возможности ее применения к различным областям знания также являются достоинствами. Система позволяет относительно легко и быстро разрабатывать порталы, ориентированные на конечного пользователя. Из недостатков ИСИР следует выделить отсутствие интеллектуальных алгоритмов определения тематики публикаций и других объектов. Отсутствует возможность генерировать онтологии

¹ <http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

области знания автоматически, а также – возможность ввода поискового запроса к системе в свободной форме. Тематику желаемых результатов поиска можно задать только явным образом, указав коды классификатора ВИНТИ. На рис. 1.1 представлена общая схема работы ИСИР.

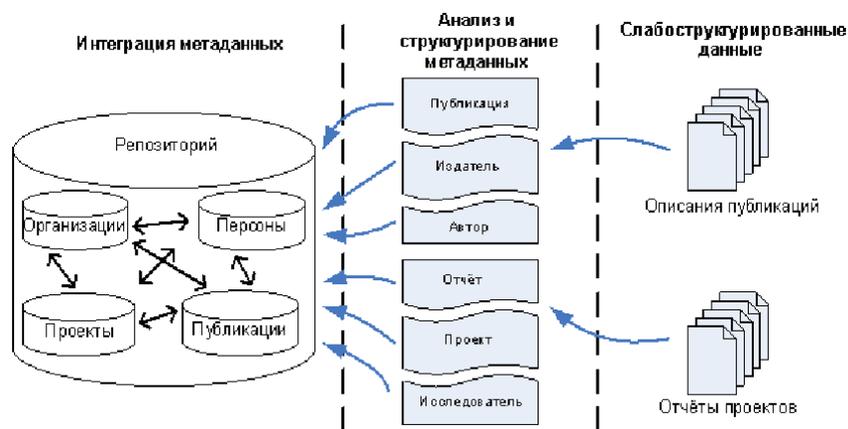


Рис. 1.1. Схема работы системы ИСИР.

Значительный вклад в решение задач на рассматриваемом направлении вносят работы, выполненные в **Институте систем информатики Сибирского отделения РАН** [5, 6]. Целью этих работ является создание интернет-портала знаний, представляющего научную дисциплину и интегрирующего ресурсы по тематике (публикации, сайты, организации и другие) в единое информационное пространство. Отметим, что на основе предлагаемой методики были созданы два предметных портала, которые функционируют в настоящее время. Один из них посвящен компьютерной лингвистике¹, а другой – археологии². Достоинствами системы являются: использование нескольких онтологий, что облегчает ее настройку на произвольную область знания; выполнение сложных аналитических запросов; возможность задать запрос как в свободной форме, так

¹<http://uniserv.iis.nsk.su/cl>

²<http://www.sati.archaeology.nsc.ru/classarch2>

и в терминах онтологии; возможность редактирования онтологии в процессе эксплуатации портала, а также развитая методика визуализации онтологии. Из недостатков системы следует отметить отсутствие модуля автоматизированного построения онтологии области знания, необходимость формирования подборки документов, релевантных области знания, для поддержки автоматизированного создания словарей экспертами, особенности методики скачивания документов, которые не позволяют находить наиболее актуальную информацию по тематике. Основные концепты онтологий научной деятельности и научного знания представлены на рис. 1.2.

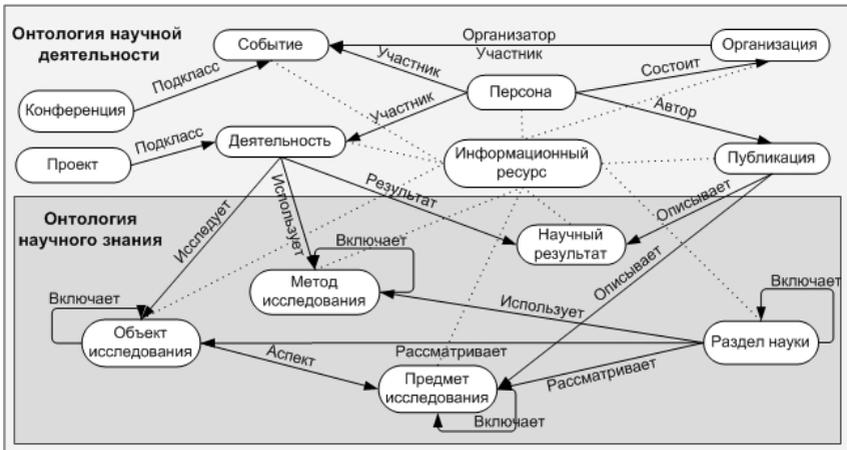


Рис. 1.2. Схема основных онтологий системы ИСИ СО РАН.

Анализ двух конкретных интернет-порталов (по компьютерной лингвистике и археологии), разработанных на основе подхода ИСИ СО РАН, выявил такие недостатки, как малый охват области знания, неудобная визуализация объектов и связей между ними, отсутствие описания модели области знания, что затрудняет навигацию и поиск, недостаток механизмов для интеллектуальной обработки поискового запроса, сложность интерфейса расширенного поиска и отсутствие логического вывода при поиске. В частности, реализация расширенного поиска на портале компьютерной лингвистики ограничивает возможности пользователя хорошо ориентироваться в про-

странстве понятий и отношений онтологии. Фрагмент интерфейса расширенного поиска на портале представлен на рис. 1.3.

Завершая краткое описание систем ИСИР и ИСИ СО РАН, заметим, что они обладают большим числом достоинств, перечисленных выше. Вместе с тем отметим, что главное отличие этих систем от системы ИСТИНА, которая рассматривается в настоящей работе, состоит в формулировке их основной цели. Цель проекта ИСИР – интегрировать информационные и программные ресурсы отдельных научных учреждений. Цель портала ИСИ – предоставить адекватное описание области знания. Главной целью системы ИСТИНА является создание инструментария, обеспечивающего сбор, поиск, систематизацию, интеллектуальный анализ данных о научно-инновационной и педагогической деятельности организации (объединения организаций) на всех уровнях (от персоналий до организации или корпорации в целом), как средства эффективного управления ею. С этих позиций наиболее близкой по целевым установкам и используемой методологии к системе ИСТИНА является **система Pure**¹, которая разрабатывается и внедряется корпорацией Elsevier². Цель проекта Pure состоит в привнесении доказательного (evidence-based) подхода к научной стратегии учреждения, к оценке эффективности его деятельности и к повседневным бизнес-решениям. Система предназначена для управленческого персонала научных организаций, ученых и сотрудников научных библиотек. Программный продукт Pure разработан на языке Java. Система внедрена в ряде европейских университетов. В России она внедряется в Уральском федеральном университете³. Система Pure включает следующие функции.

- Добавление информации о результатах научной деятельности. Данные в систему могут вводить сами пользователи и ответственные. Поддерживается импорт данных в формате

¹<http://www.elsevier.com/online-tools/research-intelligence/products-and-services/pure>

²<http://www.elsevier.com>

³<http://urfu.ru>

Описывает-Результат	
Научные результаты и продукты	
Название результата	<input type="text"/> <input type="checkbox"/> Точное значение
Описание результата	<input type="text"/> <input type="checkbox"/> Точное значение
Дата получения	г.: <input type="text"/> м.: <input type="text"/> Д.: <input type="text"/> вр.: <input type="text"/> : <input type="text"/> <input type="radio"/> содерж. <input type="radio"/> не сод. <input type="radio"/> = <input type="radio"/> != <input type="radio"/> ≤ <input type="radio"/> ≥ <input type="radio"/> < <input type="radio"/> > <input type="radio"/> [<input type="radio"/>] <input type="radio"/> (<input type="radio"/>) <input type="radio"/> (.)
URL	<input type="text"/> <input type="checkbox"/> Точное значение
Аббревиатура	<input type="text"/> <input type="checkbox"/> Точное значение
Описывает-Объект	
Объекты исследования	
Название объекта	<input type="text"/> <input type="checkbox"/> Точное значение
Описание объекта	<input type="text"/> <input type="checkbox"/> Точное значение
Описывает-Метод	
Методы и средства исследования	
Название метода	<input type="text"/> <input type="checkbox"/> Точное значение
Другие названия метода	<input type="text"/> <input type="checkbox"/> Точное значение
Описание	<input type="text"/> <input type="checkbox"/> Точное значение
URL	<input type="text"/> <input type="checkbox"/> Точное значение
Дата (возникновения)	г.: <input type="text"/> м.: <input type="text"/> Д.: <input type="text"/> вр.: <input type="text"/> : <input type="text"/> <input type="radio"/> содерж. <input type="radio"/> не сод. <input type="radio"/> = <input type="radio"/> != <input type="radio"/> ≤ <input type="radio"/> ≥ <input type="radio"/> < <input type="radio"/> > <input type="radio"/> [<input type="radio"/>] <input type="radio"/> (<input type="radio"/>) <input type="radio"/> (.)
Публикация о Деятельности	
Деятельность	
Название деятельности	<input type="text"/> <input type="checkbox"/> Точное значение
Описание деятельности	<input type="text"/> <input type="checkbox"/> Точное значение
Дата начала	г.: <input type="text"/> м.: <input type="text"/> Д.: <input type="text"/> вр.: <input type="text"/> : <input type="text"/> <input type="radio"/> содерж. <input type="radio"/> не сод. <input type="radio"/> = <input type="radio"/> != <input type="radio"/> ≤ <input type="radio"/> ≥ <input type="radio"/> < <input type="radio"/> > <input type="radio"/> [<input type="radio"/>] <input type="radio"/> (<input type="radio"/>) <input type="radio"/> (.)
Дата окончания	г.: <input type="text"/> м.: <input type="text"/> Д.: <input type="text"/> вр.: <input type="text"/> : <input type="text"/> <input type="radio"/> содерж. <input type="radio"/> не сод. <input type="radio"/> = <input type="radio"/> != <input type="radio"/> ≤ <input type="radio"/> ≥ <input type="radio"/> < <input type="radio"/> > <input type="radio"/> [<input type="radio"/>] <input type="radio"/> (<input type="radio"/>) <input type="radio"/> (.)
Аббревиатура	<input type="text"/> <input type="checkbox"/> Точное значение
Научное направление	
Разделы Науки	
Название раздела	<input type="text"/> <input type="checkbox"/> Точное значение
Описание раздела	<input type="text"/> <input type="checkbox"/> Точное значение
Ключевые термины	<input type="text"/> <input type="checkbox"/> Точное значение
Описывает Класс Методов исследования	
Методы и средства исследования	
Название метода	<input type="text"/> <input type="checkbox"/> Точное значение
Другие названия метода	<input type="text"/> <input type="checkbox"/> Точное значение
Описание	<input type="text"/> <input type="checkbox"/> Точное значение
URL	<input type="text"/> <input type="checkbox"/> Точное значение
Дата (возникновения)	г.: <input type="text"/> м.: <input type="text"/> Д.: <input type="text"/> вр.: <input type="text"/> : <input type="text"/> <input type="radio"/> содерж. <input type="radio"/> не сод. <input type="radio"/> = <input type="radio"/> != <input type="radio"/> ≤ <input type="radio"/> ≥ <input type="radio"/> < <input type="radio"/> > <input type="radio"/> [<input type="radio"/>] <input type="radio"/> (<input type="radio"/>) <input type="radio"/> (.)

Рис. 1.3. Интерфейс расширенного поиска в системе ИСИ СО РАН (фрагмент).

BibTeX. Разбор библиографических ссылок в текстовом формате не предусмотрен.

- Формирование отчетов по организации, подразделению, проекту, отдельному ученому.
- Оценка эффективности научной деятельности, в том числе – по стандартным схемам оценки, принятым в Великобритании, Дании, Австралии и Голландии. В Pure поддерживается EuroCris – европейский стандарт на результаты научных исследований. Научные проекты, получающие финансирование в Европе, обязаны публиковать отчеты с использованием этого стандарта.
- Управление профилями ученых, пользователи сами могут дополнять свои профили.
- Управление полным циклом научного проекта (от заявки до реализации).
- Анализ сетей связей ученых.
- Поиск экспертов.
- Интеграция с подсистемами университета: кадры, управление студентами, бухгалтерия, награды. Она осуществляется следующим образом. В Pure задан формат, в котором данные могут туда поступать. Разработчики-представители вуза должны написать программы-коннекторы, которые выгружают данные из подсистем университета в этот формат и тогда данные будут попадать в Pure.
- Автоматическая интеграция (импорт данных) со Scopus, Web of Science, PubMed, Mendeley, arXiv и других источников. Импорт статей на русском языке пока не осуществляется.
- Функции разграничения доступа к данным реализованы на основе ролей с возможностью назначения меток конфиденциальности. Отметим, что отсутствуют возможности назначать права на отдельный объект (например, дать возможность ученому секретарю редактировать информацию только о его диссертационном совете).

- Использование онтологии CERIF (рис. 1.4).
- Автоматический поиск источников финансирования (модуль пока находится в разработке).
- Слияние дубликатов статей.
- Для аутентификации пользователей могут использоваться Active Directory, LDAP, Single Sign-in и другие стандартные схемы.
- Программный интерфейс для разработчиков (API), SOAP и REST-интерфейсы.

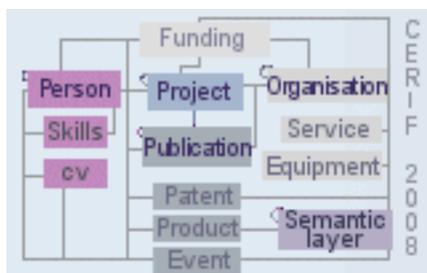


Рис. 1.4. Онтология CERIF (фрагмент).

Подобно методологии, которая реализуется в системе ИСТИНА, одной из главных задач в Pure является «очистка» и верификация данных. При импорте информации ученые должны пройти процедуру проверки для каждой публикации. Для сотрудников, не работающих в организации, также автоматически создаются профили. Однако, если фамилия автора-сотрудника пишется в разных публикациях по-разному, то для разных описаний фамилий создадутся разные профили. Чтобы этого избежать, необходимо заранее внести все свои альтернативные написания фамилии (как и в ИСТИНЕ), а также альтернативные написания названий подразделений. Заметим, что в системе ИСТИНА, как будет представлено далее, присутствует функция слияния дубликатов сотрудников. С учетом изложенного отметим, что по методологии разработки и по используемым базовым механизмам система Elsevier Pure во многом похожа на систему ИСТИ-

НА. Вместе с тем, необходимо подчеркнуть следующие преимущества системы ИСТИНА, которые более подробно будут представлены далее:

- возможность формирования отчетов, принятых в российских реалиях, например, годовых отчетов организаций, отчетов о деятельности диссертационных советов и других;
- поддержка большего количества видов результатов научной деятельности (более 20; в Pure поддерживается обработка информации о публикациях, грантах, дипломах и диссертациях, оборудовании);
- возможность автоматизированного импорта (разбора) данных в текстовом формате библиографических ссылок;
- русскоязычный интерфейс (в Pure он пока в разработке);
- возможность использования дискреционной модели разграничения доступа к объектам для задания индивидуальных разрешений для конкретных объектов (например, каждому диссертационному совету назначается свой редактор);
- опыт успешного внедрения в крупнейшем вузе страны (МГУ).

Подводя итог представленным выше подходам к решению поставленной в настоящей работе задачи, а также к созданным в рамках ее решения методам и средствам управления научной информацией, перечислим основные недостатки систем обработки и анализа научных данных, которые могут рассматриваться как возможные решения основной задачи. Они включают такие особенности, как сложность ввода данных, сложность поиска информации, использование «жестких» схем классификации и схем данных, ориентация на приоритет обработки информации из Интернет, на распределенный характер хранения и доступа к данным, а также недостаточное внимание к интеллектуализации алгоритмов загрузки, обработки и поиска информации.

1.3. Архитектура системы ИСТИНА

Взаимодействие пользователей с системой ИСТИНА осуществляется через веб-интерфейс, и архитектура Системы является типовой для современных веб-приложений (рис. 1.5).

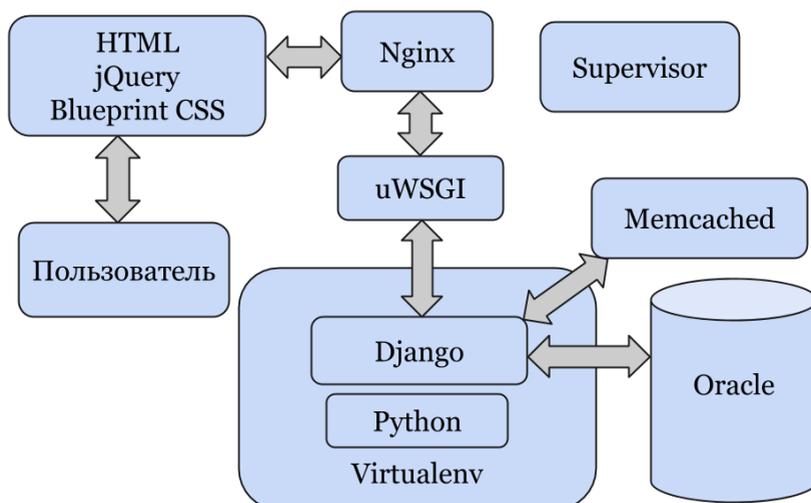


Рис. 1.5. Архитектура системы ИСТИНА.

В основе Системы лежит сервер приложений. Главным языковым средством для разработки приложений в Системе на настоящее время является программное средство (каркас) Django¹, который написан на языке Python. Для изоляции пакетов на сервере приложений и предотвращения конфликтов их версий используется модуль виртуального окружения virtualenv² и вспомогательный модуль virtualenvwrapper³. В качестве веб-сервера применяется

¹<https://www.djangoproject.com/>

²<http://www.virtualenv.org>

³<http://virtualenvwrapper.readthedocs.org>

HTTP-сервер Nginx¹. Взаимодействие между веб-сервером и Python-кодом приложений осуществляется по стандарту WSGI² с помощью программы uWSGI³. В качестве СУБД используется Oracle, физически расположенный на отдельной машине. Для ускорения доступа к данным применяется система управления кэшем memcached⁴. Для контроля состояния основных процессов на сервере приложений используется программа Supervisor⁵. В случае аварийного завершения какого-либо из процессов он автоматически перезапускается. Ответ пользователю, сгенерированный на сервере приложений с использованием информации из базы данных, в общем случае является HTML-документом. Для управления отображением элементов в HTML-документах применяется библиотека Blueprint CSS⁶, а для задания поведения элементов на стороне клиента – Javascript-библиотека jQuery⁷.

Как было отмечено выше, основной код Системы написан с использованием инструментального средства Django, написанного на языке Python. Оно реализует вариант шаблона проектирования «модель-представление-поведение» (Model-View-Controller, MVC⁸). Инструментарий Django также включает средства объектно-реляционного отображения (ORM), позволяющие автоматически генерировать SQL-запросы на основе Python-кода, включает встроенный язык шаблонов и поддерживает независимое задание URL-адресов для веб-сайта. Отметим, что эти функции в том или ином виде входят в большинство современных программных каркасов для разработки веб-приложений. Необходимо заметить, что шаблон проектирования MVC позволяет разделить на отдельные компоненты модель данных приложения, пользовательский интерфейс (представление) и связь между пользователем и

¹<http://nginx.org>

²<http://legacy.python.org/dev/peps/pep-3333>

³<http://uwsgi-docs.readthedocs.org>

⁴<http://memcached.org>

⁵<http://supervisord.org>

⁶<http://www.blueprintcss.org>

⁷<http://jquery.com>

⁸<http://heim.ifi.uio.no/trygver/themes/mvc/mvc-index.html>

Системой (поведение/контроллер). Такое разделение позволяет повысить возможность повторного использования кода и облегчить его модификацию. Выбор именно Django обусловлен, кроме исторических причин, рядом достоинств этого программного обеспечения, которые выделяли его на фоне других библиотек на момент принятия решения в 2011 году. Во-первых, привязка к языку Python дает возможность использовать богатый набор других библиотек языка, которые реализуют различные функции, причем не только связанные с веб-интерфейсом Системы. Во-вторых, в Django изначально заложена экосистема повторно используемых приложений (third-party apps), реализующих функциональные возможности, разделяемые различными веб-сайтами. Например, в системе ИСТИНА применяется стороннее Django-приложение для регистрации пользователей. Более подробный список используемых приложений представлен в следующих разделах. В-третьих, в Django на основе заданных программистом моделей автоматически генерируется полноценный веб-интерфейс администратора, с помощью которого он может просматривать, добавлять, изменять и удалять всю необходимую информацию в базе данных. Наконец, эффективность и масштабируемость Django подтверждается тем фактом, что библиотека используется в ряде крупнейших веб-проектов с десятками и даже сотнями миллионов пользователей, таких как Instagram¹, Pinterest², Disqus³, Mozilla⁴, Bitbucket⁵, Mahalo⁶, Rdio⁷, Lanyrd⁸, The Onion⁹, на веб-сайтах крупнейших изданий, таких как The Washington Post¹⁰, The New York Times¹¹,

¹<http://instagram.com>

²<https://www.pinterest.com>

³<http://disqus.com>

⁴<http://www.mozilla.org>

⁵<https://bitbucket.org>

⁶<http://www.mahalo.com>

⁷<http://www.rdio.com>

⁸<http://lanyrd.com>

⁹<http://www.theonion.com>

¹⁰<http://www.washingtonpost.com>

¹¹<http://www.nytimes.com>

The Guardian¹.

На рис. 1.6 представлена общая схема обработки HTTP-запроса в Django. Запрос пользователя, первоначально обработанный веб-сервером, попадает через программу-коннектор (в данном случае – uWSGI) в модуль `urls.py`, в котором происходит сравнение URL запроса со списком регулярных выражений, заданном в приложении. Каждому регулярному выражению соответствует Python-функция, называемая `view`-функцией, которая исполняется в том случае, если адрес запроса соответствует этому выражению. Отметим, что исполняется только одна функция, которая определяется первым в списке таких выражений. `View`-функции находятся в файлах `views.py` внутри приложений. Эти функции принимают на вход объект `request`, содержащий данные HTTP-запроса, а также параметры запроса из его адресной строки. В `view`-функции может происходить обращение к методам сущностей-моделей, определенных в виде классов в файле `models.py`. Во `view`-функциях, а также в моделях выполняются запросы к базе данных с помощью ORM-отображения. В файле `settings.py` находятся общие параметры проекта, включающие настройки базы данных. Во `view`-функции формируются данные для отображения пользователю. Эти данные передаются в заданный шаблон (написанный, например, на языке HTML) и заполненный HTML-документ отправляется через программу-коннектор и веб-сервер обратно пользователю.

Архитектура сервера приложений Системы, которая представлена на рис. 1.7, отражает требования к ее функциональным возможностям, в общем виде перечисленным в Введении.

Основу Системы составляет **ядро**, реализующее базовые (системообразующие) функции, которые напрямую (во многом) определяют и качество (показатели качества) Системы в целом. Оно содержит базовые классы, методы, формы и функции, используемые в других модулях. Ядро обеспечивает единый интерфейс и базовый каркас добавления, просмотра, редактирования и удаления результатов научно-педагогической деятельности сотрудников. В ядре со-

¹<http://www.theguardian.com>

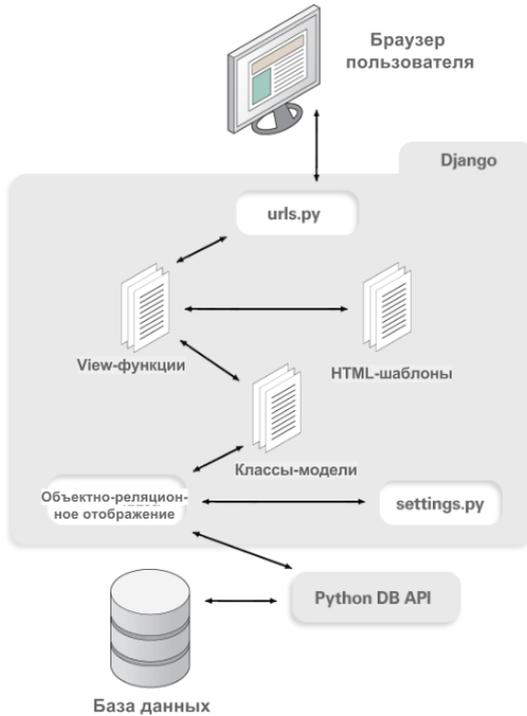


Рис. 1.6. Схема обработки HTTP-запроса в Django.

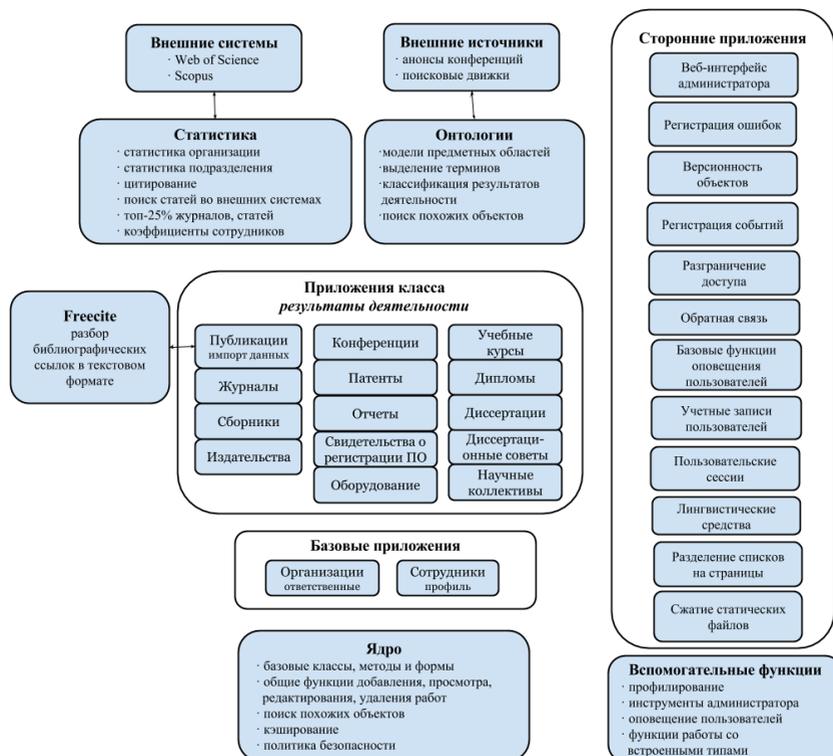


Рис. 1.7. Архитектура сервера приложений системы ИСТИНА.

держатся функции поиска похожих объектов, которые используются для подбора, в частности, похожих сотрудников, журналов и статей. Важной частью ядра является механизм кэширования. Он используется для ускорения доступа к различным данным Системы. В ядре реализуются также и другие общие функции, например, поддерживающие политику безопасности Системы, в частности, механизмы разграничения доступа к различным категориям данных Системы.

На следующем уровне архитектуры расположены два базовых приложения, **модуль организаций** и **модуль сотрудников**. Отметим, что в настоящем разделе под приложением понимается логически и физически отдельный программный модуль Системы, реализующий ту или иную базовую или прикладную функцию. При этом каждое приложение одновременно является и приложением в терминологии Django («Django app»). Модуль организаций включает классы и логику действий, связанную с данными на уровне организаций. В частности, в этот модуль включен механизм реализации действий сотрудников, которым делегирована роль ответственных от организации и отдельных ее подразделений. Он обеспечивает добавление, просмотр, редактирование и удаление информации об ответственных, которые призваны решать вопросы заданного пользователя в зависимости от его места работы. Приложение «сотрудники» содержит базовые классы, связанные с сотрудниками, в частности, их профиль в Системе и список работ. Модули организаций и сотрудников являются базовыми приложениями потому, что с ними связаны все остальные приложения Системы, которые находятся выше по иерархии.

Основным уровнем архитектуры являются приложения класса **«результаты деятельности»**. Каждое приложение отвечает за отдельный тип результатов научно-инновационной и преподавательской деятельности сотрудников (например, публикации) или за общую сущность, связанную с таким типом (например, журналы). Эти приложения опираются на общий каркас классов и функций Системы, которые представлены в ее ядре. Данный подход позволяет сравнительно легко добавлять новые типы результатов деятельности персоналий в Систему. Такие функциональные

возможности как добавление, просмотр, редактирование и удаление, а также привязка результата деятельности к сотрудникам-авторам, уже реализованы в ядре. Программисту требуется лишь настроить каркас на конкретный тип результатов деятельности, указав специфицирующие его свойства. Отметим, что наиболее сложным приложением из рассматриваемого класса являются публикации. Оно использует модуль разбора библиографических ссылок в текстовом формате Freecite¹, который был усовершенствован с целью предоставления пользователю возможности обработки текстов на русском языке.

На следующем уровне иерархии архитектуры Системы расположены два модуля, отвечающие за анализ данных, накопленных на предыдущих уровнях. **Модуль статистики** содержит функции преимущественно количественного анализа данных Системы. Он включает в себя функции анализа данных на уровне организации, подразделения и отдельного сотрудника. В этом модуле также реализован тематический анализ показателей результативности сотрудников, который выполнен сравнительно простыми средствами. А именно, статья в журнале считается принадлежащей тематической рубрике, к которой принадлежит сам журнал. Приложение «статистика» включает функции получения и обработки показателей цитирования отдельных статей из Web of Science, из Scopus, а также поиск статей в этих системах. Кроме этого, в модуль статистики входят механизмы расчета принятых в организации коэффициентов эффективности деятельности сотрудников на основе списка их статей в высокорейтинговых журналах. Эти коэффициенты в дальнейшем могут использоваться при расчете поощрительных надбавок сотрудникам.

Приложение «онтологии» имеет целью выполнение более глубокого тематического анализа данных в Системе. В частности, в механизмах этого модуля используются модели предметных областей, построенные в автоматизированном режиме. Для их построения используются анонсы научных конференций, а также результаты запросов к поисковым системам, таким как Bing. Эти модели содер-

¹<http://freecite.library.brown.edu/>

жат термины, характерные для соответствующей предметной области. По этим терминам затем выполняется классификация результатов деятельности сотрудников (например, статей), а также подбор похожих объектов для удобной навигации по Системе. Отметим, что большинство функций этого модуля пока находятся в стадии исследовательской разработки. Однако эти исследования рассматриваются разработчиками Системы, как одно из самых перспективных направлений ее развития.

Кроме основной вертикали приложений, которая представлена в левой части рис. 1.7, в Системе используются следующие вспомогательные приложения и функции:

- профилирование, которое включает оценку производительности и других характеристик Системы (в частности, для оценки скорости загрузки краткой информации о сотруднике и списка последних добавленных в Систему работ);
- инструментальные средства администратора для использования в командной строке и в веб-интерфейсе Системы;
- функции оповещения пользователей о различных событиях в Системе;
- вспомогательные функции работы со встроенными типами языка Python (такими, как строки, кортежи, списки и словари);
- веб-интерфейс администратора, предоставляемый библиотекой Django;
- подсистема регистрации ошибок на сайте (приложение `djangosentry`¹);
- подсистема регистрации и хранения версий объектов, используемая для обеспечения безопасности данных Системы, а также для ведения дополнительного журнала событий на уровне всей Системы и отдельного пользователя (приложение `djangoreversion`²);

¹<https://github.com/getsentry/sentry>

²<https://github.com/etianen/django-reversion>

- подсистема регистрации и отображения событий различного уровня в Системе (приложение `django-activity-stream`¹);
- механизмы разграничения доступа к отдельным объектам в Системе (приложение `django-object-permission`²);
- функции обратной связи с разработчиками Системы и регистрации сообщений пользователей;
- базовые функции оповещения пользователей (приложение `messages`, входящее в состав Django);
- приложение для работы с учетными записями пользователей (приложения `django-profile`³ и `django.contrib.auth`);
- приложение для работы с пользовательскими сессиями (`sessions`);
- средства для работы с русским языком (например, склонение числительных) (приложение `pytils`⁴);
- функции разделения текста на страницы (пагинации) для удобного отображения (приложение `django-paging`⁵);
- приложение для сжатия и поддержки механизмов версииности статических файлов (`django-compressor`⁶).

1.4. Отличительные особенности

В настоящем разделе отметим основные преимущества, которыми на настоящее время обладает система ИСТИНА, по сравнению с другими системами, предназначенными для управления научной информацией.

¹<https://github.com/justquick/django-activity-stream>

²<https://pypi.python.org/pypi/django-object-permission>

³<https://code.google.com/p/django-profile>

⁴<https://github.com/j2a/pytils>

⁵<https://github.com/dcramer/django-paging>

⁶https://github.com/jezdez/django_compressor

Широкий охват результатов научной деятельности

В Системе учитываются не только публикации, как в большинстве других систем, но и участие в конференциях, патенты, диссертации, дипломные работы, учебные курсы, членство в редколлегиях журналов и другие виды результатов деятельности. Система включает общую архитектуру для быстрого расширения учитываемых типов результатов научно-педагогической и инновационной деятельности. Таким образом достигается высокий уровень полноты данных, описывающих результаты работы организации.

Контроль пользователей над данными

Несмотря на потенциально существующую возможность импорта данных из других источников («сверху вниз»), Система концептуально (и в архитектурном, и в техническом плане) ориентирована:

- на пополнение своих данных от конечных пользователей (по принципу «снизу вверх»);
- на анализ, уточнение, модификацию данных и их организацию в хранилище, в значительной степени полагаясь на заинтересованное участие в этом процессе конечных пользователей.

Децентрализованный ввод данных в Систему по принципу «снизу вверх» определяет следующие ее преимущества:

- сотрудник более чем кто-либо заинтересован в том, чтобы доступная для всех заинтересованных лиц его страница в Системе (и сети Интернет) содержала наиболее точные и полные данные;
- только сам сотрудник может знать полный список результатов своей научной и учебно-педагогической деятельности;
- таким образом, в Систему поступает информация, которую крайне трудно или невозможно получить другим путем (из других источников).

Перечисленные свойства Системы обеспечивают высокую точность и полноту данных, которые адекватно характеризуют деятельность организации в целом.

Удобство ввода информации

Большое внимание при разработке Системы было уделено интерфейсу ее взаимодействия с пользователем. Для ускорения ввода информации о публикациях служит модуль разбора библиографических ссылок, который позволяет добавлять строку с описанием работы и не заполнять большое число полей. Еще одной полезной функцией является однократный ввод данных, который означает, что, например, публикация, добавленная одним сотрудником, автоматически добавляется всем ее соавторам.

Использование онтологий для учета специфики различных областей знаний

Использование в будущем апробированного в Системе подхода к описанию научного знания с помощью онтологий позволит применять дифференцированные механизмы поиска и анализа данных в зависимости от области знаний, к которой эти данные принадлежат. Такие механизмы поиска и анализа позволят более адекватно отражать и учитывать особенности именно интересующей пользователя области знания.

Важным преимуществом является также возможность привлечения большого числа высококвалифицированных ученых из МГУ и других ведущих вузов и научных центров России для участия в составлении формальных моделей (онтологий) практически всех областей науки.

Полезные функции для сотрудников

Система представляет практический интерес не только для административно-управленческих структур организаций науки и

образования, но и для конечных пользователей. Она позволяет:

- сделать более удобным (эргономичным) процесс управления (ввод, отображение, анализ, резервное копирование) данными о результатах своей научной и педагогической деятельности (персональными данными);
- в автоматическом режиме формировать отчеты, которые необходимы для выполнения формальных административных процедур (годовой отчет, подача заявок на гранты, информационный лист для переизбрания);
- выдавать списки публикаций в различных стилях, что, в свою очередь, дает возможность копировать их в другие документы (в заявки на гранты, на публикации в журналах и т.д.).

Широкие возможности анализа для руководителей

Система ИСТИНА может использоваться как инструментарий управления ВУЗом, НИИ, а также является дополнительным средством повышения его репутации (рейтинговых показателей) за счет представления данных о деятельности сотрудников организации в открытом доступе в Интернет.

Здесь также необходимо отметить следующие преимущества Системы:

- возможность сравнения показателей подразделений организации;
- возможность анализа развития отдельной области науки в организации, и просмотр статистики по ее подобластям;
- возможность учета и анализа актуальных на текущий момент времени данных по различным запросам в динамике (в реальном времени) по мере наполнения базы новыми данными.
- использование оригинальных алгоритмов семантического анализа научно-технических текстов, разработанные в МГУ;

Внедрение и масштабирование

В рамках вопроса о потенциальных возможностях внедрения и масштабирования Системы, следует подчеркнуть следующие ее преимущества:

- опыт внедрения Системы в такой крупной организации как МГУ, адекватно моделирующей структуру российских научных и образовательных учреждений, а также корпоративных структур;
- возможность масштабирования решений, применяемых в Системе, на большое число внешних по отношению к МГУ организаций, желающих подключиться к Системе путем получения доступа к единой информационной среде, в том числе – в режиме распределенной системы с «облачной» архитектурой.

Глава 2

Модули, реализующие основные функции

2.1. Персонализация

Одной из основных отличительных особенностей системы ИСТИ-НА является ее ориентация на работу с конечным пользователем, а именно – с ученым, инженером-исследователем, педагогом и т.п. Как было упомянуто ранее, не исключая и другие способы сбора данных, именно персоналии (сотрудники) являются главным источником информации, размещают и корректируют данные о результатах своей деятельности в Системе. В этой связи одной из основных задач при ее разработке является построение удобного интерфейса взаимодействия с конечным пользователем. Следует отметить, что многие потенциальные пользователи Системы являются людьми старшего поколения. Именно эти много лет отдавшие науке и образованию люди, как правило, имеют большое количество публикаций и других результатов научно-образовательной и инновационной деятельности. Однако далеко не все из них по объективным причинам обладают в должном объеме навыками использования современных средств вычислительной техники и технологий. В связи с этим обстоятельством, простота интерфейса взаимодействия пользователя с

Системой, а также наличие подробных сопровождающих такое взаимодействие инструкций также являются одними из приоритетных требований к Системе.

Цикл работы пользователя (ученого, сотрудника) с Системой может включать следующие шаги.

- Регистрация в Системе.

При регистрации пользователь указывает свое имя в Системе (логин), адрес электронной почты и пароль. Затем на его адрес электронной почты высылается письмо, необходимое для подтверждения принадлежности этого адреса пользователю. Пользователь нажимает на ссылку в письме, его учетная запись активируется, и он получает право пользоваться Системой.

- «Привязка» к сотруднику.

Следующим этапом работы пользователя с Системой является поиск себя в базе сотрудников. Дело в том, что если соавторы ученого уже зарегистрировались в Системе и добавили в ее базу данных результаты своей деятельности (работы), то вполне вероятно, что для ученого уже создана страница с его результатами. По этой причине новый пользователь может найти себя в Системе и «привязать» (далее этот термин используется без кавычек) этого сотрудника (а также все результаты деятельности, отраженные на его странице) к себе. Для этого ему нужно указать свою фамилию, имя и отчество, после чего Система выдает список похожих сотрудников, найденных в базе.

- Добавление альтернативных имен.

Альтернативные имена – это другие варианты написания имени ученого, например, в сокращенной форме, на другом языке или девичья фамилия. Альтернативные имена используются при добавлении работ для связи автора работы с сотрудником для того, чтобы работа появилась на личной странице тех ученых, которым она принадлежит.

Если ученый собирается добавлять работы на иностранном

языке, ему необходимо предварительно указать в качестве альтернативного свое имя на этом языке. Только после этого добавляемые работы будут привязываться к нужному сотруднику и, следовательно, отобразятся на его личной странице. Впоследствии он, при желании, может отредактировать или удалить альтернативные имена.

- **Указание места работы.**
Пользователь может указать организацию, подразделение, дату начала и окончания работы, а также отметить тот факт, что он работает по совместительству. В Системе имеется возможность указывать несколько мест работы, а также – удалять место работы.
- **Ввод дополнительных сведений.**
Кроме имени, альтернативных имен и места работы, ученый может добавить в свой профиль соответствующий ему идентификатор в системе ResearcherID, ученые степени и звания (с возможностью указания области науки, специальности, кафедры, академии, членом которой он является, и года присуждения звания), а также фотографию.
- **Добавление информации о результатах деятельности.**
После регистрации и заполнения своего профиля ученый может перейти к добавлению в Систему информации о результатах своей деятельности.

После регистрации у сотрудника появляется личная страница, на которой отображается информация о нем и о всех результатах его деятельности. Отметим, что страница существует в двух вариантах – в полном и сокращенном. В сокращенном варианте (рис. 2.1) на ней отображается лишь последние 5 добавленных (самим сотрудником или его соавторами) результатов деятельности, что облегчает просмотр новых данных.



Изменить
фотографию
отправить
сообщение

Голомазов Денис Дмитриевич

пользователь

МГУ имени М.В. Ломоносова, Научно-исследовательский институт механики, 404
Лаборатория автоматизации экспериментальных исследований, с 8 июля 2008

кандидат физико-математических наук с 2012 года

Соавторы: Афонин С.А., Козицын А.С., Васенин В.А., Ганкин Г.М., Бахтин А.В.

10 статей, 27 докладов на конференциях, 7 тезисов докладов, 4 НИР, 3

свидетельства о регистрации прав на ПО, 2 отчета, 1 диссертация, 1 учебный курс

Количество цитирований статей в журналах по данным Scopus: 3, Web of Science: 3
откуда взялись эти числа?

ResearcherID: G-6025-2012

Работа с системой

Мои работы
Редактировать профиль
Мои отчеты
Мои достижения
Мои отобранные статьи
Последние добавленные мной работы
Экспорт публикаций в BibTeX
Мои возможные дубликаты
Мои потерянные статьи
Поиск и редактирование журналов
Последние добавленные журналы
Инструкция по работе с системой

Последние добавленные работы [Все работы](#) [помощь по этой странице](#)

Доклады на конференциях

2013 Система автоматизации учета результатов научной деятельности

Авторы: Козицын А.С., Афонин С.А., Голомазов Д.Д.

XVIII Международная конференция по Вычислительной механике и современным прикладным программным системам (ВМСППС2013), Алупка, Крым

2011 Использование систем семантического анализа для организации поиска научно-технической информации

Авторы: Козицын А.С., Афонин С.А., Голомазов Д.Д.

Вероссийская конференция с международным участием "Знания-Онтологии-Теория" (ЗОНТ-2011), Новосибирск

НИРы

Разработка алгоритма распараллеливания теплогидравлического кода (CMS) полномасштабной модели 1 сентября 2011 - 31 декабря 2011

Научно-исследовательский институт механики

Участники НИР: Астапов И.С., Афонин С.А., Васенин В.А., Голомазов Д.Д., Занчурич М.А., Итес А.А., Казьмин О.О., Козицын А.С., Коршунов А.А., Кривчиков М.А., Морозова Н.А., Першин И.С., Пучков Ф.М., Рагулин А.Д., Роганов В.А., Титов А.С., Шалченко К.А., Шундеев А.С.

Анализ принципов построения системы мониторинга цифрового контента с целью решения задач безопасности 28 октября 2011 - 6 сентября 2012

Научно-исследовательский институт механики

Участники НИР: Афонин С.А., Бахтин А.В., Васенин В.А., Голомазов Д.Д., Занчурич М.А., Козицын А.С., Коршунов А.А., Шалченко К.А., Шундеев А.С., Ященко В.В.

Теоретические и экспериментальные исследования эффективности системы обеспечения безопасности информации в критически важных системах информационной инфраструктуры, построенных на основе технологий Grid Computing и Cloud Computing 1 июня 2012 - 1 сентября 2013

Научно-исследовательский институт механики

Участники НИР: Астапов И.С., Афонин С.А., Васенин В.А., Галатенко А.В., Галатенко В.В., Голомазов Д.Д., Занчурич М.А., Зензинов А.А., Итес А.А., Казьмин О.О., Козицын А.С., Коршунов А.А., Костюхин К.А., Кривчиков М.А., Першин И.С., Пучков Ф.М., Роганов В.А., Шалченко К.А., Шокуров А.В., Шундеев А.С.

Адрес Вашей общедоступной страницы: <http://istina.msu.ru/profile/goldan/>

Рис. 2.1. Личная страница сотрудника, сокращенный вариант.

2.2. Ввод данных

Во многих крупных библиографических системах используются средства автоматического сбора информации о результатах научной деятельности ученых. В качестве источников информации обычно выступают либо базы данных издательств, с которыми заключены соглашения о сотрудничестве, либо веб-сайты в Интернет. С одной стороны, этот подход позволяет автоматически (без участия самих авторов) обработать большие объемы данных и построить систему, охватывающую значительное число публикаций исследователей. С другой стороны, подобная автоматизация неизбежно приводит к существенному ухудшению точности (корректности, достоверности) и полноты извлекаемой информации. Под полнотой будем понимать:

- количество сотрудников по отношению к общему их числу в подконтрольной организации, результаты (показатели) деятельности которых введены в систему;
- количество типов результатов деятельности, обрабатываемых системой (например, публикации, патенты, отчеты, доклады на конференциях), по отношению к тем ее типам, которые могут быть потенциально востребованы;
- количество результатов деятельности по каждому из ее типов для отдельного ученого.

Под точностью (корректностью) будем иметь в виду следующие характеристики системы:

1. количество правильно (неправильно) введенных данных о результатах деятельности сотрудников (ученых) (правильность названий публикаций, номеров страниц и других данных) по отношению к общему числу данных;
2. эффективность (отношение корректно введенных к общему числу) механизмов разрешения неоднозначностей имен авторов работ (публикаций, патентов, диссертаций и других).

Анализ информации из нескольких наиболее крупных и широко используемых программных систем интеллектуального анализа данных научно-технического содержания, таких как Thomson Reuters Web of Knowledge, Scopus, Google Scholar, российские eLibrary.ru и MathNet.ru, позволяет сделать следующие выводы. В случае, когда информация извлекается из данных издательств-партнеров, ее полнота ограничена фиксированным перечнем их изданий. Например, даже в самых крупных комплексах подобного назначения, включая Web of Knowledge и Scopus, целые категории источников, в частности - российские журналы, представлены крайне слабо. В таких системах учитывается большое число работ многих ученых, однако при условии, что они попадают под заданные критерии, в частности, напечатаны в определенных журналах в течение некоторого промежутка времени. В eLibrary.ru, например, учитываются статьи с 1970 года. Отметим, что эти вопросы во многом решены в системе Google Scholar, в которую данные поступают из открытых источников Интернет. Однако и в ней обрабатывается информация только о публикациях и патентах. Кроме того, следует заметить, что механизмы верификации данных, обеспечивающие их корректность, в последних двух упомянутых системах очень ограничены.

Загрузка и проверка корректности крупных объемов данных требует трудоемкой работы экспертов, либо высокоэффективных интеллектуальных автоматических алгоритмов. Разработка таких алгоритмов – отдельная и очень сложная задача. Вместе с тем, их применение неизбежно снижает точность информации. Если данные поступают в структурированном виде, например, от издательств, то они обладают высоким уровнем корректности. При загрузке полуструктурированных данных из Интернет их корректность намного ниже. При этом в обоих случаях второй показатель, а именно – корректность разрешения неоднозначностей имен авторов, наиболее сложно удержать в нужном (приемлемом) диапазоне. Так, в существующих системах, как правило, встречаются ситуации, когда одному автору с распространенной фамилией (например, Ли, Смит или Иванов) приписаны десятки тысяч публикаций, на самом деле принадлежащих

десяткам или даже сотням различных авторов.

Для повышения точности и полноты информации в системе ИСТИНА выбран подход, отличающийся от тех, которые используются в существующих системах. Он основан на следующих соображениях. Предполагается, что автор работы (публикации, патента и т.п.) наиболее точно знает ее выходные данные, а также то, какие именно ученые перечислены в ее соавторах. Информация о соавторах является наиболее ценной, так как позволяет сопоставить авторов добавляемой работы с конкретными объектами в базе данных Системы. Процесс сопоставления, в частности, включает выбор искомого объекта, соответствующего автору, из множества однофамильцев в базе данных. Для эффективного автоматизированного выполнения подобной «привязки» используется следующий алгоритм. При добавлении работы, например, публикации, в Систему поступает список имен ее авторов в виде текстовой строки. Необходимо разбить эту строку на части и сопоставить им объекты (персоналии) из базы данных. Для этого выполняется поиск похожих объектов (ученых) в базе данных Системы, в процессе которого вычисляются несколько функций, именуемых «функциями похожести» (далее термин используется без кавычек), разделенных на два класса – функции сильной и слабой похожести. К функциям сильной похожести относятся проверки: совпадения полного имени; совпадения фамилии и инициалов; совпадения полного имени после добавления или пропуска (опускания) слов (например, «Иванов Иван Иванович» и «Иванов Иван»); совпадения фамилии и инициалов после добавления и пропуска инициалов (например, «Петров А.» и «Петров А.В.»). Функции слабой похожести основаны на алгоритме вычисления расстояния Левенштейна [7] и расширении алгоритма Рэтклиффа-Обершелпа [8] для сравнения двух последовательностей элементов. Каждому кандидату, ученому из базы данных Системы, присваивается определенный вес от 0 до 1, который равен максимальному значению функций похожести между ним и именем автора работы. После отбора и оценки всех кандидатов для каждого автора выбирается одно из трех возможных решений Системы. Во-первых, если для автора не найдено кандидатов с оценкой, превышающей пороговое значение, то Систе-

ма предлагает пользователю добавить в базу данных нового ученого. Это можно сделать непосредственно в процессе добавления публикации, выбрав соответствующую опцию. Во-вторых, если наибольшее значение функции похожести было достигнуто только на одном кандидате, то Система отображает всех кандидатов и выбирает по умолчанию лучшего из них. Наконец, если нескольким кандидатам присвоено одинаковое (и наибольшее) значение функции похожести, то Система отображает всех кандидатов и не делает выбора по умолчанию, предлагая пользователю выбрать нужного сотрудника самому.

Отметим, что у каждого пользователя есть возможность заранее добавить в свой профиль список своих «альтернативных имен», то есть других вариантов написания своего полного или краткого имени, которые используются в выходных данных его работ. Например, можно указать свою девичью фамилию, полное или краткое имя на английском или других языках и другие варианты имен. Эта информация используется при добавлении работ и позволяет эффективно привязать их авторов к нужным исследователям в Системе.

В Системе обрабатывается информация о результатах научной и педагогической деятельности следующих типов¹:

1. статьи в журналах;
2. статьи в сборниках;
3. книги;
4. доклады на конференциях;
5. участие в НИР;
6. патенты;
7. свидетельства о регистрации прав на программное обеспечение;

¹<http://istina.msu.ru/add/>

8. научные отчеты;
9. членство в редколлегиях журналов;
10. членство в редколлегиях сборников;
11. членство в программных комитетах конференций;
12. членство в диссертационных советах;
13. руководство и авторство диссертаций;
14. руководство дипломными работами;
15. авторство учебного курса;
16. преподавание (чтение) учебного курса;
17. научные коллективы;
18. результаты, связанные с использованием оборудования;
19. награды;
20. почетные членства в организациях;
21. членства в научных обществах;
22. стажировки.

Информацию о результатах каждого типа можно добавлять в Систему, просматривать, редактировать и удалять. Список типов постепенно расширяется. Кроме перечисленных выше типов, в Системе присутствуют функции добавления, просмотра и удаления так называемых «общих» объектов, которые связаны с этими типами. Далее представлен список общих объектов, для которых доступны часть или все эти функции:

1. журналы;
2. сборники;

3. конференции;
4. издательства;
5. диссертационные советы;
6. НИР.

Одним из главных результатов научной деятельности ученого является публикация. Многие ученые имеют значительное количество публикаций, поэтому крайне важно обеспечить удобный интерфейс ввода сведений о своих работах, позволяющий вносить данные в Систему за минимальное время. Для этого в систему ИСТИНА включен модуль распознавания библиографической ссылки (рис. 2.2), основанный на программе Freecite¹.

Шаг 1. Введите информацию о публикации

Что можно вводить? При вводе нескольких библиографических ссылок разделяйте их пустыми строками.

Внимание! Перед началом ввода данных, пожалуйста, ознакомьтесь с общими принципами работы с системой. Это займет у Вас 3 минуты, но сэкономит время в дальнейшем и поможет избежать распространенных ошибок.

В.А. Васенин, С.А. Афонин, А.С. Козицын. Автоматизированная система тематического анализа информации. Информационные технологии, 4, 1-32, 2009.

Ввод ссылки в виде обычной текстовой строки

Продолжить добавление публикации Очистить форму

Рис. 2.2. Добавление работы в систему ИСТИНА, шаг 1 - разбор данных.

Он позволяет ученому не заполнять большое число полей, как во многих других системах, а предоставляет возможность указать одну или несколько библиографических ссылок на свои публикации в виде текстовых строк. При этом на втором шаге добавления соответствующие поля уже будут автоматически заполнены. Пользователю

¹<http://freecite.library.brown.edu>

остается лишь проверить корректность разбора и, при необходимости, исправить ошибки (рис. 2.3), после чего он переходит на третий шаг.

Шаг 2. Отредактируйте данные о работе

В.А. Васенин, С.А. Афонин, А.С. Козицын. Автоматизированная система тематического анализа информации. Информационные технологии, 4, 1-32, 2009.

Основная информация

Тип работы: статья в журнале статья в сборнике книга

Авторы: Васенин В.А., Афонин С.А., Козицын А.С.

Название статьи: Автоматизированная система тематического анализа информации

Информация о статье в журнале

Журнал: Информационные технологии

Том:

Номер: 4

Год издания: 2009

Первая страница: 1

Последняя страница: 32

Разное

Результаты автоматического разбора оставляют желать лучшего:

Продолжить добавление публикации Вернуться на шаг 1 Очистить форму

информация автоматически разбивается по полям

значения можно откорректировать

Рис. 2.3. Добавление работы в систему ИСТИНА, шаг 2 - редактирование данных.

На третьем шаге (рис. 2.4) выполняется привязка имен авторов работы к объектам (научным сотрудникам) в базе данных. Система предлагает найденных похожих сотрудников и, если алгоритм позволяет сделать однозначный выбор, то выбирает одного из сотрудников по умолчанию. Если при этом в базе найдено несколько ученых с высокой степенью схожести на имя одного из авторов публикации, то пользователю предлагается сделать выбор самостоятельно. Он также может исправить выбор Системы.

Для того, чтобы облегчить выбор нужного похожего сотрудника, пользователь может оперативно, не покидая страницу, просмотреть краткую информацию о каждом из найденных сотрудников-кандидатов для выбора. Более подробное описание алгоритма подбора похожих ученых в базе данных для заданного имени автора

Шаг 3. Проверьте введенные данные

Информация о публикации

Автоматизированная система тематического анализа информации [статья в журнале](#)

журнал: Информационные технологии

Том: 4

Номер журнала: 4

Год издания: 2009

Первая страница: 1

Последняя страница: 32

Авторы. Выберите сотрудников из найденных в системе или добавьте новых

Автор	Похожие сотрудники в системе
Васенин В.А.	Васенин В.А. // добавить нового сотрудника // затрудняюсь ответить
Афонин С.А.	Афонин С.А. // добавить нового сотрудника // затрудняюсь ответить
Козицын А.С.	Козицын А.С. // добавить нового сотрудника // затрудняюсь ответить
Трофимов В.	Трофимов В.А. // Трофимов В.Т. // добавить нового сотрудника // затрудняюсь ответить

[Сохранить публикацию в систему](#) [Вернуться на предыдущий шаг](#)

Рис. 2.4. Добавление работы в систему ИСТИНА, шаг 3 - привязка авторов.

представлено ранее в настоящем разделе.

Многие исследователи ведут список своих работ в электронном виде, например, в файле на персональном компьютере или в различных информационных системах. Для того, чтобы им не приходилось вручную повторно вводить свои данные в систему ИСТИНА, в ней предусмотрены механизмы импорта данных. Поддерживаются следующие форматы и источники:

- библиографические ссылки, упомянутые ранее;
- BibTeX;
- Web of Knowledge / Web of Science;
- Scopus;
- Google Scholar;
- ResearcherID.com;
- PubMed;
- MathSciNet;
- Zentralblatt Math;
- MathNet.ru;
- веб-сайты издательств, например, Elsevier, Springer и другие.

При необходимости пользователь может не применять функции разбора или импорта и заполнить поля вручную. При этом, вне зависимости от того, какой способ ввода данных используется, для каждой работы пользователю необходимо пройти второй шаг добавления (проверка корректности заполнения полей) и третий шаг (привязка авторов к сотрудникам в Системе). При вводе публикаций к ним можно добавить аннотации и прикрепить файлы, например, с полным текстом работы или с сопутствующими материалами.

В Систему можно добавлять сразу много публикаций. Для этого данные о всех публикациях в соответствующем формате вводятся в форму, соответствующие механизмы Системы разбирают эти данные, и пользователь переходит к добавлению первой публикации.

Остальные публикации попадают в очередь добавления. После добавления текущей публикации пользователь может перейти к добавлению следующей публикации из очереди, а также он может очистить очередь добавления.

Отдельного упоминания заслуживает ввод информации о научных коллективах. После ввода данных пользователь может завершить работу над вводом информации о данном научном коллективе (функция «подписки»), и оповещение о новом научном коллективе отправляется в ректорат вуза или в дирекцию НИИ для проверки. Пользователь может также вывести специальную страницу для распечатки этих данных (рис. 2.5).

Экспериментальное определение фундаментальных термодинамических характеристик минералов, синтетических фаз и материалов

Experimental determination of fundamental thermodynamic characteristics of minerals, synthetic phases and materials

Описание	Методами высокотемпературной (1000oC) микрокалориметрии Кальве «Setaram» (Франция) проводится экспериментальное определение фундаментальных термодинамических параметров минералов, синтетических фаз и материалов: высокотемпературной теплоемкости, энтальпии образования, дегидратации, дегидроксидизации, разложения, фазовых превращений, процессов разупорядочения, образования твердых растворов. Термохимические исследования выполняются с использованием разработанных в лаборатории оригинальных экспериментальных методик высокотемпературной расплавленной калориметрии растворения и калориметрии «прямых реакций». Изучение проводится с привлечением современных диагностических методов: термического, термогравиметрического, рентгенографического, рентгенофлуоресцентного анализов, ИК-спектроскопии поглощения. Полученные термодинамические данные расширяют возможности применения термодинамического анализа к изучению физико-химических условий образования минералов, определению оптимальных условий технологических процессов, синтеза новых материалов.
Ключевые слова	Термодинамика, термохимия, калориметрия, теплоемкость, энтальпия образования, дегидратация, минерал
Описание на английском	The experimental determination of fundamental thermodynamic parameters of minerals, synthetic phases and materials (high-temperature heat capacity, enthalpies of formation, dehydration, dehydroxilation, decomposition, phase transformations, processes of disorder, mixing of solid solutions) is performed with high-temperature (1000oC) Calvet microcalorimetry «Setaram» (France). The thermochemical studies are made using the developed in the laboratory original experimental techniques of high-temperature melt solution calorimetry and "direct reaction" calorimetry. The investigation is performed using modern diagnostic methods: thermal, thermogravimetric, X-ray diffraction and X-ray fluorescence analysis, infrared absorption spectroscopy. The obtained thermodynamic data increase opportunities of the thermodynamic analysis to the study of physicochemical conditions of formation of minerals, determination of the optimal conditions of technological processes, synthesis of new materials.
Ключевые слова на английском	Thermodynamics, thermochemistry, calorimetry, heat capacity, enthalpy of formation, dehydration, mineral
Ведущее подразделение	Геологический факультет, Кафедра минералогии
Рубрики ГРНТИ	Физика минералов
Рубрики Scopus	Geology

Руководители и основные участники

1	Руководитель	Киселева Ирина Александровна	Доктор наук, старший научный сотрудник	ведущий научный сотрудник	Кафедра минералогии, Лаборатория физических методов исследования минералов
2	Участник	Огородова Любовь Петровна	Кандидат наук, старший научный сотрудник	старший научный сотрудник	Кафедра минералогии, Лаборатория физических методов исследования минералов
3	Участник	Мельчакова Любовь Васильевна	Кандидат наук, старший научный сотрудник	старший научный сотрудник	Кафедра минералогии, Лаборатория физических методов исследования минералов

Рис. 2.5. Информация о научном коллективе, версия для печати.

Важной особенностью Системы является однократный ввод данных. Такой ввод означает, что при добавлении новой работы одним из пользователей она одновременно появляется на страницах и у всех остальных соавторов работы. При этом пользователи могут легко заметить появление новой работы, которую добавил их соавтор, на специальной странице «мои последние добавленные работы». В случае неправильной привязки публикации соавтором пользователь может отказаться от ее авторства, и она исчезает с его страницы.

Для облегчения ввода в Системе используется механизм автодополнения различных данных. Результат его использования означает, что когда пользователь вводит, например, первые буквы фамилии автора, то Система подсказывает ему список существующих авторов с фамилией, которая начинается на эти символы. Автодополнение работает для фамилий сотрудников, названий издательств, проектов, организаций, журналов, рубрик журналов, сборников, серий, публикаций, диссертационных советов и конференций. Этот перечень в будущем может быть расширен.

Пользователи имеют возможность вводить данные в Систему, заполняя поля форм вручную. Однако, такой способ не всегда бывает удобным. Например, в случае, когда надо добавить все статьи, опубликованные пользователем в каком-либо журнале, ручной ввод большого количества данных может занять продолжительное время. Для устранения подобных неудобств ИСТИНА поддерживает возможность автоматизированного добавления библиографических ссылок из файлов.

Пользователь, как правило, может сравнительно легко получить список своих публикаций в журналах в виде, например, BibTeX-файла. Для добавления всех записей из этого файла ему достаточно передать файл Системе на обработку через форму на странице импорта. После этого все данные будут проанализированы Системой, и соответствующие записи будут добавлены в базу. Эти записи создаются без «привязки» к уже имеющимся в Системе объектам. Однако, при необходимости, пользователь может передавать в качестве данных идентификаторы уже существующих сущностей в базе. В этом случае Система при добавлении сделает

«привязку» новых данных к существующим объектам.

Модуль разбора библиографических данных в системе ИСТИНА поддерживает следующие форматы: библиографические ссылки в текстовом виде, BibTeX, MODS, данные из систем Elibrary, Web of Science, Scopus, ResearcherID.com. По мере развития функциональных возможностей Системы будут разработаны и добавлены механизмы поддержки других типов входных данных.

2.3. Отображение данных

Одним из основополагающих принципов системы ИСТИНА является открытость информации. Его суть в том, что сведения о работах, добавляемые пользователями Системы, в настоящее время открыты для просмотра всем пользователям Интернет. Для просмотра данных пользователю нет необходимости регистрироваться в Системе. При этом некоторые статистические отчеты по подразделениям и всей организации доступны только руководителям этих подразделений и организации. Для реализации политики разграничения доступа к функциям Системы используется механизм модераторов, прикрепленных к вершинам дерева подразделений организации.

Доступными для просмотра всеми пользователями Интернет являются следующие страницы.

- Страница сотрудника с краткой информацией о нем и списком всех добавленных его работ (рис. 2.6).
- Список всех организаций, зарегистрированных в Системе, с краткой информацией о них.
- Страница организации/подразделения со списком дочерних подразделений (рис. 2.7).
- Список всех зарегистрированных в Системе сотрудников (или пользователей) организации/подразделения с краткой информацией о каждом из них (рис. 2.8).



отправить
сообщение

Садовнический Виктор Антонович пользователь

МГУ имени М.В. Ломоносова, Механико-математический факультет, Отделение математики, Кафедра математического анализа, с 1 сентября 1967

доктор физико-математических наук с 1974 года

академик РАН с 1 сентября 1997 г.

профессор по кафедре математического анализа

Соваторы: Дубровский В.В., Александров В.В., Окунев Ю.М., Подольский В.Е., Винокуров В.А., Архипов Г.И., Самсонов В.А., Ильин В.А., Моисеев Е.И., Чубариков В.Н., Белокуров В.В., Акаев А.А., Розов Н.Х. [вексель полезность...](#)

654 статьи, 150 книг, 26 докладов на конференциях, 81 тезисы докладов, 10 патентов, 4 свидетельства о регистрации прав на ПО, 2 отчета, 9 членств в редакционных журналах, 34 членства в редакционных сборниках, 5 членств в програмных комитетах, 3 членства в диссертационных советах, 36 диссертаций, 4 учебных курса

Количество цитирований статей в журналах по данным Scopus: 161, Web of Science: 255

краткая информация о сотруднике

Деятельность обычный стиль: [обычный](#) | [ГОСТ](#) | [plain](#) | [abbr](#) | [acm](#) | [alpha](#) | [amslaTeX](#) | [amsplain](#) | [arXiv](#) | [ieeetr](#) | [siam](#)

СТАТЬИ В ЖУРНАЛАХ

- 2014** Application of molecular modeling to urokinase inhibitors development
Султмов В.Б., Каткова Е.В., Оферкин И.В., Султмов А.В., Романов А.Н., Рощин А.И., Белоглазова И.Б., Плеханова О.С., Трачук В.А., Садовнический В.А.
в журнале *BioMed Research International* DOI
- 2014 Hamiltonian Functional Integrals Representing the Traces of Differential Operators of Highest Orders
Sadovnichiy V.A., Smolyanov O.G., Shargalidze E.T.
в журнале *Doklady Mathematics*, издательство *Maik Nauka/Interperiodica Publishing (Russian Federation)*, том 456, № 1, с. 23-27
- 2014 Орторкурсивные разложения
Галатенко В.В., Лукашенко Т.П., Садовнический В.А.
в журнале *Современные проблемы математики и механики*, том 9, № 2, с. 13-25
- 2014 Параметрический хаос в нелинейных флаттерных системах
Садовнический В.А., Колесов А.Ю., Розов Н.Х.
в журнале *Доклады Академии Наук РФ*, том 456, № 2, с. 150-154 DOI
- 2014 "Space Experiments aboard the Lomonosov MSU Satellite"
Sadovnichiy V.A., Amelnyushkin A.M., Angelopoulos V., Bengin V.V., Bogomolov V.V., Garpov G.K., Gorbovskoy E.S., Grossan V., Klimov P.A., Khrenov B.A., Lee J., Lipunov V.M., Ita G.W., Panasyuk M.I., Park I.H., Petrov V.L., Russell C.T., Svertlov S.I., Sigavea E.A., Smoot G.F., Shprits Y., Veden'kin N.N., Yashin I.V.
в журнале *Cosmic Research (English translation of Kosmicheskie Issledovaniya)*, издательство *Maik Nauka/Interperiodica Publishing (Russian Federation)*, том 52, с. 250-250
- 2013 A Remark on the most informative EEG Signal Components in a super-scalable method for functional state classification based on the wavelet decomposition
Galatenko V.V., Lvshitz E.D., Staroverov V.M., Lukashenko T.P., Galatenko A.V., Podolskii V.E., Sadovnichiy V.A., Lebedev V.V., Isaychev S.A., Chernorizov A.M., Zinchenko Yu.P.
в журнале *Procedia - Social and Behavioral Sciences*, том 86, с. 18-23 DOI
- 2013 Automated real-time classification of functional states: significance of individual tuning stage
Galatenko Vladimir V., Lvshitz Evgeniy D., Chernorizov Alexander M., Zinchenko Yury P., Galatenko Alexey V., Staroverov Vladimir M., Isaychev Sergey A., Lebedev Vyacheslav V., Menshikova Galina Ya, Gusev Alexey N., Lobacheva Ekaterina M., Gabidullina Pozaliya F., Podolskii Vladimir E., Sadovnichiy Victor A.

в профиле отображаются все введенные данные о работах всех типов

Рис. 2.6. Страница сотрудника в системе ИСТИНА.

МГУ имени М.В. Ломоносова организация

10219 зарегистрированных сотрудников [статистика]

[Список диссертационных советов](#)

[Подразделения](#)

[Биологический факультет](#)

[Биотехнологический факультет](#)

[Высшая школа бизнеса](#)

[Высшая школа государственного администрирования](#)

[Высшая школа государственного аудита](#)

[Высшая школа инновационного бизнеса](#)

[Высшая школа культурной политики и управления в гуманитарной сфере](#)

[Высшая школа перевода](#)

[Высшая школа современных социальных наук](#)

[Высшая школа телевидения](#)

[Высшая школа управления и инноваций](#)

[Географический факультет](#)

[Геологический факультет](#)

[Государственный астрономический институт имени П.К.Штернберга](#)

Рис. 2.7. Страница организации в системе ИСТИНА.

МГУ имени М.В. Ломоносова

Список сотрудников

Страницы: << предыдущая 1 2 3 4 5 6 следующая >>

**Зефилов Николай Серафимович**

zefirov

МГУ имени М.В. Ломоносова, Химический факультет, с 1 августа 2014
 Сопроводители: Палулин В.А., Баскин И.И., Зык Н.В., KOZMIN A., Кузнецова Т.С., Potekhin K.A., Аверина Е.Б., ZHDANKIN V. и др.
 1825 статей, 4 книги, 199 докладов на конференциях, 228 тезисов докладов, 1 НИР, 76 патентов, 26 отчетов, 2 награды, 1 членство в научном обществе, 10 членств в редколлегиях журналов, 21 членство в программных комитетах, 2 членства в диссертационных советах, 55 диссертаций, 1 учебный курс

**Трофимов Виктор Титович**

TrofimovVT

МГУ имени М.В. Ломоносова, Геологический факультет, Кафедра инженерной и экологической геологии, с 1 сентября 1960
 Сопроводители: Зиллинг Д.Г., Харькина М.А., Барабошкина Т.А., Королёв В.А., Богословский В.А., Красилов Н.С., Аверина Т.И., Герасимова А.С. и др.
 925 статей, 83 книги, 39 докладов на конференциях, 147 тезисов докладов, 2 патента, 4 отчета, 10 членств в редколлегиях журналов, 18 членств в редколлегиях сборников, 34 членства в программных комитетах, 2 членства в диссертационных советах, 14 диссертаций, 13 дипломных работ, 33 учебных курса

**Урсул Аркадий Дмитриевич**

UrsuAD

МГУ имени М.В. Ломоносова, Факультет глобальных процессов, Кафедра глобалистики, с 1 сентября 2013
 Сопроводители: Урсул Т.А., Ильин И.В., Ващенко Н.П., Семенов Э.П., Готт В.С., Романович А.Л., Школенко Ю.А., Лось В.А. и др.
 945 статей, 198 книг, 55 докладов на конференциях, 51 тезис докладов, 21 награда, 16 членств в научных обществах, 2 почетного членства в организациях, 15 членств в редколлегиях журналов, 10 членств в редколлегиях сборников, 6 членств в программных комитетах, 12 диссертаций, 4 учебных курса

**Ишханов Борис Саркисович**

ishkhanov

МГУ имени М.В. Ломоносова, Физический факультет, Отделение ядерной физики, Кафедра общей ядерной физики, с 1 апреля 1967
 МГУ имени М.В. Ломоносова, Научно-исследовательский институт ядерной физики имени Д.В. Скобельцина, Отдел электромагнитных процессов и взаимодействия атомных ядер, с 16 апреля 1964
 Сопроводители: Капитонов И.М., Варламов В.В., Шведун В.И., Пискарев И.М., Бобошин И.Н., Мокеев В.И., Орлин В.Н., Романовский Е.А. и др.
 605 статей, 70 книг, 167 докладов на конференциях, 162 тезисов докладов, 1 НИР, 3 патента, 14 отчетов, 8 наград, 1 членство в научном обществе, 3 членства в редколлегиях журналов, 3 членства в редколлегиях сборников, 5 членств в программных комитетах, 2 членства в диссертационных советах, 33 диссертации, 2 дипломные работы, 37 учебных курсов

**Хохлов Алексей Ремович**

KhokhlovAR

МГУ имени М.В. Ломоносова, Физический факультет, Отделение физики твердого тела, Кафедра физики полимеров и кристаллов, с 14 июля 1979
 Сопроводители: Khalatov P.G., Филиппова О.Е., Махаева Е.Е., Стародубцев С.Г., Vasilevskaya V.V., Крамаренко Е.Ю., Галлямов М.О., Keshtov M.L. и др.
 745 статей, 3 книги, 19 докладов на конференциях, 19 тезисов докладов, 2 НИР, 22 патента, 2 свидетельства о регистрации прав на ПО, 3 членства в редколлегиях журналов, 1 членство в диссертационном совете, 13 диссертаций, 1 дипломная работа, 5 учебных курсов

**Золотов Юрий Александрович**

ZolotovYuA

МГУ имени М.В. Ломоносова, Химический факультет, Кафедра аналитической химии, с 1 января 1978
 Сопроводители: Дмитриенко С.Г., Цзинь Г.И., Моросанова Е.И., Петрушин О.М., SPYIAKOV B., Алпри В.В., Кузьмин Н.М., Шлигин О.А. и др.
 676 статей, 20 книг, 71 доклад на конференциях, 64 тезисов докладов, 17 патентов, 3 отчета, 1 награда, 1 членство в диссертационном совете, 4 диссертации, 1 учебный курс

**Трофимов Вячеслав Анатольевич**

Vyacheslav_Trofimov

МГУ имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Кафедра вычислительных методов, с 2 мая 1983
 Сопроводители: Варенцова С.А., Лысак Т.М., Сухоруков А.П., Trofimov V.V., Карамзин Ю.Н., Логинова М.М., Volkov A.G., Федотов М.В. и др.
 486 статей, 2 книги, 294 доклада на конференциях, 252 тезисов докладов, 1 членство в научном обществе, 2 членства в редколлегиях журналов, 4 членства в программных комитетах, 1 членство в диссертационном совете, 13 диссертаций, 10 дипломных работ, 9 учебных курсов

Рис. 2.8. Страница со списком сотрудников организации в системе ИСТИНА.

- Страница со списком всех добавленных в Систему работ сотрудниками организации/подразделения за последнее время (по умолчанию - за последние две недели).
Для этой страницы существует также специальная PDF-версия для использования в качестве отчета (рис. 2.9).
- Страница со списком всех добавленных в Систему работ за последний месяц.
- Страница работы (например, публикации) с полной информацией о ней (рис. 2.10).
На этой странице сведения о публикации можно отобразить и сохранить в файл в различных стилях и форматах (BibTeX, EndNote, RIS, Word, ISI, Ads). Пользователь также может загрузить прикрепленные к работе файлы.
- Страница связанного с работами объекта, например, конференции, журнала, сборника, серии, диссертационного совета и других (рис. 2.11).
На страницах издательства, журнала и сборника отображены списки статей и книг, напечатанных в них. Эти списки можно представить в различных стилях и экспортировать в формат BibTeX.
- Страница списка (индекса) журналов, в частности, перечня ВАК (рис. 2.12), таких индексов, как Science Citation Index, Arts & Humanities Citation Index, Current Contents, Medline, Index medicus и других (всего 23 индекса).
- Страница поиска сотрудников, публикаций и журналов (рис. 2.13).
- Различные вспомогательные страницы, например, руководство пользователя (рис. 2.14).

Кроме перечисленных выше страниц, в Системе есть также страницы со списком всех добавленных публикаций, разбитых по году опубликования. Эти списки можно представить в различных стилях, а также экспортировать в формат BibTeX.

2.3. ОТОБРАЖЕНИЕ ДАННЫХ

МГУ имени М.В.Ломоносова

Механико-математический факультет
Работы, добавленные с 20 сентября по 4 октября 2014 года

№	Год	Тип	Название	Сотрудник(и)
1	2009	статья	Isomorphisms of graded endomorphidm rings of graded modules that are closed to free ones	Balaba I.N., Mikhalev A.V.
2	2005	статья	Universal central extensions of conformal Lie algebras	Mikhalev A.V., Pinchuk I.A.
3	2010	статья	Universal central extensions of conformal Lie algebras, part 2: The supercase	Mikhalev A.V., Pinchuk I.A.
4	2011	статья	Isomorphisms of general linear groups over associative rings graded by an abelian group	Atkarskaya A.S., Bunina E.I., Mikhalev A.V.
5	2014	доклад на конференции	Cohomology of Landweber-Novikov algebra and singular Virasoro vectors	Миллионщиков Д.В.
6	2009	статья	Jacobi's bound for systems of algebraic differential equations	Kondratieva M.V., Mikhalev A.V., Pankrat'ev E.V.
7	2011	статья	Isomorphisms of general linear groups over associative rings graded by an abelian group	Atkarskaya A.S., Bunina E.I., Mikhalev A.V.
8	2012	статья	The topological Jacobson radical of rings, I	Glavatskii S.T., Mikhalev A.V., Tenzina V.V.
9	2012	статья	The topological Jacobson radical of rings, II	Glavatskii S.T., Mikhalev A.V., Tenzina V.V.
10	2012	статья	Quotient rings of graded associative rings, I	Balaba I.N., Kanunnikov A.L., Mikhalev A.V.
11	2014	диплом	Эндоморфизмы полугрупп неотрицательных обратимых матриц	Бушина Е.И., Михалёв А.В.
12	2013	диплом	Оценки высоты в смысле Ширшова количества фрагментов малого периода	Белов А.Я., Михалёв А.В.
13	2013	диплом	Подъем полиномиальных автоморфизмов и размерность Гельфанда-Кириллова	Белов А.Я., Михалёв А.В.
14	2013	диплом	Первичный радикал упорядоченных лул	Михалёв А.В.
15	2012	диплом	Параметрв кодов, построенных на основе кольца матриц с помощью функции следа	Марков В.Т., Михалёв А.В., Нечаев А.А.
16	2012	диплом	Периодичность морфических слов	Белов А.Я., Михалёв А.В.
17	2011	диплом	Числа Фробениуса полугрупп	Михалёв А.В.
18	2011	диплом	Автоморфизмы групп Шевалле типа G_2 над локальными кольцами без $1/2$	Бушина Е.И., Михалёв А.В.
19	2011	диплом	Стандартные и нестандартные автоморфизмы групп Шевалле над кольцами	Бушина Е.И., Михалёв А.В.

Рис. 2.9. Список всех добавленных в Систему работ сотрудниками подразделения за последние 2 недели, PDF-версия.



ИСТИНА

Интеллектуальная Система Тематического Исследования Научно-технической информации

Здравствуйте, Голомазов Денис Дмитриевич (goldan)!

[Главная](#) [Моя страница](#) [Добавить работу](#) [Поиск](#) [Выйти из системы](#) [О проекте](#) [Помощь](#)

Использование систем семантического анализа для организации поиска научно-технической информации [статья](#)

Авторы: Афонин С.А., Голомазов Д.Д., Козицын А.С.

Журнал: Программная инженерия

Номер: 1

Год издания: 2012

Первая страница: 29

Последняя страница: 34

Добавил в систему: Афонин Сергей Александрович

Работа с статьей

Показать публикацию в формате:
[BibTeX](#) | [EndNote](#) | [RIS](#) | [Word](#) | [ISI](#) | [ADS](#)
[Редактировать публикацию](#)
[Добавить аннотацию](#)
[Добавить DOI](#)
[Прикрепить файл](#)
[Удалить публикацию](#)

Библиографическая ссылка на публикацию в стиле [plain](#) | [abbrv](#) | [acm](#) | [alpha](#) | [amsalpha](#) | [amspain](#) | [apalike](#) | [ieeetr](#) | [siam](#)

[1] С.А. Афонин, Д.Д. Голомазов, and А.С. Козицын. Использование систем семантического анализа для организации поиска научно-технической информации. *Программная инженерия*, (1):29–34, 2012. [[http](#)]

Рис. 2.10. Страница публикации в системе ИСТИНА.

Для анализа и индексации поисковыми системами в Системе доступны также страницы со списком всех сотрудников (в кратком и расширенном формате), и списком всех сотрудников, у которых не указано место работы.

2.4. Формирование отчетов

Одним из преимуществ системы ИСТИНА является то обстоятельство, что она облегчает пользователям выполнение некоторых рутинных задач, неизбежно возникающих в процессе подготовки, организации научной работы и подготовки отчетных материалов. К таким задачам, в частности, относятся: подача заявок на проекты и гранты; формирование научных отчетов. Для решения первой задачи в Системе существует возможность отображения списка своих публикаций в различных стилях, например: в алфавитном порядке (стиль «plain»); с сокращением инициалов авторов и названий журналов (стиль «abbrv»); с использованием в качестве указателей не порядковых номеров публикаций в списке, а первых букв фамилий авторов

Nature

журнал

Издательство: Nature Publishing Group

Местоположение издательства: United Kingdom

Количество статей в журнале: 102

Редактировать информацию о журнале

Статьи, опубликованные в журнале

Библиография

стиль: **обычный** | ГОСТ | plain | abbrev | acm | alpha | amsalpha | amsplain | apalike | ieee | slam

Страницы: << предыдущая 1 2 3 4 5 6 следующая >>

- ★ 2014 Interacting supernovae from photoionization-confined shells around red supergiants
 - 👤 Mackey J., Mohamed S., Gvaramadze V.V., Kotak R., Langer N., Meyer D.M.-A., Moriya T.J., Nelson H.R.
 - 📄 в журнале Nature, издательство Nature Publishing Group (United Kingdom), том 512, № 7514, с. 282-285 DOI

- ★ 2014 Neutrinos from the primary proton–proton fusion process in the Sun
 - ✓ Bellini G., Benziger J., Bick D., Bonfini G., Bravo D., Caccianiga B., Cadonati L., Calaprice F., Caminata A., Cavalcante P., Chavarria A., Chepurinov A., D'Angelo D., Davini S., Derbin A., Empl A., Etenko A., Fomenko K., Franco D., Gabriele F., Galbiati C., Gazzana S., Ghiano C., Giammarchi M., Göger-Neff M., Goretti A., Gromov M., Hagner C., Hungerford E., Aldo Ianni, Andrea Ianni, Kobychew V., Korabiev D., Korga G., Krym D., Laubenstein M., Lehnert B., Lewke T., Litvinovich E., Lombardi F., Lombardi P., Ludhova L., Lukyanchenko G., Machulin I., Manecki S., Maneschg W., et al S.Marocci, Borexino Collaboration
 - 📄 в журнале Nature, издательство Nature Publishing Group (United Kingdom), том 512, № 7515, с. 383-386 DOI

- ★ 2013 Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga
 - 👤 Fict Jean-Francois, Hespelaris Boris, Xiang Li, Noel Benjamin, Arkhipova Irma, Danchin Etienne G.J., Hejnol Andreas, Henrissat Bernard, Koszul Romain, Aury Jean-Marc, Barbe Valerie, Barthelemy Roxane-Marie, Bast Jens, Bazhkin Georgii A., Chabrol Olivier, Couloux Arnaud, Marline Da Rocha, Corinne Da Silva, Gladyshev Eugene, Gouret Philippe, Hallatschek Oskar, Hecox-Lea Bette, Labadie Karine, Lejeune Benjamin, Piskurek Oliver, Poulain Julie, Rodriguez Fernando, Ryan Joseph F., Vakhtusheva Olga A., Wajsborg Eric, Wittm Benedicte, Yushenova Irina, Kelis Manolis, Kondrashov Alexey S., Welch David B.Mark, Pontarotti Pierre, Weissenbach Jean, Wincker Patrick, Jaillon Olivier, Van Doninck Karine
 - 📄 в журнале Nature, издательство Nature Publishing Group (United Kingdom), том 500, № 7463, с. 453-457 DOI

- ★ 2013 North Atlantic Ocean control on surface heat flux on multidecadal timescales
 - ✓ Gulev S.K., Latif M., Keenlyside N., Park W., Koltermann K.P.
 - 📄 в журнале Nature, издательство Nature Publishing Group (United Kingdom), том 499, № 7459, с. 464-467 DOI

- ★ 2013 Russian science: Academy reform needs a reality check
 - 👤 Razin S.V., Vassetzky E.S.
 - 📄 в журнале Nature, издательство Nature Publishing Group (United Kingdom), том 499, № 7458, с. 284-284 DOI

- ★ 2012 366 days: Images of the year
 - 👤 Daniel Cressley, Mark Perlow
 - 📄 в журнале Nature, издательство Nature Publishing Group (United Kingdom), № 492, с. 328-333

- ★ 2012 Coherent quantum phase slip
 - ✓ Astafiev O.V., Ioffe L.B., Katanov S., Pashkin Yu A., Arutyunov Yu K., Shahar D., Cohen O., Tsai J.S.
 - 📄 в журнале Nature, издательство Nature Publishing Group (United Kingdom), том 484, № 7394, с. 355-358 DOI

- ★ 2012 How to build science capacity
 - 👤 Fanaroff Bernie, Aebischer Patrick, Camacho Rafael, Serageldin Ismail, Suetin Nikolay, Carlos Henrique De Brito Cruz, Lee June Seung
 - 📄 в журнале Nature, издательство Nature Publishing Group (United Kingdom), том 490, № 7420, с. 331-334

Рис. 2.11. Страница журнала в системе ИСТИНА.

Список ВАК список журналов

Показывать по записей на странице

Поиск:

№	Журнал
131	Биофизика
132	Биохимия
133	Биоэтика
134	Биржа интеллектуальной собственности (БИС)
135	Боль
136	Ботанический журнал
137	Бурение и Нефть
138	Бутлеровские сообщения
139	Бухгалтерский учет
140	Бухучет в здравоохранении

Показаны записи с 131 по 140. Всего: 2,133.
[Предыдущая](#) [Следующая](#)

Рис. 2.12. Страница индекса журналов в системе ИСТИНА.



ИСТИНА

Интеллектуальная Система Тематического Исследования НАучно-технической информации

МГУ Здравствуйте, Голомазов Денис Дмитриевич (goldan) !

[Главная](#) [Моя страница](#) [Добавить работу](#) [Поиск](#) [Выйти из системы](#) [О проекте](#) [Помощь](#)

Найти сотрудника: Поиск введите фамилию сотрудника

Найти публикацию: Поиск введите название публикации

Найти журнал: Поиск введите название журнала

Рис. 2.13. Страница поиска в системе ИСТИНА.

Инструкция по работе с системой ИСТИНА

1. Зарегистрируйтесь
2. Найдите себя в базе сотрудников
3. Укажите свое место работы
4. Добавьте свои работы

- инструкция по добавлению работ из других систем

Общие принципы работы с системой

Часто задаваемые вопросы

Не получается зарегистрироваться или войти в систему. Что делать?

Следуйте инструкциям на странице Проблемы с регистрацией или входом в систему.

Обратная связь

Если при работе с системой у вас возникают какие-либо вопросы, проблемы или предложения, воспользуйтесь системой обратной связи. Для этого нажмите на ссылку "обратная связь" в самом низу любой страницы сайта и следуйте инструкциям на открывшейся странице.

Рис. 2.14. Страница помощи в системе ИСТИНА.

и года публикаций (стиль «alpha»), а также в других шести стилях. В Системе присутствует возможность экспорта сведений об отдельной публикации в форматы BibTeX, EndNote, RIS, Word, ISI и ADS и экспорта списка публикаций в формат BibTeX. Отметим, что эти форматы поддерживаются в большинстве распространенных систем для работы с библиографическими данными, что облегчает ученому повторное использование информации, введенной в систему ИСТИНА.

В системе ИСТИНА существует возможность формирования отчетов для сотрудников на основе введенных данных. Годовой отчет (рис. 2.15) содержит перечень результатов научной деятельности за текущий календарный год, которые были введены в Систему. Данная информация может использоваться при подаче сотрудниками годовых научных отчетов в своих подразделениях. Отметим, что в 2013 году в МГУ проводилось тестирование подачи годовых научных отчетов в электронной форме, сформированных с помощью системы ИСТИНА, наряду с подачей в традиционной бумажной форме. С 2014 года планируется постепенно полностью переводить процедуру формирования и подачи годовых отчетов в электронный вид с помощью механизмов Системы.

Информационный лист (рис. 2.16) включает краткую информа-

Васенин Валерий Александрович

Научный отчет за 2012 год

Статьи

2012 Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА)

Васенин В.А., Афонин С.А., Козицын А.С., Голомазов Д.Д., Бахтин А.В., Ганин Г.М.
в журнале *Обозрение прикладной и промышленной математики*, том 19, № 2, с. 239-240

2012 К созданию защищенной наложенной сети связи нового поколения для передачи данных

Васенин В.А., Левин В.Ю., Пучков Ф.М., Шалченко К.А.
в журнале *Программная инженерия*, № 4, с. 2-8

2012 Модернизация экономики и новые аспекты инженерии программ

Васенин В.А.
в журнале *Программная инженерия*, № 2, с. 2-17

2012 Обеспечение информационной безопасности в распределенных системах на основе технологий Grid и Cloud

Computing: традиционные средства защиты и вопросы асимметрии доверия
Васенин В.А., Иткес А.А., Пучков Ф.М., Шалченко К.А.
в журнале *Программная инженерия*, № 1, с. 28-34

2012 Эволюция технологии грид

Васенин В.А., Шундеев А.С.
в журнале *Информационные технологии*, № 1, с. 2-9

Отчеты

2012 Анализ алгоритмов мониторинга цифрового контента с целью решения задач безопасности

Авторы: Васенин В.А., Афонин С.А., Козицын А.С., Голомазов Д.Д., Коршунов А.А., Титов А.С., Шундеев А.С., Бахтин А.В.
#5151, 248 с.

Участие в редколлегии журналов

с 1 марта 2010 Программная инженерия

Рис. 2.15. Годовой отчет сотрудника в системе ИСТИНА.

цию о сотруднике и его деятельности за последние 5 лет. В частности, такая информация содержит количество статей и монографий, библиометрические показатели (например, количество цитирований его статей, индекс Хирша), краткие сведения о членстве в редколлегиях журналах и сборников, в программных комитетах конференций, в диссертационных советах, данные о преподавательской деятельности, о прочитанных за последние 5 лет учебных курсах.

Информационный лист может использоваться в качестве документа для процедуры переизбрания сотрудника на научно-преподавательские должности в своем подразделении. Отметим, что информационный лист может быть представлен в формате PDF. Причем этот файл формируется не с помощью веб-интерфейса Системы, как остальные страницы, а путем полностью автоматического создания и компиляции TeX-файла с соответствующими данными. Именно по этой причине он выглядит в большей степени соответствующим требованиям и форматам настоящих отчетов, которые, как правило, используются в научных организациях и вузах.

Кроме перечисленных двух основных отчетов, в Системе для каждого сотрудника присутствуют еще две страницы со списками статей. На первой из них перечислены статьи сотрудника, опубликованные в журналах из перечня ВАК (рис. 2.17). На второй – статьи, информация о цитировании которых была получена из внешних систем (Web of Science и Scopus).

2.5. Статистика

Важной для руководителей структурных подразделений и организации в целом функцией Системы является возможность анализа эффективности научной деятельности сотрудников и подразделений. В Системе реализованы методы интегрального анализа публикационной активности сотрудников организации и средства генерации сводных данных о деятельности сотрудников. Эти функции Системы описываются далее.

Научно-исследовательский институт механики
 404 Лаборатория автоматизации экспериментальных исследований
Избрание по конкурсу на должность заведующего отделом

ВАСЕНИН Валерий Александрович (1948 г.р.)

МГУ имени М.В. Ломоносова, Механико-математический факультет (1972)

доктор физико-математических наук (с 1998 года), профессор по кафедре вычислительной математики (с 2001 года)

Публикационная активность	количество работ	
	за 5 лет	всего
Всего статей в научных журналах	32	62
В том числе:		
в российских журналах из списка ВАК	29	49
в зарубежных журналах из списка ВАК	2	7
в журналах из top25	0	0
Статьи в сборниках	14	130
Число монографий	0	12
Учебно-методические работы	0	0
Библиометрические показатели (по данным Web of Science)		
H-индекс		1
Общее число ссылок		6
Число ссылок на статьи, опубликованные за последние 5 лет		0

Педагогическая деятельность за последние 5 лет	период	часов в год
Математические модели в обеспечении информационной безопасности обязательная, по выбору (спекурсе), лекции, Кафедра вычислительной математики	с 2011	36
Среднегодовая лекционная нагрузка		21,6

Под научным руководством защищено 83 дипломных работ, 18 кандидатских и 2 докторских диссертаций.

Членство в редколлегиях журналов или сборников (за 5 лет)		
Журнал: <i>Программная инженерия</i>	2010	по н.вр.
Журнал: <i>Нейрокомпьютеры: разработка, применение</i>	2010	по н.вр.
Журнал: <i>Проблемы информатики</i>	2006	по н.вр.

Рис. 2.16. Информационный лист сотрудника в системе ИСТИНА.

Голомазов Денис Дмитриевич

Статьи в журналах из перечня ВАК

Всего: 5.

- 1 **2013** Интеллектуальная Система Тематического Исследования Научно-технической информации (ИСТИНА)
Васенин В.А., Афонин С.А., Голомазов Д.Д., Козицын А.С.
в журнале *Информационное общество*
- 2 **2012** Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА)
Васенин В.А., Афонин С.А., Козицын А.С., Голомазов Д.Д., Бахтин А.В., Ганкин Г.М.
в журнале *Обозрение прикладной и промышленной математики*, том 19, № 2, с. 239-240
- 3 **2012** Использование систем семантического анализа для организации поиска научно-технической информации
Афонин С.А., Голомазов Д.Д., Козицын А.С.
в журнале *Программная инженерия*, № 1, с. 29-34
- 4 **2011** Использование семантических технологий для обнаружения Грид-ресурсов
Васенин В.А., Афонин С.А., Голомазов Д.Д.
в журнале *Программная инженерия*, № 7, с. 2-8
- 5 **2010** Выделение терминов из коллекции текстов с заданным тематическим делением
Голомазов Д.Д.
в журнале *Информационные технологии*, № 2, с. 8-13

Рис. 2.17. Список статей сотрудника в журналах из перечня ВАК.

2.5.1. Статистические данные публикационной активности

В качестве параметров анализа выступают следующие характеристики.

1. Тип публикации:

- статьи в журналах, входящих в лучшие по импакт-фактору 25% журналов в своей тематической рубрике согласно данным базы Thomson Reuters Journal Citation Reports¹;
- статьи в журналах, включенных в базу Thomson Reuters Journal Citation Reports;
- статьи в журналах, включенных в базу Scopus;
- все статьи в журналах;

¹http://thomsonreuters.com/products_services/science/science_products/az/journal_citation_reports/

- статьи в сборниках трудов конференций (исключая тезисы);
- все публикации, добавленные в базу данных Системы.

2. Статус публикации:

- проверенные (верифицированные) публикации, корректность выходных данных о которых была вручную проверена и подтверждена специально назначенными ответственными сотрудниками;
- все публикации.

3. Год публикации:

- публикации, вышедшие в печать в 2014 году;
- публикации, вышедшие в печать в 2013 году;
- публикации, вышедшие в печать в 2012 году;
- публикации, вышедшие в печать в 2011 году;
- публикации, вышедшие в печать в 2010 году;
- все публикации.

4. Подразделение, в котором работает сотрудник-автор публикации:

для анализа можно отобрать подразделение, находящееся на любом уровне структурной иерархии организации.

5. Тематика журнала, в котором опубликована работа:

для анализа можно отфильтровать публикации по тематике журнала. В Системе поддерживается два рубрикатора: Scopus и ГРНТИ.

6. Метрика, то есть количественная характеристика, по которой рассчитываются показатели отдельных подразделений и тематических направлений.

К числу таких показателей относятся следующие далее.

- Общее число публикаций, удовлетворяющих критериям поиска.

- Общее число цитирований публикаций по данным Web of Science. Показатели цитирования отдельных публикаций попадают в Систему из Web of Science с помощью инструментального средства Article Match Retrieval¹. Через некоторое время после добавления каждой новой статьи в журнале, которая попадает на личную страницу сотрудника, запускается следующий механизм. Система ИСТИНА связывается с системой Web of Science и по данным статьи запрашивает информацию о цитировании. В качестве параметров запроса выступает название статьи, индекс DOI, список авторов, том и номер журнала, номера страниц и т.д. Если эта статья была найдена в доступных базах цитирования системы Web of Science, то Web of Science выдает общее количество цитирований этой статьи. На основе этой первичной информации (число ссылок на каждую отдельную статью) в системе ИСТИНА рассчитываются остальные показатели: общее число цитирований статей сотрудника, индекс Хирша и т.д. Отметим, что из системы Web of Science в систему ИСТИНА попадают только статистические данные по статьям (число ссылок на каждую отдельную статью). Вся информация, касающаяся самих статей (название, список авторов, журнал и т.д.), должна быть введена пользователями в систему ИСТИНА самостоятельно. Следует заметить, что описанный механизм может также быть использован для верификации статей.
- Число цитирований публикаций по данным Web of Science в расчете на одного автора.
- Число сотрудников-авторов публикаций.
- Сумма коэффициентов вкладов авторов-сотрудников подразделения, где вклад одного автора рассчитывается как единица, деленная на число авторов работы.
- Взвешенная сумма коэффициентов вкладов авторов-сотрудников подразделения, которая отличается от

¹<http://researchanalytics.thomsonreuters.com/solutions/amr/>

предыдущей характеристики тем, что дополнительно учитывается позиция журнала, в котором опубликована работа, по импакт-фактору в своей тематической рубрике по данным Web of Science. Данная метрика позволяет учитывать все статьи в журналах из Journal Citation Reports, придавая большее значение публикациям в журналах с высоким импакт-фактором.

Комбинируя значения перечисленных параметров, руководитель организации или подразделения может получить большое число различных статистических данных по интересующим его подразделениям и тематикам. Результаты анализа представляются в виде списков и диаграмм. Отметим, что информация обновляется автоматически, без перезагрузки страницы. Например, для оценки работы различных подразделений с помощью описанного инструментария можно вывести следующие данные:

- отсортированный список структурных подразделений по количеству цитирований публикаций их сотрудников в отчетном периоде (рис. 2.18), при необходимости учитывая рейтинг соответствующего журнала по импакт-фактору;
- отсортированный список подразделений по общему количеству публикаций сотрудников в отчетном периоде;
- распределение тематик по суммарному числу цитирований публикаций (рис. 2.19), что позволяет выделить в организации области (направления) научного поиска, публикации по которым имеют суммарно наибольшую цитируемость.

Кроме общей оценки подразделений, с помощью представленного инструментария можно проанализировать их результативность в рамках заранее выбранной (интересующей) тематики. Например, можно вывести на экран список подразделений, сотрудники которых опубликовали наибольшее количество статей в журналах по физике за отчетный период (рис. 2.20). При этом отображается также разбиение этих статей по подразделам заданной тематики в рамках выбранного рубрикатора. Такие статистические отчеты позволяют со-

Информация о публикациях

Инструкция и правила расчета коэффициентов

Выберите параметры запроса

Тип публикации: Все журналы | Только проверенные | Метрика: Число ссылок WoS (на статью) | Год: Все

Обновить

Все подразд.

Name	Value
1 Химический факультет	27586
2 Физический факультет	19302
3 Научно-исследовательский институт ядерной физики имени Д.В. Скобельцына	12833
4 Научно-исследовательский институт физико-химической биологии им. А.Н. Белозерского	8866
5 Факультет наук о материалах	7581
6 Биологический факультет	4472
7 Научно-исследовательский институт механики	3761
8 Механико-математический факультет	2441
9 Государственный астрономический институт имени П.К. Штернберга	2349
10 Геологический факультет	1622

критерий: число цитирований статей

самые «успешные» подразделения

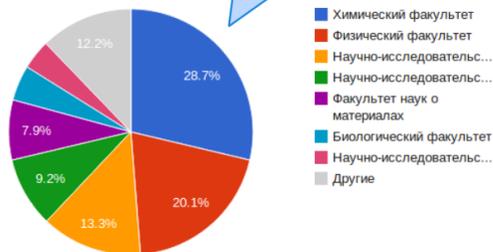


Рис. 2.18. Распределение подразделений МГУ по критерию числа цитирований статей под авторством сотрудников МГУ.

Информация о публикациях

Инструкция и правила расчета коэффициентов

Выберите параметры запроса

Тип публикации: Все журналы | Только проверенные | Метрика: Число ссылок WoS (на статью) | Год: Все

Обновить

Все подразд.

Name	Value
1 Химический факультет	27586
2 Физический факультет	19302
3 Научно-исследовательский институт ядерной физики имени Д.В. Скобельцына	12833
4 Научно-исследовательский институт физико-химической биологии им. А.Н. Белозерского	8866
5 Факультет наук о материалах	7581
6 Биологический факультет	4472
7 Научно-исследовательский институт механики	3761
8 Механико-математический факультет	2441
9 Государственный астрономический институт имени П.К. Штернберга	2349
10 Геологический факультет	1622

критерий: число цитирований статей

самые «успешные» подразделения

Все рубрики

Рубриктор: ГРНТИ | Порог: 1/100 | Фильтровать подразделения:

Метрика: Из запроса

самые развитые в организации области науки

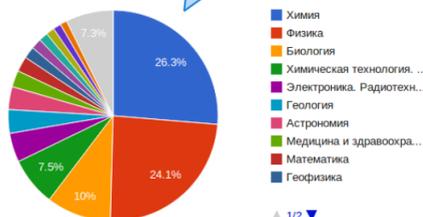


Рис. 2.19. Распределение областей знаний по критерию числа цитирований статей под авторством сотрудников МГУ.

ставить общее представление о развитии рассматриваемого тематического направления в организации в целом и в ее отдельных подразделениях, а также понять, какие из подразделов заданной области знания развиваются в организации наиболее активно, а какие - наоборот, находятся в стадии стагнации.

Информация о публикациях

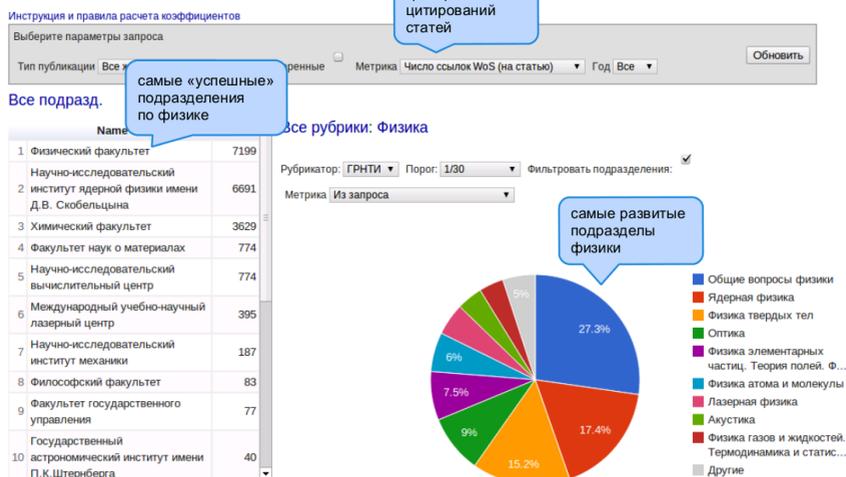


Рис. 2.20. Распределение подразделений МГУ и подразделов физики по критерию числа цитирований статей по физике под авторством сотрудников МГУ.

Отметим, что в настоящее время средства анализа в Системе во многом ориентированы на естественно-научные специальности. Такое положение обусловлено тем обстоятельством, что анализу подвергаются только публикации в журналах и сборниках трудов конференций. Однако для многих гуманитарных специальностей наиболее ценными показателями результативности деятельности сотрудников служат другие типы работ. Например, для филологов такими работами могут быть переводы книг, для историков - монографии, для журналистов - статьи в газетах с большим тиражом. Как следствие, одним из приоритетных направлений развития системы ИСТИНА является разработка средств, которые позволят различным подразделениям и организациям устанавливать свои критерии анализа, учиты-

вающие специфику их предметной области. Первым шагом в этом направлении является расширение списка типов результатов научной деятельности, которые можно добавлять в Систему. Отметим, что в текущей версии таких типов уже 20, что выгодно отличает ИСТИНУ от других распространенных систем, учитывающих главным образом публикации. Вторым шагом является расширение возможностей количественного и тематического анализа работ этих типов, в частности, с использованием онтологий для моделирования предметных областей. В настоящее время анализ деятельности подразделений выполняется только по публикациям их сотрудников и не учитывает, например, членство в редколлегиях журналов и диссертационных советах. Однако эти виды деятельности являются очень важными при оценке труда ученого. Отметим, что несмотря на этот факт, система ИСТИНА позволяет просмотреть страницу любого сотрудника и вручную оценить его деятельность и убедиться в его квалификации уже по всем 20 возможным типам работ на его странице, включая членство в редколлегиях журналов и диссертационных советах.

2.5.2. Статистика подразделения

Кроме статистики по организации и тематическому направлению, в системе ИСТИНА для анализа доступна статистическая информация по отдельному подразделению. Отметим, что можно вывести информацию как за отдельный календарный год, так и за все годы. В частности, на странице статистики подразделения отображаются:

- общее количество результатов деятельности, добавленных сотрудниками подразделения, распределенные по типам (рис. 2.21);
- общее количество результатов деятельности, добавленных сотрудниками организации, распределенные по типам;
- наукометрические показатели сотрудников подразделения, включающие индекс Хирша, общее число цитирований, число цитирований за 5 лет, число статей из Топ-25% журналов по

Механико-математический факультет, сводные данные подразделение

Отфильтровать по году: [все](#) | [2014](#) | [2013](#) | [2012](#) | [2011](#) | [2010](#) | [2009](#)

Число публикаций различных типов

№	Тип	Количество
1	Доклады на конференциях	1844
2	Монографии	1039
3	Патенты	169
4	Статьи в журналах (всего)	11060
5	Статьи в журналах из Top-25	600
6	Статьи в журналах из списка ВАК	4269
7	Статьи в журналах из списков SCOPUS, WoS	3437
8	Статьи в сборниках	3024
9	Стереотипные издания	38
10	Тезисы	2004
11	Учебные пособия	148

Рис. 2.21. Общее количество результатов деятельности, добавленных сотрудниками механико-математического факультета МГУ.

импакт-фактору Web of Science с возможностью сортировки по каждому показателю (рис. 2.22);

- наукометрические показатели подразделений за период (по умолчанию - 2006-2013 годы), что позволяет сравнить показатели различных подразделений организации;
- распределение количества цитирований статей сотрудников подразделения по году публикации статьи, что позволяет оценить наиболее продуктивные периоды деятельности подразделения (рис. 2.23);
- статистика по участию сотрудников подразделения в конференциях, включающая число докладов, число соавторов и число участников конференций (рис. 2.24);
- статистика по участию сотрудников подразделения в программных комитетах конференций;
- список всех публикаций сотрудников подразделения (рис. 2.25);
- статистика по участию сотрудников подразделения в редколлегиях журналов (рис. 2.26);

Поиск (по всем полям): WOS							
Сотрудник	Источник	Н-индекс	Всего статей	Число ссылок	Статей за 5 лет	Ссылок на статьи 5-ти лет	Число статей из top25
Соколов Дмитрий Дмитриевич	WOS	36	206	4111	61	381	96
Измоденов Владислав Валерьевич	WOS	28	99	2303	32	685	69
Баранов Владимир Борисович	WOS	15	41	1315	4	20	15
Богачев Владимир Игоревич	WOS	15	78	669	19	64	21
зайцев михаил владимирович	WOS	14	60	637	16	37	12
Ольшанский Максим Александрович	WOS	16	52	587	27	168	38
Ширяев Альберт Николаевич	WOS	12	25	425	6	13	8
Левин Владимир Алексеевич	WOS	7	100	398	24	13	3
Шкаликос Андрей Андреевич	WOS	9	47	382	10	35	1
Протасов Владимир Юрьевич	WOS	10	48	378	20	81	10
Мохов Олег Иванович	WOS	10	62	357	9	2	4
Kotesh Alexander	WOS	11	45	357	13	37	21
Гусейн-Заде Сабир Меджидович	WOS	10	69	330	16	10	17
Артемов Сергей Николаевич	WOS	8	48	323	11	7	5
Мальшев Вадим Александрович	WOS	9	47	310	4	8	7
Питербарг Владимир Ильич	WOS	9	46	287	12	27	9
Фоменко Анатолий Тимофеевич	WOS	9	41	270	3	3	2
Ильященко Юлий Сергеевич	WOS	7	28	261	10	14	14
Садовничий Виктор Антонович	WOS	9	137	255	30	48	3
Фалин Геннадий Иванович	WOS	9	57	241	3	10	5
Колпаков Роман Максимович	WOS	5	21	225	7	6	2
Дьячков Аркадий Георгиевич	WOS	7	32	211	5	7	6
Нестерено Юрий Валентинович	WOS	5	16	206	3	2	0
Неретин Юрий Александрович	WOS	7	49	186	10	7	10
Чечин Григорий Александрович	WOS	7	33	186	17	50	8

Рис. 2.22. Наукометрические показатели сотрудников механико-математического факультета МГУ.

МГУ имени М.В. Ломоносова, Механико-математический факультет

Показывать по 10 строк на странице		Поиск (по всем полям):	
Год публикации статьи	Число ссылок		
1996	1176		
2000	1019		
1995	963		
2004	962		
1999	918		
2005	916		
2002	915		
2009	878		
2006	870		
2003	828		

Показаны строки с 1 по 10, Всего: 51 [Предыдущая](#) [Следующая](#)

Рис. 2.23. Распределение числа цитирований статей сотрудников механико-математического факультета МГУ по году публикации статьи.

ГЛАВА 2. МОДУЛИ, РЕАЛИЗУЮЩИЕ ОСНОВНЫЕ ФУНКЦИИ

					Поиск (по всем полям): <input type="text"/>
Доклады	Соавторов	Сотрудников	Страна	Название	
17	19	7	Россия	Ломоносовские чтения	
7	8	6	Россия	ЛОМОНОСОВСКИЕ ЧТЕНИЯ	
2	6	3	Испания	5th International Workshop on Applied Probability (IWAP 2010) 5-8 July, Colmenarejo, Madrid, Spain	
2	5	3	Италия	III Symposium on semi-Markov models: Theory and Applications	
4	4	3	Россия	X международный семинар "Дискретная математика и ее приложения"	
3	4	3	Россия	Научная конференция "Ломоносовские чтения". Секция механики. МГУ имени М. В. Ломоносова.	
3	3	3	Украина	Modern Stochastics: Theory and Applications III	
3	3	3	Россия	Спектральная теория операторов и ее приложения	
2	3	3	Россия	XXII международная конференция им. И. Г. Петровского "Дифференциальные уравнения и смежные вопросы"	
2	3	3	Россия	Междунар. науч. симпоз. по проблемам механики деформируемых тел, посвященный 95-летию со дня рождения А.А.Ильюшина	
2	3	3	Россия	Международная конференция «Современные проблемы математики, механики и их приложений», посвященная 70-летию ректора МГУ академика В.А.Садовничева	
2	3	3	Россия	Международная научная конференция "Современные проблемы анализа и преподавания математики" посвященная 105-летию академика Сергея Михайловича Никольского	
2	3	3	Россия	Международный алгебраический симпозиум, посвященный 80-летию кафедры высшей алгебры механико-математического факультета МГУ и 70-летию профессора А.В. Михалева	

Рис. 2.24. Статистика по участию в конференциях сотрудников механико-математического факультета МГУ.

Тип	Авторы	Название	Источник	Год
Статьи в журналах	Карликов В.П.	Решение линеаризованной ассимметричной задачи о точечном взрыве в среде с переменной плотностью	Доклады Академии наук СССР 101	1955
Статьи в журналах	Карликов В.П.	О движении газа после придания ему начального вращения	Оборонгиз 19	1956
Статьи в сборниках	Черный Г.Г., Гонор А.Л., Иванова Е.Л.	Течения идеального газа при большой сверхзвуковой скорости	Технический отчет филиала ЦИАМ. Номер 2794	1956
Статьи в журналах	Черный Г.Г., Гонор А.Л.	О телах наименьшего сопротивления при больших сверхзвуковых скоростях	Изв. АН СССР ОТН 7	1957
Статьи в журналах	Karpelevich F.I., Tutubalin V.N., Shur M.G.	Limit theorems for the composition of distributions in the Lobachevsky plane and space	Theory of Probability and its Applications 4 4 10.1137/1104039	1959
Статьи в журналах	Долженко Е.П.	О граничных теоремах единственности и о поведении аналитических функций вблизи границы	Доклады Академии наук СССР 1 129	1959
Статьи в журналах	Долженко Е.П.	Построение на нигде не плотном континууме нигде не дифференцируемой функции, разлагающейся в ряд по рациональным функциям	Доклады Академии наук СССР 5 125	1959
Статьи в журналах	Карликов В.П.	Линеаризованная задача о распространении сильного взрыва в неоднородной атмосфере	Вестник Московского университета	1959
Статьи в журналах	Карпелевич Ф.И., Тутубалин В.Н., Шур М.Г.	Предельные теоремы для композиций распределений в плоскости и пространстве Лобачевского	Теория вероятностей и ее применения 4 4	1959
Статьи в журналах	Леонов В. П., Ширяев А. Н.	К технике вычисления семинвариантов	Теория вероятностей и ее применения 1 4 10.1137/1104031	1959

Показаны строки с 1 по 10. Всего: 11.793

[Предыдущая](#) [Следующая](#)

Рис. 2.25. Список всех публикаций сотрудников механико-математического факультета МГУ.

Журнал	Сотр. в редколлегии
Фундаментальная и прикладная математика	16
Интеллектуальные системы	12
Central European Journal of Mathematics	5
Вестник Московского университета. Серия 1. Математика и механика	5
Moscow Mathematical Journal	4
Известия Российской академии наук. Механика твердого тела	3
Труды московского математического общества	3
Фундаментальная и Прикладная Математика	3
Int. J. Acta Mechanica, Springer, Austria	2
Int. J. Archive of Applied Mechanics, Springer, Germany.	2
Moscow Journal of Combinatorics and Number Theory	2
The Scientific World Journal	2

Рис. 2.26. Статистика по участию сотрудников механико-математического факультета МГУ в редколлегиях журналов.

- статистика по участию сотрудников подразделения в диссертационных советах (рис. 2.27);

Название	Сотрудников
Д 501.002.16	12
Д 501.001.89	9
Д 501.001.84	8
Д 501.001.91	7
Д 501.001.85 МГУ им. М.В.Ломоносова	6
Д 002.022.03 Математический институт им. В.А. Стеклова РАН	3
Д 002.059.01 Институт машиноведения им. А.А. Благонравова РАН, Москва	2
Д 212.278.02 Ульяновский государственный университет	2
Д 501.002.07 МГУ им. М.В. Ломоносова	2
Экспертный совет по математике и механике ВАК РФ Министерство образования и науки РФ	2

Рис. 2.27. Статистика по участию сотрудников механико-математического факультета МГУ в диссертационных советах.

- список журналов, в которых наиболее часто публикуются сотрудники подразделения (рис. 2.28);
- список сотрудников подразделения, имеющих публикации в журналах из Топ-25% журналов Web of Science по импакт-фактору с указанием коэффициента их участия (рис. 2.29);
- распределение числа зарегистрированных сотрудников подразделения по кафедрам и лабораториям (рис. 2.30);
- список сотрудников подразделения, отсортированный по коли-

Список журналов, в которых часто публикуются сотрудники

№	Журнал	Количество статей	Количество авторов
1	Вестник Московского университета. Серия 1. Математика и механика	577	161
2	Успехи математических наук	411	106
3	Математические заметки	335	107
4	Доклады Российской Академии наук	266	75
5	Дифференциальные уравнения	242	33
6	Математический сборник	235	79
7	Russian Mathematical Surveys	202	64
8	Фундаментальная и прикладная математика	183	83
9	Mathematical Notes	177	75
10	Теория вероятностей и ее применения	165	33
11	Journal of Mathematical Sciences	133	79
12	Интеллектуальные системы	125	60
13	Doklady Mathematics	120	46
14	Функциональный анализ и его приложения	120	31
15	Известия Российской академии наук. Механика жидкости и газа	115	26
16	Differential Equations	109	29
17	Russian Journal of Mathematical Physics	103	31
18	Theory of Probability and Its Applications	93	20
19	Moscow University Mathematics Bulletin	93	40
20	Дискретная математика	88	38
21	Fluid Dynamics	86	16
22	Труды Математического института им.В.А.Стеклова РАН	86	45
23	Sbornik Mathematics	86	41
24	Известия Российской академии наук. Механика твердого тела	84	28
25	Прикладная математика и механика	76	23

Рис. 2.28. Список журналов, в которых наиболее часто публикуются сотрудники механико-математического факультета МГУ.

Сотрудники, имеющие публикации в журналах из Топ-25%

№	Сотрудник	Коэффициент	Подразделение
1	Kotech Alexander	9,74	Кафедра теории вероятностей
2	Агеев Олег	4	Механико-математический факультет
3	Александров Владимир Васильевич	0,66	Кафедра прикладной механики и управления
4	Алексеев Дмитрий Владимирович	0,12	Кафедра математической теории интеллектуальных систем
5	Алимов Алексей Ростиславович	1,5	Кафедра теоретической информатики
6	Аржанцев Иван Владимирович	1,03	Кафедра высшей алгебры
7	Артемов Сергей Николаевич	2,53	Кафедра математической логики и теории алгоритмов
8	Арушанян Игорь Олегович	0,5	Кафедра вычислительной математики
9	Афанасьев Валерий Иванович	0,25	Кафедра математической статистики и случайных процессов
10	Баранов Владимир Борисович	7,16	Кафедра аэромеханики и газовой динамики
11	Белошапка Валерий Константинович	0,5	Кафедра теории функций и функционального анализа
12	Богаевский Илья Александрович	1	Кафедра теории динамических систем
13	Богатый Семеон Антонович	1,83	Кафедра общей топологии и геометрии
14	Богачев Владимир Игоревич	8,41	Кафедра теории функций и функционального анализа
15	Булинская Екатерина Вадимовна	9	Кафедра теории вероятностей
16	Булинский Александр Вадимович	1	Кафедра теории вероятностей
17	Бухштабер Виктор Матвеевич	4,5	Кафедра высшей геометрии и топологии
18	Васильева Анастасия Андреевна	1	Кафедра общих проблем управления
19	Ватугин Владимир Алексеевич	0,25	Кафедра математической статистики и случайных процессов
20	Верещагин Николай Константинович	2,5	Кафедра математической логики и теории алгоритмов
21	Веселов Александр Петрович	3,16	Механико-математический факультет
22	Галатенко Владимир Владимирович	0,13	Кафедра математического анализа
23	Георгиевский Дмитрий Владимирович	1	Кафедра механики композитов
24	Голо Войслав Любомирович	1,61	Кафедра дифференциальной геометрии и приложений
25	Голован Андрей Андреевич	0,5	Кафедра прикладной механики и управления

Рис. 2.29. Список сотрудников механико-математического факультета МГУ, имеющих публикации в журналах из Топ-25% журналов Web of Science по импакт-фактору.

Число зарегистрированных сотрудников

№	Подразделение	Количество зарегистрированных сотрудников
1	Химический факультет	66
2	Кафедра аналитической химии	49
3	Кафедра английского языка химического факультета	0
4	Кафедра высокомолекулярных соединений	54
5	Кафедра коллоидной химии	27
6	Кафедра лазерной химии	22
7	Кафедра неорганической химии	103
8	Кафедра общей химии	55
9	Кафедра органической химии	152
10	Кафедра радиохимии	51
11	Кафедра физической химии	231
12	Кафедра химии нефти и органического катализа	43
13	Кафедра химии природных соединений	43
14	Кафедра химической кинетики	49
15	Кафедра химической технологии и новых материалов	80
16	Кафедра химической энзимологии	75
17	Кафедра электрохимии	47
18	Лаборатория вычислительных методов в химии	2
19	Подготовительное отделение химического факультета	0
	Всего	1140

Рис. 2.30. Распределение числа зарегистрированных сотрудников химического факультета МГУ по кафедрам и лабораториям.

честву статей и тезисов докладов, который позволяет получить представление о наиболее продуктивно работающих сотрудниках подразделения (рис. 2.31);

МГУ имени М.В. Ломоносова, Механико-математический факультет

Список сотрудников

Страницы: << предыдущая 1 2 3 4 5 6 следующая >>



Садовничий Виктор Антонович

SVA

МГУ имени М.В. Ломоносова, Механико-математический факультет, Отделение математики, Кафедра математического анализа, с 1 сентября 1967
Соваторы: Дубровский В.В., Александров В.В., Окунев Ю.М., Подольский В.Е., Винокуров В.А., Архипов Г.И., Самсонов В.А., Ильин В.А. и др.
654 статья, 150 книг, 26 докладов на конференциях, 81 тезисы докладов, 10 патентов, 4 свидетельства о регистрации прав на ПО, 2 отчета, 9 членств в редколлегиях журналов, 84 членства в редколлегиях сборников, 5 членств в программных комитетах, 3 членства в диссертационных советах, 36 диссертаций, 4 учебных курса



Левин Владимир Алексеевич

Levin

МГУ имени М.В. Ломоносова, Научно-исследовательский институт механики, 112 Лаборатория газодинамики взрыва и реагирующих систем, с 14 июля 2009
Соваторы: Марков В.В., Георгиевский П.Ю., Журавская Т.А., Мануйлович И.С., Osinkin S.F., Хмельский А.Н., Громов В.Г., Starik A.M. и др.
295 статей, 2 доклада на конференциях, 62 тезисов докладов, 3 патента, 12 отчетов, 2 членства в редколлегиях журналов, 1 членство в диссертационном совете, 5 диссертаций, 4 дипломные работы



Шамолин Максим Владимирович

shamolin

МГУ имени М.В. Ломоносова, Научно-исследовательский институт механики, 302 Лаборатория навигации и управления, с 15 октября 1992
Соваторы: Георгиевский Д.В., Самсонов В.А., Айдаулов Р.Р., Селиванова Н.Ю., Борисенко И.Т., Шебаршов Д.В., Окунев Ю.М., Трофимов В.В. и др.
159 статей, 7 книг, 5 докладов на конференциях, 193 тезисов докладов, 9 отчетов, 4 членства в редколлегиях журналов, 1 членство в программном комитете, 4 диссертации



Соколов Дмитрий Дмитриевич

sokoloff

МГУ имени М.В. Ломоносова, Физический факультет, Отделение экспериментальной и теоретической физики, Кафедра математики, 1 апреля 1972 - 1 января 2015
Соваторы: Puzmalu ikin A.A., Shukurov A., Moss D., Frick P., Kuzanyan K., Beck P., Kleorin N., Zhang Y.H. и др.
324 статья, 6 книг, 21 доклад на конференциях, 6 тезисов докладов, 1 награда, 5 членств в редколлегиях журналов, 3 членства в программных комитетах, 3 членства в диссертационных советах, 3 диссертации, 4 учебных курса



Сабитов Иджд Хакевич

ISabitov

МГУ имени М.В. Ломоносова, Механико-математический факультет, Отделение математики, Кафедра математического анализа, с 15 октября 1964
Соваторы: Максимов И.Г., Александров А.Д., Астрелин А.В., Иванова-Каратопраклиева И., Перлова Н.Г., Ivanova-Karatopraklieva I., Александров В.А., Аминов Ю.А. и др.
241 статья, 10 книг, 76 тезисов докладов, 4 награды, 3 членства в научных обществах, 4 членства в редколлегиях журналов, 25 членств в программных комитетах, 4 диссертации, 12 дипломных работ, 3 учебных курса



Фоменко Анатолий Тимофеевич

ATFomenko

МГУ имени М.В. Ломоносова, Механико-математический факультет, Отделение математики, Кафедра дифференциальной геометрии и приложений, с 1 октября 1967
Соваторы: Болснисов А.В., Мищенко А.С., Носовский Г.В., Трофимов В.В., Матвеев С.В., Новиков С.П., Тружинин А.А., Калашников В.В. и др.
273 статьи, 86 книг, 12 докладов на конференциях, 26 тезисов докладов, 4 членства в редколлегиях журналов, 1 членство в программном комитете, 2 членства в диссертационных советах, 52 диссертации, 30 дипломных работ, 5 учебных курсов



Ковалев Валерий Леонидович

Kovalev

МГУ имени М.В. Ломоносова, Механико-математический факультет, Отделение механики, Кафедра газовой и волновой динамики, с 1 сентября 1987
Соваторы: Крупнов А.А., Ячунчиков А.Н., Пососбеян М.Ю., Афонина Н.Е., Громов В.Г., Суслев О.Н., Буцаенко А.А., Тирский Г.А. и др.
158 статей, 4 книги, 7 докладов на конференциях, 80 тезисов докладов, 14 отчетов, 2 членства в редколлегиях журналов, 2 членства в диссертационных

Рис. 2.31. Список сотрудников механико-математического факультета МГУ.

- список добавленных в Систему статей сотрудниками подразделения, который позволяет получить представление об их научной работе в последнее время или о деятельности, направленной

ной на более полное отражение своих результатов в Системе (рис. 2.32);

Механико-математический факультет

[подразделение](#)

Работы, добавленные за последние 2 недели [версия для печати](#)

Страницы: << предыдущая 1 2 3 4 5 6 следующая >>

- сегодня 2014 Cohomology of Landweber-Novikov algebra and singular Virasoro vectors Миллиончиков Д.В.
- сегодня 2009 Jacobi's bound for systems of algebraic differential equations Kondratieva M.V., Mikhalev A.V., Pankrat'ev E.V.
- сегодня 2011 Isomorphisms of general linear groups over associative rings graded by an abelian group Atkarskaya A.S., Bunina E.I., Mikhalev A.V.
- сегодня 2012 The topological Jacobson radical of rings, I Glavatskii S.T., Mikhalev A.V., Tenzina V.V.
- сегодня 2012 The topological Jacobson radical of rings, II Glavatskii S.T., Mikhalev A.V., Tenzina V.V.
- сегодня 2012 Quotient rings of graded associative rings, I Balaba I.N., Kanunnikov A.L., Mikhalev A.V.
- сегодня 2014 Эндоморфизмы полугрупп неотрицательных обратимых матриц Бунина Е.И., Михалёв А.В.
- сегодня 2013 Оценки высоты в смысле Ширшова количества фрагментов малого периода Белов А.Я., Михалёв А.В.
- сегодня 2013 Подъем полиномиальных автоморфизмов и размерность Гельфанда-Кириллова Белов А.Я., Михалёв А.В.
- сегодня 2013 Первичный радикал упорядоченных лул Михалёв А.В.
- сегодня 2012 Параметр кодов, построенных на основе кольца матриц с помощью функции следа Марков В.Т., Михалёв А.В., Нецаев Е.А.
- сегодня 2010 Периодичность морфических слов Белов А.Я., Михалёв А.В.
- сегодня 2011 Числа Фробениуса полугрупп Михалёв А.В.
- сегодня 2011 Автоморфизмы групп Шевалле типа G_2 над локальными кольцами без $1/2$ Бунина Е.И., Михалёв А.В.
- сегодня 2011 Стандартные и нестандартные автоморфизмы групп Шевалле над кольцами Бунина Е.И., Михалёв А.В.
- сегодня 2010 Градуированные кольца частных и ортогональное пополнение Михалёв А.В.
- сегодня 2014 Изоморфизмы линейных групп над ассоциативными кольцами Аткарская А.С., Михалев А.В., Бунина Е.И.
- сегодня 2014 Критерий элементарной эквивалентности групп автоморфизмов и колец эндоморфизмов абелевых r -групп Бунина Е.И., Михалев А.В., Ройзнер М.А.
- сегодня 2014 Киселев А.Б. в программном комитете конференции Проблемы газовой и волновой динамики и ракетной техники (2014) Киселев А.Б.
- вчера 2014 Проблема Рисса-Радона-Фреше характеристики радоновских интегралов как линейных функционалов Захаров В.К.
- вчера 2014 Проблема Рисса-Радона-Фреше характеристики радоновских интегралов как линейных функционалов Родионов Т.В.
- вчера 2014 Характеристики поворачиваемости решений дифференциальных систем Сергеев И.Н.

Рис. 2.32. Список последних добавленных в Систему статей сотрудниками подразделения механико-математического факультета МГУ.

- список всех статей сотрудников подразделения в формате XML;
- список всех научных коллективов в организации;
- список всех единиц уникального, научного оборудования в организации, которое, в том числе, может использоваться научными коллективами разных подразделений;
- список всех грантов, полученных разными научными коллективами.

Отметим, что подобная статистика доступна по каждому подразделению организации любого уровня (факультету, кафедре или лаборатории) с соответствующим разграничением доступа.

Механико-математический факультет подразделение

Список публикаций 2008-2012 годов в журналах, индексируемых в международных системах цитирования (Приложение 5)

Показывать по записей на странице

Поиск (по всем полям):

№ пп	Полное библиографическое описание статьи	Библиографическая база, в которой индексируется журнал	Входит в Перечень ведущих рецензируемых научных журналов и изданий, рекомендованный ВАК	Импакт-фактор журнала	Количество цитирований	Электронный адрес размещения	Цифровой идентификатор объекта DOI
281	Global observations of the interstellar interaction from the interstellar boundary explorer (ibex) / D. McComas, F. Allegrini, P. Bochsler et al. // Science. — 2009. — Vol. 326, no. 5955. — P. 959.	Web of Science	да	31,377	141	http://www.scopus.com/record/display?url?origin=recordpage&eid=2-s2.0-72449184780	10.1126/science.1180906
243	Comparison of interstellar boundary explorer observations with 3d global heliospheric models / N. Schwadron, C. Prested, M. Bzowski et al. // Science. — 2009. — Vol. 326, no. 5955. — P. 966–968.	Web of Science	да	31,377	82	http://www.scopus.com/record/display?url?origin=recordpage&eid=2-s2.0-72449165025	10.1126/science.1180986
	Evolution of magnetic fields in galaxies and						

Рис. 2.33. Список публикаций за 2008–2012 гг. сотрудников механико-математического факультета МГУ в журналах, индексируемых в международных системах цитирования (фрагмент).

В научно-образовательной сфере часто возникает необходимость формировать отчеты организации в различные органы управления наукой и образованием. Одним из последних примеров такого отчета стали сведения, востребованные в мае 2013 г. Минобрнауки РФ в связи с мониторингом деятельности сети диссертационных советов. Разработчиками Системы были оперативно написаны функции формирования необходимых отчетов в требуемом Министерством виде как по отдельным сотрудникам, так и по подразделениям, с возможностью экспорта данных в формат Excel. Пример такого отчета представлен на рис. 2.33. Список всех отчетов по подразделению для мониторинга сети диссертационных советов приведен на рис. 2.34.



ИСТИНА

Интеллектуальная Система Тематического Исследования Научно-технической информации

[Главная](#) [Моя страница](#) [Добавить работу](#) [Поиск](#) [Статистика](#) [О проекте](#) [Помощь](#)

Голомазов Денис Дмитриевич (goldan)
[Выйти из системы](#)

МГУ имени М.В. Ломоносова

Механико-математический факультет подразделение

Отчет "Об организации мониторинга деятельности сети диссертационных советов"

1. Сведения о научных и научно-педагогических работниках (Приложение 1)
[Просмотреть на сайте](#) [Скачать Excel-файл](#)
2. Монографии, изданные за 2008-2012 годы (Приложение 4)
[Просмотреть на сайте](#) [Скачать Excel-файл](#)
3. Список публикаций 2008-2012 годов в журналах, индексируемых в международных системах цитирования (Приложение 5)
[Просмотреть на сайте](#) [Скачать Excel-файл](#)
4. Список публикаций 2008-2012 годов в научных журналах, входящих в Перечень ВАК (Приложение 6)
[Просмотреть на сайте](#) [Скачать Excel-файл](#)
5. Патенты на изобретения, полученные в 2003-2012 годах (Приложение 7)
[Просмотреть на сайте](#) [Скачать Excel-файл](#)

Примечания

1. Вспомогательные материалы, которые представлены в системе ИСТИНА, рассчитываются только по сотрудникам, зарегистрированным в системе и явно указавшим свое место работы в МГУ.
2. В отчеты попадает информация о результатах деятельности, в авторах которых есть хотя бы один пользователь, указавший место работы в подразделении или его дочернем подразделении.
3. Результаты деятельности (публикации и т.д.) в отчетах не разбиты по группам специальностей, так как в системе в настоящее время нет этой информации.
4. Библиографические описания публикаций отображаются по ГОСТ 7.1.
5. В качестве электронного адреса размещения указывается адрес публикации в Scopus (если статья была найдена в Scopus), иначе указывается адрес ее страницы в системе ИСТИНА.
6. Импакт-факторы по Scopus обновлены до последней версии 2012 года.
7. Импакт-факторы по РИНЦ взяты со страницы [eLibrary.ru](#), показатель "пятилетний импакт-фактор с учетом переводной версии".
8. Пустые ячейки в отчетах означают, что соответствующих данных в настоящее время нет в системе ИСТИНА.

Рис. 2.34. Список отчетов по мониторингу сети диссертационных советов.

С тем, чтобы для сбора необходимой по подобным запросам информации в будущем не приходилось каждый раз обращаться к разработчикам Системы, возникла идея создать инструментальное средство автоматического («по кнопке») формирования отчетов различных (заданных оператором Системы) видов. В настоящее время оно находится в разработке. Предполагается, что сотрудник организации, имеющий соответствующие права доступа, сможет выполнять произвольные SQL-запросы к данным Системы. Для того, чтобы данные по ошибке не были повреждены, а также для снижения нагрузки на базу данных, эти запросы, во-первых, включают только запросы на чтение, а во-вторых, выполняются к

копии рабочей базы данных.

2.6. Методы и средства повышения качества данных

Система ИСТИНА предназначена для сбора, обработки, хранения, анализа и выдачи по запросу информации о результатах научной и педагогической деятельности сотрудников научных и образовательных учреждений. Важной характеристикой такой системы является качество данных, под которым понимается их точность и полнота. Точность характеризуется такими показателями, как корректность информации и количество дублирующих данных. Под полнотой будем понимать:

- количество ученых, результаты (показатели) деятельности которых введены в систему;
- количество типов результатов деятельности, обрабатываемых системой (например, публикации, патенты, отчеты, доклады на конференциях);
- количество результатов деятельности по каждому из типов для отдельного ученого.

Настоящий раздел посвящен методам и средствам повышения качества информации, которые разрабатываются и используются в системе ИСТИНА. Для повышения точности информации в Системе необходимо рассмотреть следующие вопросы.

- Классификация некорректных и дублирующих данных по категориям для облегчения последующего их анализа.
- Идентификация нежелательных ситуаций, которые приводят к появлению таких данных в Системе.
Таким образом предлагается сосредоточить усилия не только на исправлении ошибок, но и на причинах их возникновения.

Результаты анализа информации из хранилища экземпляра Системы, который используется в Московском университете, показывают, что во многих случаях некорректные или дублирующие данные возникают из-за недостатков интерфейса взаимодействия пользователя с Системой.

- Механизмы, которые предотвращают возникновение нежелательных ситуаций.
- Средства исправления некорректных и дублирующих данных.
- Формирование политики разграничения доступа к функциям Системы, особенно к тем, которые позволяют модифицировать существующую информацию. Очевидно, например, что нельзя давать право редактировать и удалять любую статью любому пользователю.
- Механизмы запуска операций, повышающих качество данных, то есть определение, куда и каким образом поступает информация о некорректных или дублирующих сведениях, и кто осуществляет их проверку и исправление.

Основные идеи, связанные с механизмом запуска таких операций состоят в следующем. Во-первых, предлагается предоставить максимально возможный спектр действий, направленных на исправление данных, всем пользователям, не подвергая при этом информацию других пользователей риску быть испорченной. Во-вторых, следует создать особую группу пользователей («ответственные сотрудники»), которые, будучи прикрепленными к заданной вершине иерархии подразделений организации, получают повышенные привилегии по редактированию данных сотрудников, принадлежащих подразделению из поддерева этой вершины.

Некорректные и дублирующие данные в Системе делятся на следующие категории.

- Некорректные данные о результатах деятельности (напр., статьях):

- неправильное название, том, номер и т.п.;
- неправильная привязка авторов;
- Дубликаты:
 - пользователей;
 - соавторов;
 - результатов деятельности (статей, докладов,...);
 - общих объектов (журналов, конференций).

Среди ситуаций, вследствие которых уменьшается точность информации в Системе, выделяются следующие:

- ошибки в указанных пользователем сведениях о результате научной деятельности при добавлении;
- неправильное соотнесение (привязка) автора работы к сотруднику в базе данных Системы вследствие того, что нужный сотрудник не был найден в базе данных Системы;
- неправильное соотнесение автора работы к сотруднику вследствие того, что пользователем был сделан неправильный выбор нужного сотрудника;
- возникновение в хранилище дубликатов объектов различных типов.

В Системе используются следующие механизмы предотвращения отмеченных выше нежелательных ситуаций:

- уведомление о существующем похожем результате деятельности в Системе при добавлении;
- поиск похожих общих объектов;
- поиск похожих сотрудников.

Отметим, что поиск похожих объектов и сотрудников является важной задачей на широком научном направлении «разрешение неоднозначностей именных групп» (Named Entity Disambiguation).

Для исправления возникших некорректных и дублирующих данных в системе ИСТИНА существуют перечисленные далее механизмы их исправления.

- Модификация общих сведений о результате деятельности.
- Модификация сопоставления («привязки») авторов результатов деятельности с сотрудниками в базе данных Системы.
- Слияние дубликатов следующих объектов.

- Пользователей.

- Сотрудников.

Если пользователь попадает на страницу сотрудника, не связанного с другим пользователем, и Система считает, что данные пользователя и сотрудника (в частности, фамилия, имя и отчество) похожи, то она предлагает пользователю «привязать» этого сотрудника к своему профилю. Если пользователь соглашается, он нажимает кнопку, и все данные сотрудника добавляются к профилю пользователя. Поискать сотрудников, которые могут быть его дубликатами, пользователь может на специальной странице «возможные дубликаты».

- Результатов деятельности.

- Общих объектов.

В частности, в Системе разработан отдельный интерфейс слияния дубликатов журналов (рис. 2.35). Слияние могут осуществлять ответственные по подразделениям, которые лучше всего осведомлены о существующих журналах в своей предметной области. Кроме слияния, в Системе реализована функция редактирования администратором альтернативных имен журнала для уменьшения будущего дублирования. Ответственный сотрудник также может просмотреть список последних добавленных в Систему журналов для облегчения первоначального поиска кандидатов на слияние.

- Функция привязки статьи, книги и тезисов докладов.

2.6. МЕТОДЫ ПОВЫШЕНИЯ КАЧЕСТВА ДАННЫХ

В результатах поиска показывается не более 300 журналов.

Для сортировки по одному полю нажмите на название поля один или два раза (сортировка по возрастанию или убыванию). Для сортировки по нескольким полям нажмите на название поля при нажатой клавише SHIFT. Сортировка производится в порядке выбора полей (сначала по первому выбранному полю, потом - по второму). Чтобы добавить журнал в список сливаемых, нажмите на соответствующую ему строку. Чтобы удалить журнал из списка сливаемых, нажмите на соответствующую ему строку в списке сливаемых.

Отобрано на слияние: 2

Показывать по строк на странице

Поиск (по названию):

Искать в найденном

ID	Название	Издательство	Кол-во статей	ISSN	Индексы	Последним редактировал	Статус
1052607	Аналитический вестник Совета Федерации Федерального собрания Российской Федерации		5	✗	✗	kulkov	✗
1060621	Вестник UNESCO, (Commission of the Russian Federation for UNESCO)		0	✗	✗	LeonovaAB	✗
1074962	«Экономика, статистика и информатика», МЭСИ, научно-практический журнал, Вестник УМО		1	✗	✗	chaya	✗
1083874	Вестник КРАУНЦ		1	✗	✗	Lubnina_Natalia	✗
1085294	Альма матер. Вестник высшей школы		4	✗	✗	Asmolev	✗
1089193	Вестник Института Геологии (РАН, Уральское отд. Коми научный центр)		1	✗	✗	neremin	✗
1112642	Вестник Академии прогнозирования		0	✗	✗	Rachkovskaya	✗
1117070	Вестник КРАУНЦ. Гуманитарные науки		1	✗	✗	D.A.Leontiev	✗
1120174	Вестник Воронежского университета. Серия: Философия		2	✗	✗	a-lexx	✗
1154057	Вестник ВНЦ РАН и РСО-А		1	✗	✗	Rogozhin_EA	✗

Показаны строки с 1 по 10. Всего: 300

[Предыдущая](#) [Следующая](#)

Список сливаемых журналов.

ID	Название	Издательство	Кол-во статей	ISSN	Индексы	Последним редактировал	Статус
1083874	Вестник КРАУНЦ		1	✗	✗		✗ Убрать
1117070	Вестник КРАУНЦ. Гуманитарные науки		1	✗	✗		✗ Убрать

Показаны строки с 1 по 2. Всего: 2

Рис. 2.35. Интерфейс слияния журналов в системе ИСТИНА.

Если пользователь попадает на страницу статьи, и Система считает, что его данные (в частности, фамилия, имя и отчество) похожи на данные одного из авторов статьи, то Система предлагает пользователю «привязать» статью к своему профилю. Это предложение делается только в том случае, если похожий автор еще не привязан к другому пользователю (так называемый «свободный» автор). Если пользователь соглашается, он нажимает кнопку, и статья появляется в его профиле. Таким образом, он может легко (не редактируя статью) находить публикации, которые по ошибке не попали в его профиль, и привязывать их к себе.

- Поиск «потерянных» статей.

У каждого пользователя существует отдельная страница, на которой перечислены все публикации, которые он может привязать к себе. У каждой из таких публикаций есть похожий на пользователя свободный автор. Пользователь может перейти на страницу любой статьи и привязать ее к себе.

- Функция отказа от статьи, книги и тезисов докладов.

Зайдя на страницу статьи, книги или тезисов докладов, прикрепленных к его профилю, пользователь может отказаться от авторства данной работы, если он считает, что она попала к нему по ошибке. Работа открепляется от его профиля, но при этом не удаляется из Системы. Другой подходящий пользователь может прикрепить эту статью к своему профилю.

- Страница с последними добавленными пользователем статьями (и другими результатами деятельности), на которой он может найти статьи, по причине неправильной привязки не попавшие на его личную страницу, и исправить ошибку.

Важной задачей для обеспечения точности и полноты данных в системе ИСТИНА является разработка и внедрение механизма разграничения доступа пользователей к таким функциям Системы, как изменение данных и выполнение административных действий. Внедрение корректным образом настроенного механизма разграничения доступа позволяет ограничить неконтролируемое

изменение данных в Системе и уменьшить объем работ по проверке актуальности внесенных в нее данных.

Механизм разграничения доступа принимает решение запретить или разрешить определенное действие в Системе на основе задаваемого набора правил разграничения доступа. Корректное выполнение таких правил заложено в программную реализацию механизма разграничения доступа, а сами правила настраиваются при введении Системы в строй и в процессе ее дальнейшей эксплуатации. В основу формирования правил разграничения доступа в системе ИСТИНА положены следующие общие принципы:

- пользователям предоставляется широкий набор функций Системы по работе с их собственными данными в целях эффективного использования заложенного в Систему потенциала распределенного редактирования данных, принимая во внимание компетентность пользователей в отношении принадлежащих им результатов научной деятельности;
- ответственные по подразделениям, назначаемые в процессе эксплуатации Системы, обладают повышенными привилегиями в рамках соответствующего подразделения для поддержки актуальности сведений о результатах научной деятельности и других данных сотрудников подразделения;
- в целях обеспечения целостности данных ограничиваются действия, которые могут привести к преднамеренной или случайной порче данных пользователями Системы.

Реализация надежного механизма разграничения доступа, который соответствует актуальным потребностям Системы, должна выполняться на основе специально разрабатываемой формальной модели. Необходимо также учитывать, что реализация рассматриваемого механизма должна быть достаточно производительной и масштабируемой с учетом большого количества пользователей и объектов данных в Системе. Указанный аспект актуален в силу того обстоятельства, что проверка возможности доступа должна выполняться при всех действиях пользователей в рамках Системы. Кроме того,

механизм разграничения доступа целесообразно использовать при организации ограничивающего интерфейса: например, пользователю могут предлагаться только допустимые для него действия. Такой подход позволяет дополнительно упорядочить взаимодействие с Системой и сократить количество случайных и преднамеренных действий по нежелательному изменению данных в ней. Отмеченным аспектам формальной модели, правил и программной реализации механизма разграничения доступа посвящена глава 5.

2.7. Взаимодействие с другими системами

Одной из характеристик, определяющих высокий уровень эффективности современной информационной системы, является возможность интеграции размещенных в ней данных с информацией из других, сторонних по отношению к ней, систем. В частности, публикации многих ученых уже проиндексированы такими системами, как Web of Science, Scopus, Google Scholar, eLibrary.ru и другими. Естественным образом возникают вопросы импорта данных этих хранилищ в систему ИСТИНА, экспорта данных из системы ИСТИНА и сопоставления информационных объектов в системе ИСТИНА с соответствующими объектами в других системах.

2.7.1. Возможности импорта и экспорта данных

В системе ИСТИНА поддерживается импорт и экспорт данных о публикациях в формате BibTeX, который поддерживается многими другими программными комплексами. Это позволяет экспортировать и импортировать данные о публикациях в системе ИСТИНА без необходимости разработки для этих целей дополнительных средств.

В настоящее время кроме BibTeX, в систему ИСТИНА можно в автоматизированном режиме добавлять данные из следующих хранилищ:

- Web of Knowledge / Web of Science;
- Scopus;
- Google Scholar;
- ResearcherID.com;
- PubMed;
- MathSciNet;
- Zentralblatt Math;
- MathNet.ru;
- веб-сайты издательств, например, Elsevier, Springer и другие.

На основе методов импорта данных в формате BibTeX и алгоритмов автоматического разрешения неоднозначности авторства публикаций (раздел 3.2) разработаны средства автоматического наполнения базы данных из внешних источников, поддерживающих программный интерфейс поиска новых публикаций сотрудников выбранной организации. К числу таких систем относятся информационные системы Scopus и Springer.

В качестве отдельного сервиса в Системе также реализовано преобразование библиографической ссылки в BibTeX-запись (без записи в базу данных). Сотрудник также может экспортировать данные о всех своих публикациях со своей страницы в формат BibTeX, или отобразить эти данные в различных стилях.

2.7.2. Сопоставление информационных объектов

В настоящее время в мире существует целый ряд информационных систем, определяющих различные наукометрические характеристики публикаций. В первую очередь это показатели цитирования статей в периодических изданиях и сборниках. Наиболее известными примерами таких систем являются Web of Science и Scopus. Для публикаций на русском языке такой системой является eLibrary. В отдельных областях науки существуют собственные системы расчета

показателей цитирования, например, MathSciNet¹ в области математики или электронная библиотека АСМ² в области компьютерных наук. Такие системы проводят индексирование основных научных изданий в своей области и определяют число ссылок на публикации. Специализированные системы, как правило, имеют более широкое покрытие источников.

Как отмечено в 2.5.1, наукометрические показатели отдельных статей являются важной частью статистического анализа публикационной активности отдельных сотрудников организации или её подразделений. Разработка собственного индекса цитирования является сложной задачей, поскольку для этого требуется обработка данных журналов и конференций всего мира. В связи с этим возникает необходимость *автоматического сопоставления* информационных объектов, представленных в системе ИСТИНА, с записями в других системах. Фактически такое сопоставление сводится к установлению соответствий между идентификаторами статей в двух системах.

В системе ИСТИНА реализованы методы автоматического сопоставления статей с системам Web of Science и Scopus. Для этого используются программные интерфейсы данных систем. В частности, интеграция с Web of Science осуществляется с использованием инструментального средства Article Match Retrieval (AMR). После установления соответствия в систему ИСТИНА с заданной периодичностью поступают данные о цитировании отдельных публикаций, на основе которых рассчитываются различные наукометрические показатели публикационной активности ученых и подразделений, например, индекс Хирша или общее количество цитирований. Этот же инструментарий может использоваться для проверки корректности ввода данных о статьях по DOI следующим образом. Если в статье указан DOI, то производится запрос в AMR, который возвращает все сведения о статье (авторы, название, журнал и другие). Эти сведения затем сравниваются с теми, которые ввел пользователь, и, при наличии расхождений, ему может посылаться уведомление о возможной ошибке.

¹<http://http://www.ams.org/mathscinet/>

²<http://dl.acm.org/>

В Системе присутствует ряд функций, облегчающих ее интеграцию с другими информационными системами в организации. Например, списки сотрудников организации или подразделения могут быть экспортированы в формат XML.

В качестве дальнейших шагов развития Системы на направлении ее интеграции с другими системами на ближайшее будущее рассматриваются следующие задачи:

- настройка корректной индексации данных из ИСТИНЫ в Google Scholar, EBSCO и других аналогичных системах;
- разработка публичного программного интерфейса (API) для обеспечения возможности создания программных средств работы с Системой сторонними разработчиками, например, разработчиками Web-сайтов;
- более тесная интеграция Системы с платформами КИАС и eLibrary.ru.

2.8. Техническая реализация

Настоящий раздел посвящен различным аспектам программной реализации Системы, которые включают ее базовую составляющую (ядро), вопросы хранения и эффективного доступа к данным, кэширование информации, профилирование и автоматическое тестирование программного кода, а также модуль поиска и слияния дубликатов объектов в Системе. Представленные далее методы, механизмы и инструментальные средства системы ИСТИНА обеспечивают ее преимущества перед другими подобными разработками. В самом ядре Системы заложена расширяемость спектра видов результатов научной деятельности, поддерживаемых в Системе. Базовые модели, формы и мастера форм позволяют сравнительно легко добавлять новые виды результатов, повторно используя общий код. Именно благодаря этим функциям, в Системе на настоящее время поддерживается 20 видов результатов деятельности. Механизмы и интерфейс слияния объектов в Системе, также основанные на едином базовом

наборе функций, позволяют пользователям и ответственным сотрудникам самим выполнять очистку данных, повышая, тем самым, точность и корректность данных в Системе. Методы и алгоритмы сопоставления и привязки авторов при добавлении работ служат для того, чтобы вопросы синонимии фамилий решались пользователями при поддержке алгоритмов Системы сразу же при добавлении, и статьи и другие результаты деятельности попадали на страницы к нужным сотрудникам. Таким образом, все описанные методы и средства направлены на то, чтобы в системе ИСТИНА: содержалась наиболее точная и полная информация о результатах научной и педагогической деятельности сотрудников научно-образовательных учреждений; пользователи управляли данными о себе и имели возможность формировать различные отчеты; руководители всех уровней имели все необходимые функции для анализа деятельности своих сотрудников как по статистическим показателям, так и по тематическим направлениям их исследований.

2.8.1. Ядро системы

В основе Системы лежит ядро – программный модуль, реализующий классы, методы, формы и функции, общие для многих типов объектов (рис. 1.7). Более подробная информация о составе и функциях ядра представлена в [9]. Далее кратко отметим лишь основные вопросы, касающиеся технической реализации ядра. Ядро Системы обеспечивает следующие функциональные возможности:

- единый интерфейс добавления, просмотра, редактирования и удаления результатов деятельности сотрудников;
- поиск похожих объектов, например, сотрудников, журналов и статей;
- слияние объектов, например, сотрудников или журналов;
- кэширование;
- общие функции, реализующие политику безопасности Системы, в частности, механизмы разграничения доступа к результатам деятельности сотрудников;

- вспомогательные функции.

Далее эти функции описаны подробнее.

Общий код для работы с результатами деятельности

Важнейшим отличием системы ИСТИНА от других похожих систем является изначальная ориентация не только на статьи в журналах и другие публикации, но и на многие другие виды результатов научной, научно-педагогической и научно-административной деятельности. Механизмы для поддержки многих видов подобных сущностей заложены в ядре Системы. Эти механизмы позволяют автоматически получать работающий интерфейс добавления, просмотра, редактирования и удаления нового типа результатов деятельности на основе его описания в модели. Далее представлено более подробное описание этого подхода, который включает следующие части:

- иерархия базовых классов-моделей, от которых наследуются классы, представляющие отдельные виды результатов деятельности;
- иерархия базовых форм для ввода данных пользователем;
- иерархия мастеров форм (wizards), которые обеспечивают логику и интерфейс добавления и редактирования объектов.
- функции поиска похожих объектов;
- функции слияния объектов;
- другие вспомогательные функции.

Кратко опишем первые три части этого подхода, а именно – базовые классы, формы и мастера форм. В классах-моделях описываются общие атрибуты объектов этих классов. Основная часть иерархии базовых классов-моделей в системе ИСТИНА представлена на рис. 2.36. Базовым классом, который определен в Системе и наследуется от Django-класса `Model`, является класс `MyModel`. В нем определены разделяемые другими

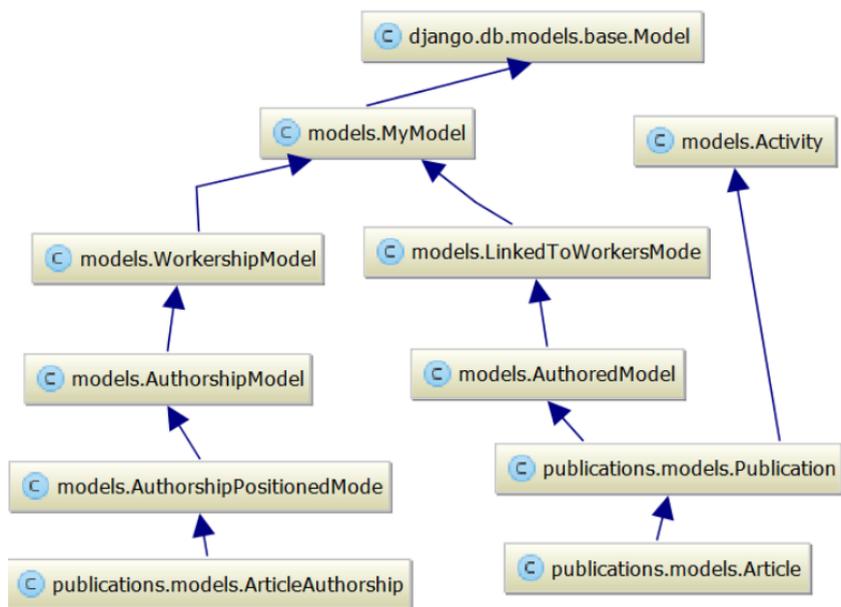


Рис. 2.36. Иерархия базовых классов-моделей в системе ИСТИНА (фрагмент).

классами самые общие атрибуты и методы. Далее иерархия классов разделяется на две цепочки. Первая из них, которая начинается с класса `WorkshipModel`, служит для задания объектов-результатов деятельности, связанных только с одним сотрудником, например, «авторство статьи», «авторство патента», «руководство диссертацией», «членство в научном обществе», «членство в редколлегии журнала». Вторая цепочка начинается с класса `LinkedToWorkersModel` и служит для задания объектов-результатов деятельности, которые могут быть связаны с группой сотрудников, например, статья, патент, отчет, НИР, доклад на конференции, учебный курс. Объекты этих типов связываются с классом `Worker` (сотрудник) как раз через объекты классов, наследованных от `WorkshipModel`, например, статья связана с сотрудниками-авторами через авторства статьи. В классах `WorkshipModel` и `LinkedToWorkersModel` заданы основные методы, относящиеся к связи «сотрудник-результат деятельности», например, получение списка сотрудников по объекту-результату деятельности в различных видах, добавление сотрудников к объекту, изменение связи между объектами и сотрудником.

На следующем уровне иерархии классов находятся модели `AuthorshipModel` и `AuthoredModel`. Они служат для задания результатов деятельности, у которых роль связанных сотрудников называется «автор». Примерами могут служить авторства статей (наследуется от `AuthorshipModel`)/статьи (наследуется от `AuthoredModel`), авторства патентов/патенты, авторства учебных курсов/учебные курсы. Эти две модели созданы для уменьшения повторного кода и не содержат сложных функций. От класса `AuthorshipModel` наследуется класс `AuthorshipPositionedModel`, который отличается от базового класса тем, что служит для задания результатов деятельности, для которых важен порядок авторов, например, статей и патентов. Отдельным классом, наследующимся напрямую от `MyModel`, является модель `Activity`. Этот класс не содержит никаких функциональных возможностей и служит своеобразной меткой

для классов, задающих отдельные виды результатов деятельности. Если такой класс наследуется, в том числе, и от `Activity`, этот факт означает, что он является «окончательным» классом, обозначающим результат деятельности. Именно эти типы объектов перечислены в качестве вариантов на странице добавления нового результата деятельности. Кроме этого, факт «окончателности» используется для формирования списка типов объектов на странице сотрудника, а также в общей статистике Системы. Отметим, что «окончательными» классами могут быть как подклассы `WorkershipModel` (например, членства в научных обществах), так и подклассы `LinkedToWorkersModel` (например, статьи, патенты или диссертации).

На нижнем уровне иерархии находятся классы, задающие отдельные виды результатов деятельности сотрудников, например, авторство статьи (`ArticleAuthorship`), отчет, диссертация. Отметим, что для публикаций создан промежуточный класс `Publication`, в котором содержатся общие функциональные возможности для двух классов публикаций в Системе, статей (в журналах и сборниках) и книг.

Основная часть иерархии базовых форм в системе ИСТИНА представлена на рис. 2.37. На верхнем уровне иерархии находится класс `MyBaseForm`, в котором задаются общие `css`-классы для отображения полей определенных типов во всех формах. Например, если в какой-либо форме на сайте поле называется `date`, `startdate` или `enddate`, то для него при отображении полностью автоматически генерируется всплывающий календарь для удобного выбора даты. Отметим, что при необходимости это поведение (как и другое, заданное в базовых классах) можно отключить для отдельной формы. Далее следует класс `MyModelForm`, в котором определяются общие методы, вызываемые при сохранении объектов по заполненным в форме данным, например, запись создателя объекта. В следующем, вспомогательном классе `Select2ModelFormMixin` определяются функции, связанные с интерактивным интерфейсом выбора сотрудников для связи с результатом деятельности. Он основан на стороннем модуле `django-`

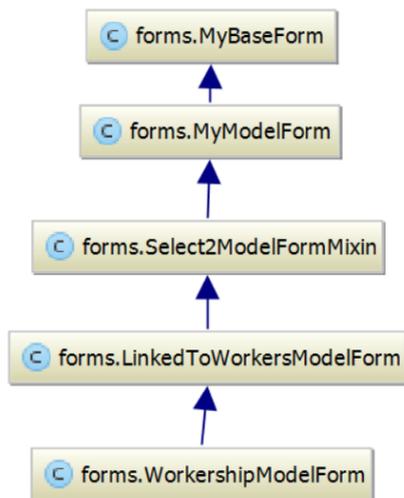


Рис. 2.37. Иерархия базовых форм в системе ИСТИНА (фрагмент).

select2¹. Наконец, в классах `LinkedToWorkersModelForm` и `WorkershipModelForm` задаются методы сохранения связи между результатами деятельности (объектами-потомками классов `LinkedToWorkersModel` и `WorkershipModel` соответственно) и выбранными пользователем сотрудниками, а также задаются поля для автодополнения имен сотрудников в этих формах.

Основная часть общей логики, связанной с добавлением и редактированием результатов деятельности, находится в классах, задающих мастера форм. Мастера форм (wizards в терминологии Django) – это классы, облегчающие реализацию форм на веб-сайте, состоящих из нескольких шагов. Пример отдельного шага мастера форм представлен на рис. 2.38. Мастера форм позволяют сохранять какую-либо информацию между шагами, динамически включать или исключать шаги на основе введенных пользователем данных, а также сохранять всю введенную информацию после того, как все необходимые формы были заполнены и проверены. В системе ИСТИНА определены два

¹<https://github.com/applegrew/django-select2>

Шаг 1 из 2. Введите информацию о учебном курсе

Информация о учебном курсе

Авторы	Воскресенский С.С., Леонтьев О.К., Симонов Ю.Г.
Название курса	Динамическая геоморфология
Год создания	1986
Организация	МГУ имени М.В. Ломоносова
Описание	

Продолжить редактирование курса

Рис. 2.38. Мастер ввода информации об учебном курсе в системе ИСТИНА, первый шаг.

основных класса мастеров форм (рис. 2.39), `MyModelWizardView` и `LinkedToWorkersModelWizardView`. В первом из них

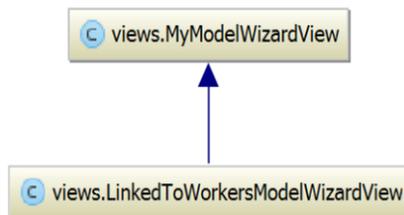


Рис. 2.39. Иерархия мастеров форм в системе ИСТИНА (фрагмент).

реализованы общие функции, например, проверка прав доступа, механизмы динамического включения-исключения шагов, сохранение и получение общих данных между шагами. Во втором задана логика, отвечающая за шаг разрешения неоднозначностей имен сотрудников. На этом шаге по фамилии и инициалам, введенным пользователем (например, Иванов А.А.) подбираются похожие сотрудники из базы данных Системы и отображается краткая информация о них, чтобы пользователь мог выбрать

нужного сотрудника из однофамильцев. Система также отображает сотрудников с фамилиями, отличающимися на одну или две буквы, учитывая тот факт, что пользователь мог ошибиться при вводе фамилии. Во многих случаях Система может с высокой степенью уверенности сделать выбор самостоятельно, и тогда один из сотрудников отображается пользователю как выбранный по умолчанию. Пользователь может, при необходимости, скорректировать выбор. В классе `LinkedToWorkersModelWizardView` также определены методы, сохраняющие выбранные связи между введенными фамилиями, инициалами и сотрудниками в базе данных. Все производимые операции записываются в журнал событий. Отдельно следует отметить мастер форм `MergeWizardView`, который реализует общие функции и интерфейс слияния различных объектов. Более подробная информация о нем представлена далее.

2.8.2. Кэширование

Целью, которую преследуют механизмы кэширования, является снижение нагрузки на базу данных, что ведет к ускорению загрузки страниц. Для этого используются программы `Memcached`¹ и `Pylibmc`², хорошо зарекомендовавшие себя в современной практике разработки приложений. Программа `Memcached` позволяет хранить кэшируемые значения в выделенной под это оперативной памяти сервера, при этом не обязательно того же, который используется для работы основной системы. Модуль `Pylibmc` обеспечивает доступ к программному интерфейсу `Memcached` в библиотеке `Django`.

Сброс кэшируемых данных

Сброс (удаление) кэшируемых данных происходит по истечении заданного срока (на текущий момент – через месяц после записи в кэш, что является максимальным значением, которое позволяет устано-

¹<http://memcached.org/>

²<http://sendapatch.se/projects/pylibmc/index.html>

вить Memcached) или же в случае, если был послан сигнал, что данные в кэше устарели и их необходимо удалить.

Построение ключа

Доступ к элементам кэша осуществляется по ключу. При кэшировании частей HTML-шаблонов ключ генерируется функцией библиотеки Django, куда передаются произвольное число дополнительных аргументов, включающих название файла-шаблона, а также дополнительные аргументы, однозначно задающие элемент кэша. Например, при кэшировании списка соавторов и результирующей информации на странице пользователя, в качестве таких аргументов используется идентификатор сотрудника.

Кэшируемые данные

На настоящее время в кэш попадают список соавторов, информация о деятельности сотрудника на его странице. Список соавторов и количество работ каждого автора является отдельным элементом кэша. В качестве дополнительных аргументов для построения ключа используется идентификатор сотрудника.

Сброс (удаление) кэшируемого списка соавторов выполняется в следующих случаях:

- добавление/редактирование/удаление авторства работы, например, пользователем, либо через интерфейс администратора;
- изменение имени одного из соавторов.

Информация о результатах деятельности сотрудника на его странице разбита по категориям. Список результатов в каждой категории каждого сотрудника является отдельным элементом кэша. В качестве дополнительных аргументов для построения ключа используется идентификатор сотрудника и название категории деятельности на английском языке. Страница сотрудника может отображаться

как со ссылками на страницы работ (на странице пользователя) или же без таких ссылок (на странице годового отчета), поэтому эти два набора данных кэшируются отдельно. Также кэшируются отдельно данные, включающие весь список результатов деятельности сотрудника, и включающие только последние несколько работ для отображения в списке похожих сотрудников при добавлении новой работы.

Сброс (удаление) кэшируемых результатов деятельности сотрудника выполняется в следующих случаях:

- добавление/редактирование/удаление работы, при этом у ее авторов сбрасывается кэш соответствующей категории;
- добавление/редактирование/удаление авторства работы;
- изменение имени автора работы.

Далее перечислены примеры особых случаев, в которых необходимо удалять кэшируемые данные. К их числу относятся следующие:

- при изменении названия журнала или сборника происходит сброс кэша категории «статьи» для авторов статей в этом журнале или сборнике;
- при изменении названия конференции происходит сброс кэша категории «доклады на конференциях» для авторов докладов, прочитанных на этой конференции.

2.8.3. Профилирование

В данном подразделе представлены результаты исследований скорости загрузки отдельных страниц или времени работы отдельных компонент Системы. Тестирование проводится с целью нахождения узких мест Системы и их устранения.

Используемые программные средства

Для замера используется Siege¹ – утилита для нагрузочного тестирования веб-серверов. Утилита была создана для того, чтобы предоставить разработчикам возможность проверить ресурсоёмкость своего кода в условиях, максимально приближенных к реальным. Программа Siege может также имитировать обращения к сайту сразу нескольких пользователей. В системе ИСТИНА Siege используется для оценки эффективности кэширования и для оценки настройки Apache. Пример использования программы под ОС Linux: `siege istina.msu.ru -c 2 -r 10`. В этом примере параметры означают следующее.

- *istina.msu.ru* – адрес страницы, тестирование которой проводится.
- *-c 1* – количество пользователей, которое будет имитировать программа. При начале работы программы каждый пользователь одновременно делает запрос по указанной ссылке. В данном случае создается два пользователя.
- *-r 10* – количество запросов от каждого пользователя. Как только пользователь получает ответ от сервера, он тут же делает очередной запрос, общее количество запросов для каждого пользователя задается этим параметром. В данном случае каждый из двух пользователей сделает по 10 запросов, итоговое количество запросов будет 20.

Результаты профилирования загрузки страниц

Измерялось время загрузки следующих страниц:

- <http://istina.msu.ru> – главная страница сайта;
- <http://istina.msu.ru/organizations/214524/workers/> – список сотрудников МГУ;

¹www.joedog.org/index/siege-home

- <http://istina.msu.ru/profile/vasenin/> – страница сотрудника с большим числом добавленных работ.

Для каждой страницы проводилось два теста – 50 последовательных запросов от одного пользователя и 5 параллельных запросов от разных пользователей. Замерялось общее время на обработку запросов, а также максимальное, среднее, минимальное время обработки. Были получены следующие результаты (в секундах):

Тестируемая страница	50 посл.	5 пар.
	общ./макс./мин./ср.	общ./макс./мин./ср.
Главная страница	39.92/2.34/0.11/0.3	2.36/2.36/0.16/0.61
Список сотрудников	197.10/5.40/3.14/3.32	9.88/8.87/4.71/7.67
Страница сотрудника	80.71/3.15/0.86/1.09	5.84/5.49/4.30/4.97

Результаты профилирования поиска похожих сотрудников

Для тестирования последнего шага добавления работ, а именно поиска похожих сотрудников, замерялось время работы при добавлении публикации с шестью авторами, производилось 10 замеров. Подробнее об алгоритме можно прочитать в разделе 2.2. В ходе тестирования были получены следующие результаты (в секундах):

Общее	Минимальное	Максимальное	Среднее
132.7	10.4	13.2	11.6

2.8.4. Поиск похожих объектов

Для повышения качества данных в Системе реализован программный механизм автоматизированного (с участием человека) поиска и объединения похожих объектов. В настоящее время он используется для объединения дубликатов конференций. Однако он может также

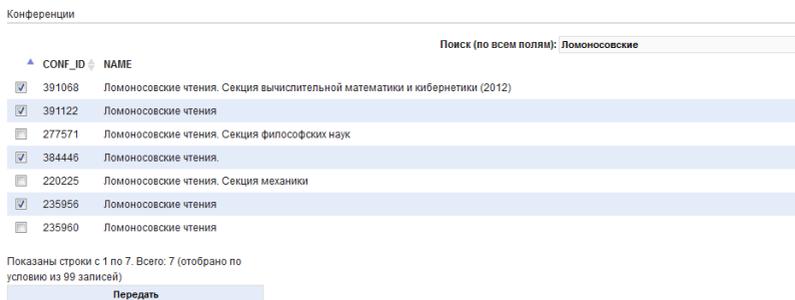


Рис. 2.40. Поиск похожих объектов.

применяться для устранения дубликатов других объектов в Системе, в том числе – различных видов результатов деятельности сотрудника, например статей или книг.

Для удобной работы с этим модулем разработан веб-интерфейс, доступный для ответственных за очистку данных в Системе. На первом шаге производится поиск объектов в базе данных по определенным полям, например, по названию (рис. 2.40). После этого пользователь выбирает возможные дубликаты объектов и переходит на страницу подтверждения, где может получить более подробную информацию о выбранных объектах (рис. 2.41). Вид информации зависит от типа объединяемых объектов. В случае конференций решение об объединении объектов происходит на основе года, места проведения, количества докладов с этой конференции, добавленных в Систему, и известных членах программного комитета. Для каждого типа объекта в Системе, для которого доступно такое полуавтоматическое устранение дубликатов, указывается описание необходимой для принятия решения информации. Эта информация и будет отображаться на данном шаге.

Рассматриваемый модуль реализован таким образом, что объединение любых объектов происходит с использованием одного и того же программного кода. В связи с этим обстоятельством модуль не требует изменений, связанных с расширением Системы и добавлением в нее новых типов объектов. Он значительно упрощает процесс поиска дубликатов и в перспективе значительно повысит качество

Вы уверены, что хотите объединить эти объекты? Желтым выделен объект, в который будут занесены данные из других объектов в этом списке, после чего они будут удалены из системы. ВНИМАНИЕ: изменения не обратимы!

ID	Название	Год	Место проведения	Кол-во докладов	Кол-во членов комитета	
384446	Ломоносовские чтения.	2012	Москва, МГУ имени М.В.Ломоносова, Шуваловский корпус.	1	0	<input type="button" value="Убрать"/>
391068	Ломоносовские чтения. Секция вычислительной математики и кибернетики (2012)	2012		2	0	<input type="button" value="Убрать"/>
277571	Ломоносовские чтения. Секция философских наук	2012	МГУ имени М.В.Ломоносова, философский факультет	2	0	<input type="button" value="Убрать"/>
391122	Ломоносовские чтения	2011	Москва	14	0	<input type="button" value="Убрать"/>
235956	Ломоносовские чтения	2001	Институт механики МГУ, г. Москва	2	0	<input type="button" value="Убрать"/>
235960	Ломоносовские чтения	1998	Институт механики МГУ, г. Москва	2	0	<input type="button" value="Убрать"/>

Показаны строки с 1 по 6. Всего: 6

Рис. 2.41. Страница подтверждения объединения дубликатов.

данных в Системе.

2.8.5. Хранение данных

В настоящем подразделе кратко представлены результаты анализа методов и средств, используемых для организации быстрого и надежного доступа к большим объемам данных в системе ИСТИНА.

Архитектура базы данных

При разработке архитектуры базы данных, предназначенной для хранения информации о результатах научно-технической деятельности учитывались следующие требования, предъявляемые к Системе.

Наличие пиковых нагрузок при обработке данных. Количество запросов и число одновременно работающих пользователей в Системе существенно изменяется в течении суток, в течении недели и в течении календарного года. Суточные и недельные пиковые нагрузки определяются рабочим графиком сотрудников. Годовые пиковые нагрузки вызваны увеличением количества обращений к Системе в период подготовки и сдачи отчетных материалов сотрудниками организации.

Необходимость резервирования данных. Поскольку хранимые

в Системе данные вводятся пользователями вручную и не могут быть получены из других источников в автоматическом режиме, необходимо обеспечить сохранность данных в случае отказов аппаратного обеспечения и повреждения физических носителей информации.

Незначительный ущерб в результате краткосрочного отключения Системы. Внесение данных в Систему производится в течение длительного времени и, как правило, этот процесс может быть приостановлен на несколько часов без существенного ущерба для отдельных пользователей и организации в целом.

Распределение нагрузки. Построение отчетных форм, например, может потребовать значительных вычислительных ресурсов, что, в свою очередь, приведет к существенному снижению производительности сервера базы данных. Поскольку подобные временные замедления работы сервера базы данных не должны оказывать существенного влияния на работу других пользователей Системы по вводу данных, необходимо распределять нагрузку на сервера базы данных.

Необходимость поддержки кратковременной версии данных. Построение сложных аналитических отчетов в Системе не должно блокировать ввод и редактирование данных другими пользователями. Сервер базы данных автоматически должен порождать несколько версий данных по состоянию на различные моменты времени. Аналитическим отчетам предоставляются в монопольном режиме версия данных по состоянию на момент запуска отчета, а пользователям предоставляются наиболее актуальная версия данных.

Требования по обеспечению безопасности данных. Необходимо обеспечить совместный доступ пользователей к пересекающимся массивам данных, разграничивая права доступа пользователей по просмотру, добавлению и изменению представленных в Системе результатов научной деятельности.

Масштабируемость. На первом этапе внедрения Системы предусматривалось подключение к ней только сотрудников МГУ имени М.В. Ломоносова. Судя по существующим запросам, на

следующем этапе не исключено подключение и создание на платформе ИСТИНА информационно-аналитических систем других вузов и научных центров России. Архитектура Системы должна предусматривать соответствующее увеличение нагрузки как на сервера приложений, так и на базу данных Системы.

Требования к операционной системе. Поскольку большинство модулей Системы реализованы на платформе операционной системы Linux, выбранная Система Управления Базой Данных (СУБД) также должна функционировать под управлением этой операционной системы.

Выбор СУБД определялся требованиями к масштабируемости Системы и базы данных, к резервированию и распределению ее нагрузки, а также требованием к операционной системе. Первым трем требованиям в полной мере удовлетворяют только две промышленные СУБД: Oracle и MS SQL. В связи с тем, что СУБД MS SQL не может работать под управлением операционной системы Linux для реализации всех необходимых функциональных возможностей по хранению и обработке данных, в качестве основной выбрана СУБД Oracle.

Учитывая наличие пиковых нагрузок в Системе, расчет вычислительных мощностей сервера СУБД и его памяти производился исходя из возможной нагрузки на базу данных в 1000 запросов в секунду. Размер кэша СУБД рассчитывался на основе допустимого порогового значения попадания в кэш $K_{bf} 0.95$, где $K_{bf} = C_{buf}/C_{all}$; C_{all} - общее количество прочитанных блоков; C_{buf} - количество прочитанных блоков, найденных в буфере. Средняя активная нагрузка на базу данных Системы на настоящее время составляет менее 40 запросов в секунду, а коэффициент попадания в кэш $K_{bf} = 0.995$. С учетом изложенного, база данных может выдерживать двадцатикратные пиковые нагрузки по интенсивности запросов пользователей.

Обеспечение резервирования осуществляется на основе двухступенчатой схемы, обеспечивающей двойную защиту данных пользователей. Актуальная копия данных создается на основе механизма StandBy. Все изменения после попадания в журнал копируются на

выделенный резервный сервер и вносятся в резервную базу данных. Таким образом, в случае уничтожения основного сервера имеется возможность произвести оперативное переключение на готовый резервный сервер с незначительной потерей данных. Кроме этого, выполняются ежедневные снимки актуального состояния базы данных. Их использование позволяет восстановить состояние данных, которое было в Системе несколько дней и даже месяцев назад.

Принимая во внимание возможность краткосрочных отключений Системы, в ее архитектуре не предусмотрено наличие серверов, выполняющих роль основного сервера при проведении регламентных работ по обслуживанию Системы с отключением сервера базы данных. Для проведения таких работ на основном сайте Системы заблаговременно размещается объявление о временном отсутствии доступа к Системе с указанием времени начала и окончания проведения таких работ.

Кратковременная версия данных поддерживается ядром выбранной СУБД Oracle и позволяет получить результаты запросов за прошедший период в течении не менее чем одного часа с момента внесения пользователем изменений данных.

Доступ к данным осуществляется на основе ролевой модели, которая реализуется на уровне СУБД. Модули Системы, которым не требуется изменять данные, получают доступ к роли с правами только на чтение данных. Кроме того, архитектура Системы предусматривает построение виртуальных баз данных. Такой подход позволяет предоставлять надежную защиту при чтении данных пользователями разного уровня доступа или разных организаций. Однако, пока Система функционирует в рамках одной организации, эта возможность не используется, и разграничение прав доступа осуществляется на уровне сервера приложений.

Масштабируемость Системы и увеличение ее производительности может обеспечиваться за счет увеличения оперативной памяти, использования дополнительных серверов хранения и обработки распределенных данных, а также переносом части запросов на резервные StandBy-сервера.

Таким образом, разработанная и представленная на рисунке 2.42 архитектура Системы хранения данных удовлетворяет представленным выше требованиям к ней и позволяет обеспечить качественное хранение и обработку данных в системе ИСТИНА.

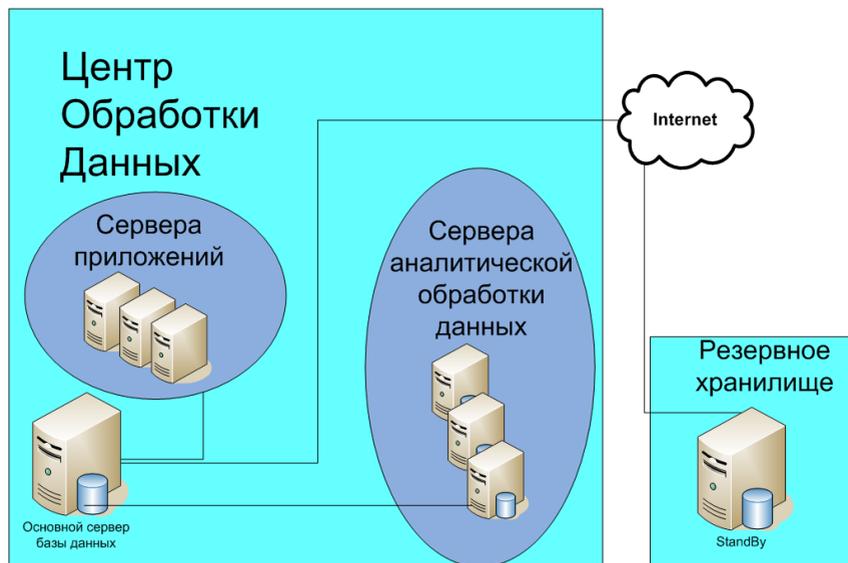


Рис. 2.42. Организация хранения данных в ИСТИНЕ.

Логическая структура базы данных

Проектирование логической структуры базы данных ИСТИНА проводилось с использованием теории нормализации отношений для реляционных СУБД. Целью нормализации является построение структуры отношений, обеспечивающих максимальное удобство использования, отсутствие большого количества логических операций для проверки допустимости вносимых пользователем изменений данных и простоту запросов. Существует несколько критериев, которые позволяют оценить качество проектирования реляционной базы данных. Эти критерии называются нормальными формами. Для построения структуры ИСТИНА использовалось

последовательное улучшение модели с использованием теории нормализации и ее проверку на соответствие представленным далее нормальным формам.

Теория нормализации основывается на наличии той или иной зависимости между полями таблицы. Определены два вида таких зависимостей, а именно — функциональные и многозначные.

Функциональная зависимость. Поле В таблицы функционально зависит от поля А той же таблицы в том и только в том случае, когда в любой заданный момент времени для каждого из различных значений поля А обязательно существует только одно из различных значений поля В. Отметим, что здесь допускается, что поля А и В могут быть составными.

Полная функциональная зависимость. Поле В находится в полной функциональной зависимости от составного поля А, если оно функционально зависит от А и не зависит функционально от любого подмножества поля А.

Многозначная зависимость. Поле А многозначно определяет поле В той же таблицы, если для каждого значения поля А существует хорошо определенное множество соответствующих значений В.

Одной из основных характеристик, определяющих качество проектирования логической структуры базы данных, является степень нормализации, показывающая, каким нормальным формам удовлетворяет разработанная логическая схема данных.

1-ая нормальная форма (1НФ). Отношение находится в 1НФ, если значения всех его атрибутов атомарны. Этот факт означает, что поиск по данному полю будет проводиться, как поиск по единому целому. Несоблюдение 1-ой нормальной формы влечет за собой необходимость поиска по части поля, что приводит к большим временным затратам на поиск.

2-ая нормальная форма (2НФ). Отношение находится во второй нормальной форме, если оно находится в 1НФ и любой неключевой атрибут полностью зависит от первичного ключа целиком.

3-я нормальная форма (3НФ). Отношение находится во второй

нормальной форме, если любой атрибут не являющийся первичным или альтернативным может быть однозначно определен только первичным или альтернативным ключом.

Нормальная Форма Области-Ключа (НФОК). Отношение находится в НФОК тогда и только тогда, когда любое ограничение следует или из определения области домена, или из свойств ключа.

Логическая структура данных, разработанная для хранения информации в системе ИСТИНА, удовлетворяет 3-ей нормальной форме, и большинство отношений удовлетворяет нормальной форме области-ключа, что является хорошим показателем качества разработанной структуры.

Модели индекса

Обеспечение быстрого поиска как по запросам пользователей, так и при проведении автоматической обработке загружаемых материалов является важным аспектом, характеризующим функциональные возможности базы данных системы ИСТИНА. При добавлении описания новых научных результатов выполняется их индексирование. Процесс индексирования включает такие шаги, как разбор данных и запись информации в базу данных. Существует несколько моделей индекса и методов ускорения поиска, к числу которых относятся перечисленные далее.

- Индексы на основе В*tree-дерева. При этом в конструкции, похожей на конструкцию В-дерева, хранятся пары (ключ, идентификатор строки), где ключ представляет собой набор полей таблицы, по которым она индексируется, а идентификатор строки — ссылка на место, в котором расположена соответствующая запись. В*tree-дерево сортируется по ключу, при этом все пары (ключ, идентификатор строки) связаны в список, что позволяет быстро просматривать записи по порядку.
- Таблицы, организованные по индексу. Таблица хранится как В*tree-дерево, что обеспечивает упорядоченность данных в

ней по определенному набору полей.

- Индексы на основе битовых карт. При таком подходе для каждого значения поля, присутствующего в таблице строится специальная структура. Каждой записи в таблице сопоставляется 0 или 1 в зависимости от значения индексируемого поля. Полученные нули и единицы записываются в одну длинную строку, которая и является битовой картой. Множество битовых карт по всем значениям поля составляют индекс.

Проведенный анализ показал, что индексы на основе битовых карт неприменимы для ИСТИНЫ, поскольку при каждом изменении иницируют блокировку большого диапазона записей в модифицируемой таблице. Таблицы, организованные по индексу также имеют существенные недостатки, обусловленные наложением в используемой СУБД существенных ограничений на работу с такими таблицами. В этой связи, в системе ИСТИНА на основном сервере используются B*tree индексы. Они наилучшим образом позволяют организовывать поиск по одиночным предикатам без установки длительных блокировок на всю таблицу при коротких, но частых, операциях модификации данных.

2.8.6. Автоматическое тестирование кода

Тестирование программного кода является важным элементом процедуры проверки соответствия предъявляемым к нему требованиям. Вопросы контроля качества разработки, верификации добавляемых данных и автоматизации тестирования приобретают особую важность в связи с постоянным добавлением новых функций и увеличением объема кода.

Типы тестов

В среде разработки Django/Python доступно несколько различных типов тестов, которые можно разделить на следующие категории:

- doc-тесты;

- модульные тесты;
- функциональные тесты.

Кратко изложим основные характеристики, присущие каждой из перечисленных категорий тестов.

Самые примитивные из отмеченных – **doc-тесты**. В языке Python специальным образом подготовленную справку по функции или классу можно превратить в тест путем выполнения записанных в ней команд. Однако требования, предъявляемые языком Python к написанию подобных тестов, сильно ограничивают их как в реализации тестовых испытаний, так и в гибкости внедрения. По этим причинам тесты этой категории используются редко.

Самые быстрые из перечисленных выше – **модульные (unit) тесты**. Они сконцентрированы на тестировании одной единицы (unit) кода, например, функции или метода. Это очень точные тесты в том смысле, что при правильном их написании, завершившийся с ошибкой тест позволяет определить место ошибки с точностью до одного-двух десятков строк кода. В состав стандартной библиотеки Python входит модуль `unittest` для написания подобных тестов и их упрощенной интеграции в приложения.

Самыми представительными по числу предоставляемых ими возможностей среди рассматриваемых категорий являются **функциональные тесты**, которые еще называют тестами интеграции. Их основная задача – протестировать функции приложения, возможно состоящего из большого количества разнородных независимых модулей, как единого целого. В силу своей специфики они более медленные и менее точные, чем unit-тесты. Однако, они позволяют не только контролировать совместную работу различных участков кода, но и тестировать отдельные функциональные возможности приложения. Последнее обстоятельство выгодно отличает их от других категорий тестов, ориентированных на тестирование отдельных частей. Инструментарий Django включает в себя готовые средства для функционального тестирования приложений, в частности, специальные классы, например `LiveServerTestCase`, для создания подобных тестов. Взаимосвязь между функциональными и модульными теста-

ми проиллюстрирована на рис. 2.43 из [10].

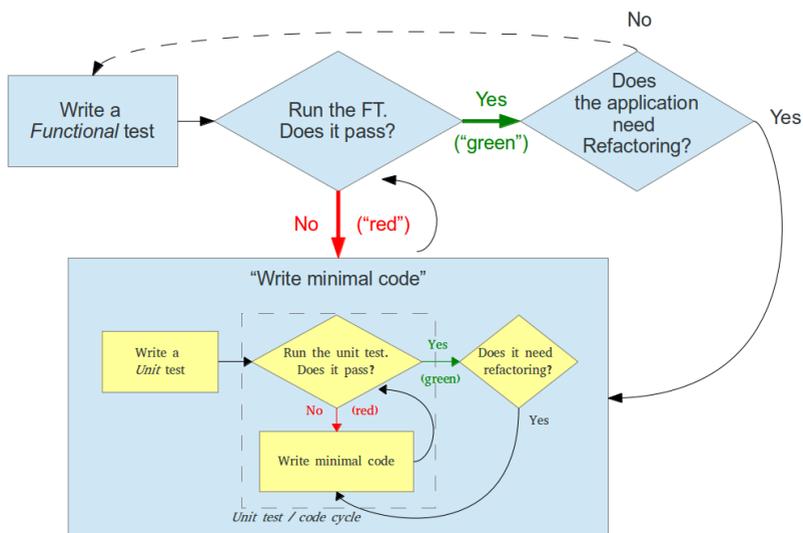


Рис. 2.43. Блок-схема тестирования.

Функциональное тестирование в Django

Остановимся более подробно на инструментальных средствах для функционального тестирования Django-приложений. Среда разработки Django имеет встроенный клиент для тестирования, который называется Django test client. Данный клиент подходит только для выполнения простейших тестовых испытаний, а именно – симуляции сетевых запросов к приложению и получения ответов, проверки корректности заполнения шаблонов веб-страниц и данных в них. Отличительными техническими характеристиками Django test client являются отсутствие поддержки скриптового языка JavaScript, который является стандартом для современных веб-приложений, а также отсутствие возможности работать с

готовыми HTML-страницами, то есть веб-страницами в том виде, в котором они отображаются для пользователей.

Более совершенным инструментарием для тестирования веб-приложений, в том числе – Django-приложений, является Selenium¹. Создатели среды разработки Django рекомендуют использовать именно его для проведения полноценного функционального тестирования. Это программное средство можно рассматривать как модуль управления веб-браузером с помощью вызовов специального API (интерфейс программирования приложений). Чтобы иметь возможность отличать API, используемое в Selenium, будем называть его Selenium API. Можно считать, что тестируемые возможности этого API ограничены только возможностями используемого веб-браузера. В настоящее время этим API поддерживаются все наиболее распространенные веб-браузеры, в том числе – Internet Explorer, Opera, Firefox и Google Chrome. Отметим также, что с 2012 года World Wide Web Consortium ведет процесс стандартизации Selenium в качестве инструментального средства для управления браузером². Реализующие Selenium API библиотеки доступны для многих языков программирования, среди которых Python, Java, Ruby, PHP, Perl. Из других важных технических особенностей следует выделить поддержку современных средств разработки веб-приложений, например AJAX и JavaScript, а также наличие подробной документации.

Далее кратко перечислим другие инструментальные средства для тестирования веб-приложений и обозначим причины, по которым было решено сделать выбор в пользу Selenium. К числу таких средств относятся:

- Watir³, который не поддерживает Python;
- twill⁴, обладающий очень ограниченными возможностями взаимодействия с элементами веб-страниц;

¹<http://docs.seleniumhq.org/>

²<http://www.w3.org/TR/webdriver/>

³<http://watir.com/>

⁴<http://twill.idyll.org/>

- Windmill¹, развитие которого прекращено;
- WebTest², который не поддерживает работу с готовыми HTML-страницами.

Библиотека `selenium-tests`

Инструментарий Selenium максимально универсален и прост в использовании. Однако это обстоятельство вызывает определенные трудности при работе с элементами веб-страниц, которые логически представляют собой единую сложную структуру. Например, чтобы считать текст из ячеек таблицы, необходимо сначала указать вручную на каждую ячейку и только потом получить текст. Подобные неудобства делают тестирование большого количества высокоуровневых HTML-элементов, таких как таблицы, формы, списки, излишне усложненным с использованием лишь встроенных в Selenium возможностей.

Чтобы преодолеть эти ограничения, был написан слой абстракции на языке Python поверх Selenium API. Позднее он был положен в основу библиотеки `selenium-tests`, который используется для проведения функционального тестирования в системе ИСТИНА. Архитектура системы тестирования представлена на рис. 2.44. Ниже мы более подробно остановимся на каждом из модулей, входящих в состав библиотеки `selenium-tests`, а также расскажем о выполненных на настоящее время тестовых испытаниях.

Возможности `selenium-tests`

Библиотека `selenium-tests` включает в себя несколько различных модулей, каждый из которых реализует собственную отдельную часть функциональных возможностей, которые необходимы для разработки программного обеспечения тестовых испытаний. Названия этих модулей имеют префикс `selenium`.

¹<http://www.getwindmill.com/>

²<http://webtest.pythonpaste.org/en/latest/>

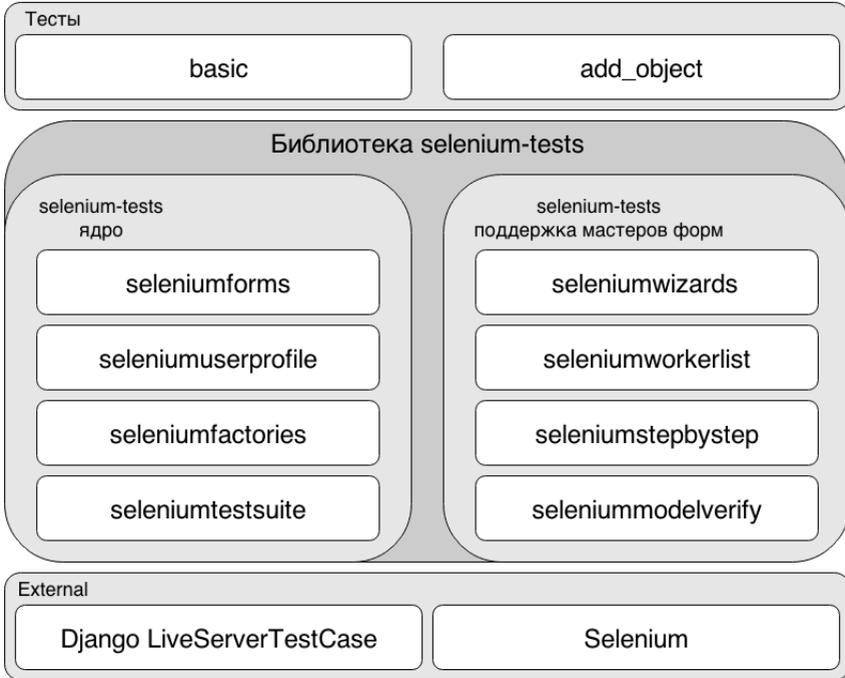


Рис. 2.44. Архитектура системы тестирования.

Одной из базовых частей `selenium-tests` является модуль **`seleniumtestsuite`**. Он реализует упомянутый выше слой абстракции над Selenium API. Собственный API, предоставляемый этим модулем, обеспечивает простое взаимодействие с высокоуровневыми элементами веб-страниц в коде тестов. С его помощью предоставляется возможность создавать тестовые испытания, в ходе которых используются такие элементы как таблицы, формы с произвольным количеством полей, переключатели, флаги и списки из них, а также ряд других элементов. Перечисленные возможности позволяют сократить усилия и временные затраты на расширение набора тестов и уменьшают вероятность программных ошибок.

К базовым частям инструментария `selenium-tests` относится модуль **`seleniumforms`**. Он содержит классы для высокоуровневой работы только с веб-формами и его функцио-

нальные возможности частично пересекаются с возможностями модуля `seleniumtestsuite`. Существенным отличием `seleniumforms` является поддержка кеширования объектов при работе с элементами страницы, что может давать заметный прирост производительности при выполнении тестовых испытаний. Кроме того, данный модуль обеспечивает взаимодействие веб-форм в HTML-страницах, которые доступны только с использованием Selenium API, и веб-форм в том виде, как они реализованы в логике Django-приложения. Эта связь между веб-формами и формами Django-приложения становится особенно важна при тестировании добавляемых данных. Причина в том, что она предоставляет программисту необходимые для создания тестов параметры, такие как типы данных, значение префикса, а также – максимально нивелирует разницу при работе с различными видами полей веб-форм.

При выполнении тестовых испытаний часто возникает потребность предварительно поместить те или иные объекты в базу данных. Для этих целей используется модуль `seleniumfactories`. Он содержит обширную коллекцию так называемых `factory`-классов, то есть классов, которые создают объекты в базе данных на основе заданного шаблона. Этот модуль опирается на инструментарий сторонней библиотеки `factory_boy`¹.

Вспомогательный модуль `seleniumuserprofile` отвечает за создание пользователя в базе данных и подготовку соответствующего теста. Его применение востребовано в тестах, которые выполняются в контексте авторизованного пользователя. Отметим, что в наборе `selenium-tests` присутствуют и тесты, которые выполняются в контексте гостя, то есть пользователя, не имеющего учетной записи в Системе.

Перечисленные выше тесты составляют ядро рассматриваемой библиотеки и уже позволяют реализовывать обширный набор тестовых испытаний с широким диапазоном тестируемых возможностей. Однако, для написания важного для системы

¹https://pypi.python.org/pypi/factory_boy/

ИСТИНА класса тестов, верифицирующих добавляемые данные, этого недостаточно. Ниже обратим внимание на модули, которые успешно решают эту задачу и обеспечивают разработчика программного обеспечения тестовых испытаний необходимым инструментарием. Напомним, что в системе ИСТИНА добавление данных организовано в виде пошагового процесса, при котором на каждом из шагов пользователь заполняет поля соответствующей формы, либо выполняет привязку к уже существующим объектам в базе данных. Верификация добавляемых данных происходит путем имитационного моделирования их ввода пользователем Системы и последующим сравнением добавленных объектов и связей между ними с эталоном.

Следующие три модуля выполняют первый этап процедуры верификации – ввод данных в Систему так, как это делал бы пользователь. Управление процессом добавления берет на себя модуль **seleniumstepbystep**, который обеспечивает контроль за успешным выполнением каждого из шагов процедуры ввода данных. Для облегчения работы программиста с функциями добавления данных, был написан модуль **seleniumwizards**. Подобно тому, как **seleniumforms** обеспечивает взаимодействие веб-форм и форм в Django-приложении, **seleniumwizards** обеспечивает взаимодействие пошаговой процедуры ввода данных и соответствующего ей объекта в Django-приложении. Такими объектами являются так называемые Django-мастера форм или Django-wizards. Мастера форм можно рассматривать как набор Django-форм с описанием логики перехода от одной формы к другой. Модуль **seleniumwizards** обеспечивает корректную инициализацию мастера форм (в силу особенностей Django это требует целого ряда операций), упрощает работу с индивидуальными шагами, используя инструментарий **seleniumforms** для соответствующих форм, а также реализует возможность синхронизации данных внутри кода теста с данными внутри пользовательской сессии. Такая синхронизация необходима, поскольку данные, генерируемые тестом, могут отличаться от данных пользовательской сессии, в контексте которой работает этот тест. Практически во всех модулях

добавления объектов в Систему на последнем шаге выполняется привязка новых данных к сотрудникам. За обработку этого шага в `selenium-tests` отвечает модуль **seleniumworkerlist**.

Второй этап процесса верификации состоит в сравнении объектов, которые были добавлены в базу в результате успешного завершения процедуры ввода данных, с эталоном. При тестировании функций добавления заранее известно какие объекты, в каком количестве, с какими значениями полей и связями друг с другом должны быть добавлены в базу. Модуль **seleniummodelverify** содержит инструментарий, который контролирует соблюдение всех этих параметров, причем универсальным образом, то есть не требует ручного указания типов полей объектов, типов связей и количества объектов. Для этого необходимо построить дерево с указанием факта связей между объектами, типами объектов и значениями полей. Такое дерево может быть указано явно, либо может быть использована встроенная в модуль эвристическая процедура построения.

Тестовые испытания, входящие в состав selenium-tests

Базовый набор тестов выделен в отдельный пакет `basic`. В этот набор входит около 50 различных тестов, которые разделены по нескольким модулям в зависимости от выполняемых задач. К их числу относятся тесты: поиска по сотрудникам и публикациям (на корректных и некорректных данных); процедур регистрации, активации и входа в Систему (на корректных и некорректных данных); тесты процедур сброса пароля и сброса активации (на корректных и некорректных данных). Этот пакет тестирует работу с Системой с точки зрения незарегистрированного пользователя.

Набор тестов для верификации добавляемых данных выделен в отдельный пакет `add_object`. В этот набор входит около 130 различных тестов. Все тесты в данном наборе осуществляют ввод только допустимых данных. С помощью данного пакета можно проверить практически каждый аспект любой процедуры добавления (в настоящее время исключением являются НИР), в том числе – механизмы поиска похожих объектов, привязки к пользо-

вателям, работа псевдонимов сотрудников и многое другое. Для всех процедур добавления проводится верификация добавляемых данных и осуществляется тестирование всех допустимых шагов.

Глава 3

Алгоритмы

3.1. Использование онтологий

Для хранения и наглядного представления всех сущностей и связей, характеризующих научную деятельность, в Системе используются решения на основе онтологии – формальной модели представления знаний, опирающейся на дескриптивную логику. Общая схема Системы, включающая онтологии, представлена на рис. 3.1. На схеме отображена онтология научной деятельности, которая является расширением широко используемой онтологии Semantic Web for Research Communities (SWRC)¹ [11]. Она включает такие концепты, как человек, организация, публикация, конференция, проект, а также связи между ними. В ней перечислены все типы работ, обрабатываемые в программном комплексе, связи между ними и учеными, конференциями, журналами и другими сущностями. В будущем предполагается разработать средство, которое будет позволять представителям различных подразделений самим формировать части онтологии, отвечающие за формулировку критериев, позволяющих оценить деятельность сотрудников этих подразделений. Затем по отмеченным частям онтологии будут автоматически или в автоматизированном режиме формироваться

¹<http://ontoware.org/swrc>

запросы к Системе, а также результаты их выполнения. Отметим, что такая методология позволит вычислять более широкий спектр запросов к данным Системы и создавать разнообразные аналитические отчеты, не затрагивая при этом ее программный код.

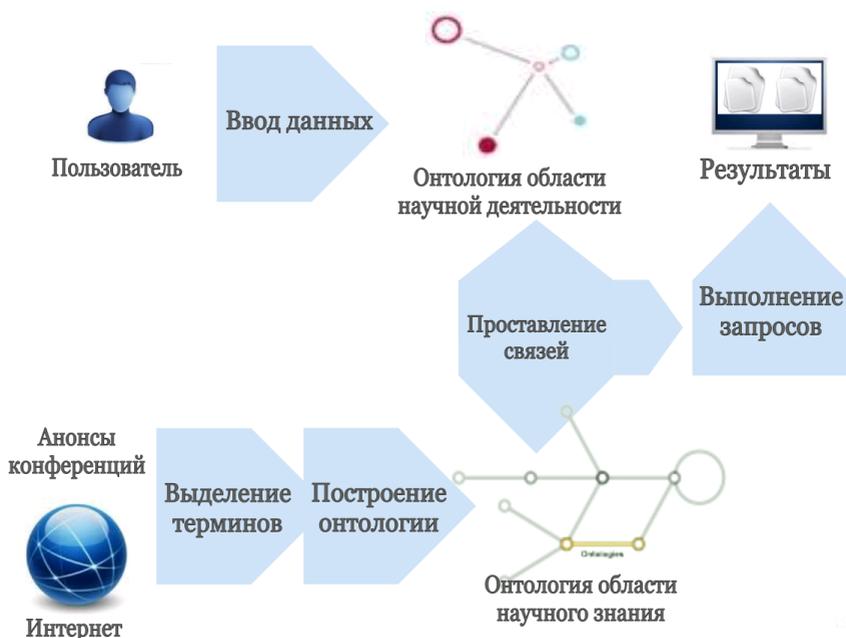


Рис. 3.1. Общая схема использования онтологий в системе ИСТИНА.

Вторым типом онтологий, которые используются в Системе, являются онтологии отдельных областей научного знания – модели конкретных областей знания на языке OWL (рис. 3.1), например, математики, информатики или физики, которые содержат основные направления, понятия и связи между ними. Для эффективного формирования таких онтологий разработан алгоритм *Sonmake* [12] автоматизированного построения онтологий областей научного знания на основе текстов анонсов научных конференций и информации из поисковых систем в Интернет. Заметим, что в нем активно используется алгоритм выделения терминов из полуструктурированных текстов *Brainsterm* [13], созданный одним из разработчиков Системы.

В качестве онтологии области знания верхнего уровня в системе ИСТИНА в настоящее время используется онтология, фрагмент которой изображен на рис. 3.2. Для удобства изложения в дальнейшем

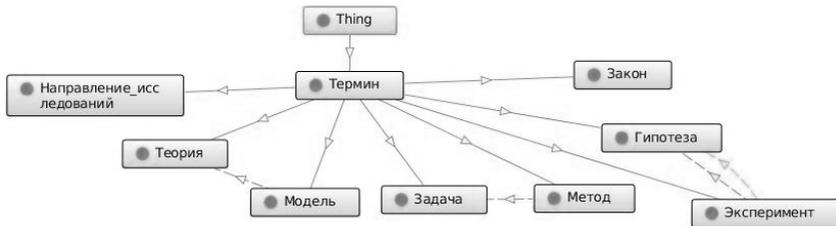


Рис. 3.2. Фрагмент онтологии научного знания.

будем называть *специальными* все понятия рассматриваемой онтологии, кроме базовых понятий *owl:Thing* и *термин*. Например, к специальным понятиям относятся *направление исследований*, *задача* и *метод*. Настройка Системы на определенную область знания требует решения задачи заполнения этой онтологии.

Для построения онтологии области научного знания в системе ИСТИНА используются методы лингвистического и статистического анализа текстов, в частности, метод шаблонов Хёрст [14] и контекстный анализ результатов поиска в Интернет с помощью поисковых систем. В качестве основного внешнего источника информации об области знания, которая является отдельным разделом науки, в настоящее время используются анонсы проходящих научных конференций, называемые в научной среде «Call for Papers»(CFP), которые можно получить из открытых источников. Документы CFP содержат основные сведения о конференциях, в том числе – название, место и даты проведения, состав программного комитета, описание конференции, список направлений области знания, работы по которым принимаются на рассмотрение. Этот подход имеет несколько преимуществ по сравнению с традиционно используемыми. В частности, один из них - тот факт, что конференции, как правило, содержат актуальную, полную и достаточно надежную информацию, отражающую текущее состояние рассматриваемого раздела науки. Они поз-

воляют также получить информацию на различных уровнях детализации при затрате сравнительно небольших усилий. На основе информации о конференциях можно выделить ключевые понятия области знания и связи между ними, то есть заполнить онтологию рассматриваемой области научного знания. Алгоритм заполнения онтологии области знания состоит из следующих этапов.

- Выделение ключевых слов (терминов) [13].
Из коллекции CFP извлекаются ключевые слова, встречающиеся в списках направлений исследований, которым посвящены конференции, а также в названиях конференций. Методика выделения основана на применении методов статистического анализа с использованием результатов поиска в Интернет с помощью поисковых систем. Каждое ключевое слово или словосочетание попадает в класс «термин» онтологии. Таким образом, после завершения данного этапа формируется онтология, заполненная экземплярами понятия «термин». На дальнейших этапах на основе лингвистического и статистического анализа ключевых слов создаются экземпляры отношений, связывающие термины между собой и со специальными классами онтологии.
- Классификация терминов по специальным понятиям онтологии.
С помощью методов лингвистического анализа (в частности, шаблонов Хёрст) для каждого ключевого слова или словосочетания проверяется, относится ли оно к одному из специальных концептов онтологии. Если предварительно удалось выделить название некоторого научного направления из названия конференции, то в онтологии создается экземпляр отношения типа «связан» между рассматриваемым ключевым словом и выделенным научным направлением. Организация терминов в иерархию в рамках каждого из специальных понятий. Для каждого специального понятия выполняется процесс организации терминов, которые являются его экземплярами, в иерархическую структуру. В частности, на этом этапе проис-

ходит построение дерева направлений исследований в рамках рассматриваемой области знания.

- Установление связей между экземплярами.

На данном этапе устанавливается связь каждого из терминов с одним или несколькими научными направлениями. Кроме того, между некоторыми ключевыми словами, являющимися экземплярами специальных понятий, устанавливаются отношения, определяемые спецификой этих понятий. Например, экземпляр «метод к ближайших соседей» концепта «метод» может находиться в отношении «используется для решения» с экземпляром «задача классификации объектов» понятия «задача».

Построенные онтологии могут использоваться, в том числе, для уточнения и расширения запросов при осуществлении пользователем полнотекстового поиска в Интернет. При уточнении запросов используется информация об онтологии интересов пользователя. Пользователь задает набор ключевых слов для поиска, а затем этот набор автоматически дополняется терминами из онтологии, соответствующими его интересам. Например, запрос «определение группы» будет преобразован в «определение группы AND (математика OR физика OR вычислительные методы)». Такое уточнение позволит Системе не показывать пользователю тексты, которые содержат определения групп, но относятся к социологии или биологии. При расширении запроса используется общая онтология. Указанное в запросе понятие дополняется списком связанных с ним дочерних понятий с учетом типа, степени и направления связи. Например, запрос «алгоритм RSA» будет дополнен понятиями «теорема Эйлера», «открытый ключ» и другими. Такое дополнение позволит найти документы, которые посвящены обсуждению связанных с алгоритмом RSA вопросов, однако не содержат его явного упоминания.

Онтологии могут использоваться и для другой цели, а именно - для тематического анализа результатов научной деятельности. Онтологию области знания можно использовать двумя способами. Во-первых, она поможет расширить возможности анализа деятельности

подразделений и организаций по тематикам. В настоящее время в качестве источника информации о тематиках используются рубрикаторы, причем тематика публикации определяется по фиксированной тематике журнала, в котором опубликована статья. Использование онтологий вместо рубрикаторов позволит более точно определить тематику отдельной работы, так как предполагается, что онтология содержит большее число понятий и отношений, чем рубрикатор, и к понятию из онтологии привязана каждая отдельная публикация, а не целый журнал. Во-вторых, использование онтологий позволит добавить в Систему такие функции, как отображение списка работ, похожих на заданную, списка ученых, занимающихся похожей тематикой, списка похожих журналов и другие тематические запросы. В настоящее время для системы ИСТИНА разрабатываются и тестируются рассмотренные средства тематического анализа, основанные на использовании онтологий.

3.2. Разрешение неоднозначности авторства публикаций

Существующие методы количественной оценки результатов научной работы ученых или организаций во многом основываются на показателях цитирования научных статей. Чем больше ссылаются на работы конкретного ученого, тем более успешной можно считать его деятельность. Получение универсальной шкалы, позволяющей сравнивать эффективность работы ученых из различных областей, связано со значительными трудностями [15]. Например, количество научных статей, которые ежегодно публикуются в области математики и биологии приблизительно совпадает. Однако количество ссылок между работами в области биологии в 10 раз больше (список источников биологической статьи в среднем в 10 раз длиннее, чем у математической работы). Этот факт означает, что абсолютные значения показателей цитирования у математиков и биологов будут различаться. Могут быть существенные различия и в рамках одной научной дисциплины, связанные с «популярностью» того или иного

направления исследований. Тем не менее, показатели цитирования рассматриваются как один из основных методов объективной оценки результативности научной деятельности [16].

Для получения показателей цитирования конкретных статей могут использоваться специализированные базы данных, такие как Web of Science, Scopus или РИНЦ. Эти системы достаточно точно определяют общее число ссылок на конкретную статью — ошибки определения статей при анализе списка литературы минимальны. Однако вычисление показателей цитирования конкретного ученого производится с большой погрешностью из-за наличия однофамильцев и неточного определения авторства статей. Например, если зарегистрировать персональный профиль в системе GoogleScholar с достаточно распространенной фамилией, то с высокой вероятностью он будет содержать статьи однофамильцев и показатели цитирования или индекса Хирша будут существенно отличаться от реального.

Рассмотрим задачу автоматического определения авторов статьи по ее библиографическим данным и основные подходы к её решению в Системе. Предполагается, что имеется Система, содержащая список ученых с некоторыми из их публикаций, для которых выполнено разрешение конфликтов авторства. Требуется разработать алгоритм, который позволит определять авторов новых публикаций, поступающих в Систему. Публикации описываются названием, списком авторов, названием журнала и другими библиографическими параметрами. Данные, которые уже находятся в Системе, считаются корректными, то есть все однофамильцы разделены и работы корректно приписаны соответствующим пользователям. В дальнейшем будем использовать слово *пользователь* для обозначения записи в базе данных, которая соответствует конкретному ученому, а слово *автор* — для обозначения строки, указанной в поле авторов новой библиографической записи. Определение авторства — это процесс сопоставления автора и пользователя. Например, одним из авторов публикации может быть указан «Иванов И.», а в системе могут быть пользователи с именами «Иванов Игорь» и «Иванов Иван». Задача алгоритма — определить, кто из этих пользователей является автором данной работы. При определении авторства могут учитываться различные па-

раметры, в частности, текстовая близость между фамилией пользователя и автора, наличие других совместных работ у выбранных пользователей, тематическая близость новой работы и предыдущих работ пользователя и другие.

Характерный сценарий применения алгоритма состоит в следующем. Текущее состояние базы данных используется в качестве обучающей выборки для алгоритма определения авторства. При добавлении новой работы, содержащей n авторов, алгоритм возвращает n рекомендаций. Каждому автору (напомним, что автор — это просто строка с его именем) может быть сопоставлен либо идентификатор некоторого пользователя, либо специальное значение, означающее что автоматический выбор сделать не удалось. Полученные данные могут использоваться для информирования соответствующих пользователей Системы о добавлении в неё новой работы. Следует отметить, что значительный практический интерес представляет смежная задача кластеризации публикаций. В данном сценарии имеется большая база библиографических записей, но никакого предварительного определения авторов не производилось.

3.2.1. Описание модели

При добавлении новой работы, содержащей N авторов (a_1, \dots, a_N) , для каждого автора a_i составляется список X_i «похожих» пользователей системы. Такой список фактически является множеством кандидатов для авторства работы и может содержать идентификаторы всех пользователей, фамилии которых совпадают с фамилиями авторов или являются их транслитерацией. Далее необходимо из каждого списка выбрать не более одного элемента таким образом, что значение некоторой функции качества F является максимальным по всем возможным наборам.

Пусть K — число пользователей системы. Каждой строке a , содержащей имя автора, можно сопоставить некоторую функцию $sim_a : \{1, \dots, K\} \rightarrow [0, 1]$, которая моделирует работу алгоритма поиска пользователей, похожих на автора a . Пусть $S = \{\cdot : \{1, \dots, K\} \rightarrow [0, 1]\}$ есть множество всех таких функций, а

P — множество конечных подмножеств S .

Оценка достоверности того, что пользователи $I = (i_1, \dots, i_q) \in 2^K$ написали статью, авторами которой указаны (a_1, \dots, a_q) , определяется функцией $F : P \times 2^K \rightarrow [0, 1]$. Первый аргумент F отражает результат применения алгоритма поиска похожих пользователей, а второй — сделанный выбор пользователей. Задача определения авторства состоит в нахождении такого набора I^* , при котором значение F максимально

$$I^* = \arg \max_{I \in 2^K} F(\langle \text{sim}_{a_1}, \dots, \text{sim}_{a_q} \rangle, \langle i_1, \dots, i_q \rangle).$$

Заметим, что первый аргумент F фиксирован и зависит только от входных данных.

При реализации указанного метода определения авторства возникают следующие задачи. Во-первых, необходимо сформировать множество параметров, влияющих на значение функции F , и дать её формальное определение. Во-вторых, необходимо реализовать эффективный алгоритм решения оптимизационной задачи. Основная задача данной работы состоит в формализации параметров F и её определение.

3.2.2. Формализация критериев оценки

Целью данного подраздела является построение формальных критериев оценки качества конкретного выбора $\langle v_1, \dots, v_N \rangle$, $v_j \in X_j$ пользователей системы, которые будут использоваться для построения функции F .

Пусть задан набор $\langle v_1, \dots, v_N \rangle$, $v_j \in X_j$, определяющий конкретный выбор пользователей системы по заданному набору N авторов. Построим по этому набору гиперграф $H = (V, E)$, где $V = \{v_1, \dots, v_N\}$ — множество вершин, а $E = \{e_1, \dots, e_m\}$ — множество ребер. Наличие гиперребра означает, что ученые, соответствующие входящим в состав гиперребра вершинам, имеют совместные публикации. То есть H отражает историю совместных публикаций пользователей V . Введем обозначения $E' = \{e \in E : |e| = 1\}$, $V' = V \setminus E'$,

$N' = |V'|$, где $|e|$ обозначает степень гиперребра (мощность множества e). Каждому $u \in 2^V$ ставится в соответствие $B(u)$ – множество общих статей авторов из u . Для любого $e \in E$ выполнено: $\exists u \in 2^V, e \subset u : B(e) = B(u)$. Это равносильно следующему условию: $\exists v \in V : B(e) = B(e \cup \{v\})$. Для каждого $e \in E$ определена функция $P(e) = |B(e)|$. Далее из всех рёбер удалим вложенные в них подрёбра и все ребра $e : |e| = 1$, то есть выполнено: $\forall e \in E \exists e' \in E, e' \subset e$ и $\forall e \in E : |e| \neq 1$. Полученное множество рёбер опять обозначим через E .

Каждая вершина v графа H однозначно соответствует автору $a(v)$ работы. Для удобства введем обозначение $p(v) = \text{sim}_{a(v)}(v)$, то есть $p(v)$ определяет степень «похожести» пользователя v соответствующему автору.

Выделим параметры гиперграфа, которые будут влиять на значение функции F :

- $r(H)$ – ранг гиперграфа H (максимальная степень ребра). Это максимальное число авторов, имеющих общие статьи. Следовательно, чем ближе $r(H)$ к N , тем больше значение F . Пусть максимум достигается на ребре g , тогда положим $C = P(g)$.
- $\bigcap_{j=1}^m e_j$ – характеризует, сколько авторов имеют соавторство со всеми авторами из V . Чем больше $\gamma(H) = \left| \bigcap_{j=1}^m e_j \right|$, тем больше значение F .
- $k(H)$ – количество компонент связности H . Нахождение подмножества авторов в одной компоненте связности гиперграфа показывает, что они могли писать статьи в соавторстве. Чем меньше $k(H)$, тем больше значение F .

Назовем *трансверсальным* множество $T \subset V'$, если $\forall e \in E: T \cap e \neq \emptyset$. $\tau(H)$ – число трансверсальности H – мощность минимального (по мощности) трансверсального множества.

- $\mathbb{T} = \mathbb{T}(H)$ – семейство всех трансверсальных множеств $T \subset V$, мощность которых равна $\tau(H)$.

Каждое из этих множеств является «базисом» гиперграфа H , так как если $T = \{a_1, \dots, a_k\}$, $k = \tau(H)$, то положив $A_j = \{e \in E : a_j \in e\}$, получаем, что $\bigcup_{j=1}^k A_j = V'$. T является минимальным набором с таким свойством. Для каждого $T \in \mathbb{T}$ аналогично строим соответствующее ему множество $\{A_j\}$ и вычисляем следующие функции:

$$f(T) = \max\{1 \leq j \leq k : \bigcap_{s=1}^j A_{i_s} \neq \emptyset\},$$

где (i_1, \dots, i_k) – перестановка $(1, \dots, k)$,

$$g(T) = \max_{1 \leq j \leq k} |A_j|$$

$$\beta_1(T) = \sum_{j=1}^k p(a_j)$$

$$\beta_2 = \max_{1 \leq j \leq k} p(a_j)$$

$$\beta_3 = \min_{1 \leq j \leq k} p(a_j)$$

Верны следующие соотношения: $1 \leq f(T) \leq \tau(H)$, $1 \leq g(T) \leq N'$, $0 < \beta_1 \leq \tau(H)$, $0 < \beta_3 \leq \beta_2 \leq 1$.

$f(T)$ показывает, насколько «близки» базисные наборы A_j . Заметим, что если $f(T) = \tau(H)$, то $k(H) = 1$, так как каждое A_j является связным множеством гиперграфа (пересечение всех рёбер, входящих в A_j непусто, оно содержит a_j по построению). Если $f(T) < \tau(H)$, то гиперграф может быть несвязным. $g(T)$ показывает, сколько максимально соавторов у «базисных» авторов. $\beta_1, \beta_2, \beta_3$ показывают, насколько базисные авторы похожи на авторов поступившей статьи.

Определим функцию

$$\Phi(H) = \frac{1}{|\mathbb{T}(H)|} \cdot \sum_{T \in \mathbb{T}(H)} \frac{g(T)}{N'} \cdot \frac{f(T)}{\tau(H)} \cdot \exp(\beta_2(T) - 1) \cdot \frac{\beta_3(T)}{\beta_2(T)} \cdot \frac{\beta_1(T)}{\tau(H)}.$$

Функция $\Phi(H)$ принимает значения в полуинтервале $(0, 1]$, ее максимум достигается в «идеальном» случае: $g(T) = N'$, $f(T) =$

$$\tau(H), \quad \beta_1 = \tau(H), \quad \beta_2 = \beta_3 = 1$$

Назовем *независимым* множество $A \subset V'$, если: $\forall u, v \in A \quad \nexists e \in E : u, v \in e$. То есть A – множество авторов, попарно не имеющих общих статей. Пусть M – мощность максимального (по мощности) независимого множества.

- $\mathbb{A} = \{A - \text{независимое}\}$.

Пусть

$$q(A) = \begin{cases} 1, & A \text{ лежит в одной компоненте связности } H, \\ 0, & \text{иначе.} \end{cases}$$

Тогда положим

$$Q(\mathbb{A}) = \frac{\sum_{A \in \mathbb{A}} q(A)}{|\mathbb{A}|}$$

- Определим функцию

$$\alpha(H) = \max_{e \in E} P(e) \cdot \left(\frac{1}{|e|} \sum_{v \in e} p(v) \right)$$

Эта функция задается только «весовыми» значениями вершин гиперграфа и ребер, такими как количество общих статей и степень похожести авторов из базы на авторов из поступившей статьи.

3.2.3. Определение функции оценки качества набора

Пусть

$$I_{\gamma(H)} = \begin{cases} 1, & \text{если } \gamma(H)=0, \\ 0, & \text{иначе.} \end{cases}$$

Функцию F определим следующим образом :

$$F(H) = \frac{1}{|k(H)|} F'(H),$$

где

$$F'(H) = (C \cdot r(H) \cdot \exp(\gamma(H)) + I_{\gamma(H)} \cdot \Phi(H)) \cdot (1 + Q(\mathbb{A})) \cdot \alpha(H)$$

Очевидно, данная функция удовлетворяет приведенным выше условиям. Она принимает значения в полуинтервале $(0, +\infty)$, причем в «идеальном» случае $F(H) = C^2 \cdot N \cdot \exp(N)$. Далее монотонно и взаимно-однозначно отображаем F на интервал $(0, 1)$ с помощью функции $y(t) = \frac{t}{1+t}$. Чем ближе полученная функция $y(F(H))$ к 1, тем лучше соответствующий ей набор авторов.

3.2.4. Реализация и тестирование

Данная функция реализована на языке C++. На вход алгоритму подается набор строк – авторов статьи.

- **1 шаг**

Для каждого автора, выполняя запрос к базе данных, строим набор, состоящий из id похожих авторов, упорядоченный по степени похожести.

- **2 шаг**

Для каждого id из всех наборов, выполняя запрос к базе данных, строим множество статей, привязанных к данному автору.

- **3 шаг**

Выбирая из каждого набора похожих авторов по одному представителю, формируем множество авторов, на котором считаем функцию $y(F)$.

- **4 шаг**

Упорядочиваем все возможные комбинации авторов по значению функции $y(F)$.

Алгоритм реализован с помощью библиотеки STL (для работы с множествами и списками) и библиотеки OPL 4.0 (для выполнения запросов к базе данных).

Проводились следующие тесты на основе данных системы ИСТИНА по состоянию на начало 2014 года. Выбирается статья из базы данных, для которой известен результат. Результат работы алгоритма сравнивается с правильным ответом. Сравнение проводится усредненное - доля правильно распознанных авторов среди всех. Получены следующие результаты.

- В качестве результата берутся 3 набора авторов с максимальным значением функции выбора.
Всего обработано 877 статей.
Средняя доля совпавших авторов составляет 97.06%.
- В качестве результата берется 1 набор авторов с максимальным значением функции выбора.
Всего обработано 877 статей.
Средняя доля совпавших авторов составляет 92.79%.

С учётом изложенного выше, можно констатировать, что в результате решения рассматриваемой задачи разработан формальный метод оценки степени достоверности разрешения неоднозначности авторства публикаций. Тестирование предлагаемого алгоритма показывает, что на реальных данных достигается более чем 92% точность определения авторов работы. Дальнейшая работа на данном направлении может быть связана с разработкой более эффективных алгоритмов минимизации целевой функции F , расширение множества используемых критериев за счет анализа выходных данных журналов, добавление элементов семантического анализа названий работ и их аннотаций.

3.3. Критерии оценки результативности научных исследований

Как уже отмечалось ранее, любые абсолютные числовые значения, вычисленные на основе анализа публикационной активности и других видов научной деятельности, не следует использовать для сравнения различных областей. Данные оценки можно рассматривать как

порядковую характеристику. Если у одной организации оценка в два раза выше, чем у другой, то этот факт означает, что потенциал первой выше, чем у второй. Однако при этом нельзя утверждать, что он реально выше в два раза. Учитывая то обстоятельство, что в рамках данной работы оценка потенциала производилась только на основании данных о публикациях в научных журналах, отображение различных предметных областей на одном графике является корректным. При этом следует учитывать следующие обстоятельства:

- в организации могут быть развиты предметные области, которые в силу ограниченного объема данных получили меньшую оценку;
- тематическое деление основано на тематической принадлежности журналов, а публикации в междисциплинарных журналы повышают потенциал организации сразу в нескольких областях.

Критерии оценки эффективности

Разработка системы оценки научного потенциала требует решения следующих вопросов:

- формирование перечня показателей о результатах научно-исследовательской и педагогической деятельности, которые могут коррелировать с уровнем научного потенциала;
- определение состава первичных данных, необходимых для проведения анализа;
- разработка методики количественной оценки или установления линейного порядка для выбранных показателей;
- разработка методики сбора первичных данных.

Основные показатели, влияющие на оценку результативности научно-педагогической деятельности

Для оценки результативности научно-педагогической деятельности отдельных учёных, научных коллективов и организаций могут использоваться перечисленные далее показатели, которые можно разделить на три категории.

1. Результативность научных исследований:

- статьи в журналах;
- статьи в научных сборниках;
- монографии, главы монографий;
- патенты, свидетельства о регистрации прав на программное обеспечение;
- доклады на конференциях;
- руководство и участие в научно-исследовательских проектах;
- защищённые диссертации.

2. Педагогическая деятельность:

- руководство диссертационными и дипломными работами;
- авторство учебных и учебно-методических изданий;
- авторство учебного курса;
- преподавание (чтение) учебного курса.

3. Признание в профессиональном сообществе:

- членство в редколлегиях журналов;
- членство в редколлегиях сборников;
- членство в программных комитетах конференций;
- членство в диссертационных советах;
- членство в экспертных советах;
- оппонирование диссертационных работ;

- наличие премий и наград;
 - почётное членство в различных научных обществах;
 - стажировка сотрудников других организаций.
4. Для оценки научного потенциала организации дополнительно можно рассматривать такую характеристику, как наличие в организации молодых сотрудников и их вовлеченность в научные исследования.

Очевидно, что перечисленные параметры отражают степень востребованности в профессиональном сообществе результатов, полученных отдельными учёными или научными коллективами. Например, если сотрудники организации привлекаются к работе экспертных советов, программных комитетов международных конференций, проводят стажировку сотрудников других организаций или публикуют свои результаты в ведущих научных журналах, то это может означать высокий научный уровень и потенциал сотрудников данной организации. Для получения качественных оценок научного потенциала необходимо производить объективную количественную оценку перечисленных параметров. Оценку научного потенциала предлагается производить на основании количественных оценок отдельных параметров. Следует подчеркнуть, что в различных областях науки могут существовать свои характерные особенности представления результатов и возможность построения единой системы оценки эффективности, которая позволяет сравнивать представителей различных областей науки, представляется маловероятной.

Состав первичных данных, необходимых для проведения объективного анализа

Для проведения автоматической оценки научного потенциала, например, региона необходимо наличие следующих первичных данных, описывающих результаты научно-педагогической деятельности сотрудников организаций.

- Сведения о научных публикациях, включающие выходные данные публикации, место работы авторов.
- Для проведения тематического анализа желательно также иметь ключевые слова и аннотации работ.
- Сведения о научных конференциях, включая списки докладчиков и членов организационного и программного комитетов.
- Сведения о защищённых в организации диссертационных работах, включая информацию об официальных оппонентах.
- Сведения о членстве сотрудников организаций в экспертных и диссертационных советах, научных обществах и наличия премий и наград.
- Данные о возрастном составе сотрудников организаций.

Особенности оценки научного потенциала организации на основании данных о публикациях в открытых периодических научных изданиях

Получение всего объёма данных, необходимого для вычисления обоснованной оценки научного потенциала представляется сложной задачей. В настоящее время не существует единого источника информации, которых охватывал бы весь спектр параметров. Доступный для анализа массив данных включает сведения о публикациях в периодических научных изданиях, что несомненно является одним из важнейших результатов научной деятельности. Информация о публикациях сотрудников организации может являться основой для оценки её научного потенциала. Отметим, что потенциал организации целесообразно оценивать относительно выбранной предметной области: если сотрудники имеют признанные результаты в области дифференциальной геометрии, то это не должно влиять на оценку потенциала организации в области истории средних веков.

Основные критерии, которые учитываются при оценке научного потенциала на основании анализа публикаций в периодических изданиях, перечислены ниже.

- Оценка уровня периодических изданий, в которых авторы организации публикуют свои результаты.
- Наличие публикаций в изданиях других регионов. Большинство университетов Российской Федерации издают собственные научные журналы, которые входят в перечень периодических изданий, рекомендованных ВАК, в частности, большинство серий «Вестник» университета включены во все необходимые списки. Можно ожидать, что публикации сотрудников университетов в таких журналах являются достаточно регулярными и в качестве одного из критериев значимости работы можно рассматривать относительное количество публикаций в журналах, выходящих в других регионах и странах.
- Наличие новых специалистов. При наличии данных о публикациях сотрудников организации за прошедшие несколько лет можно оценить число новых специалистов. Критерием оценки может быть дата первой публикации учёного. Если первая публикация учёного (в данной организации) попадает в трехлетний период, то его можно считать новым сотрудником (аспирантом или привлечённым для работы). Наличие новых сотрудников отражает положительную динамику организации в выбранной области.
- Динамика научного роста новых специалистов. Организация с высоким уровнем научного потенциала должна обладать новыми сотрудниками, количество и качество (в соответствии с введённым выше методом оценки качества научного журнала) публикаций которых увеличивается со временем.
- Наличие научных контактов. Оценка данного параметра предполагает наличие более широкого набора данных, так как для корректной оценки необходимо знать места работы всех соавторов сотрудников организации. В основе данного критерия лежит гипотеза, что организации, обладающие высоким потенциалом должны сотрудничать с другими организациями и отдельными учёными в своей области. В качестве количественной оценки объёма научных контактов можно рассматривать

процент публикаций, которые включают соавторов из других организаций.

- Наличие устойчивых творческих коллективов. Перечисленные выше критерии применимы к любой группе учёных, в том числе, ко всем сотрудникам организации. Анализ графа соавторства позволяет выделить группы авторов, которые часто публикуют совместные работы. Во многих случаях такая группа может соответствовать сотрудникам структурного подразделения организации (кафедра, лаборатория). Наличие в организации эффективных научных коллективов (показатели которых превосходят средние значения) должно являться критерием эффективности организации.
- Общее число авторов и общее число публикаций. Данный параметр отражает эффективность творческих коллективов и должен использоваться с учётом общего числа сотрудников организации.

Каждый критерий является некоторой количественной оценкой. Итоговая оценка получается как взвешенная сумма, где весовые коэффициенты отдельных параметров определяются в соответствии со спецификой предметной области и могут быть откорректированы экспертом, который проводит оценку потенциала.

Необходимо отметить следующие особенности количественной оценки результативности научных исследований, которые необходимо учитывать при интерпретации результатов анализа.

- Несмотря на то, что оценка потенциала представляется числом, сравнение между собой организаций из различных предметных областей некорректно. Причина связана с тем, что в различных областях модели цитирования и соавторства существенно различаются. Например, число ссылок на математические работы на порядок меньше, чем число ссылок на работы в экспериментальных науках.
- Показатели цитирования журнала являются косвенным свидетельством качества научных статей. В журнале с высоким по-

казателем цитирования могут быть опубликованы как высокоцитируемые работы, так и работы, которые не цитируются.

- Качественные научные результаты могут быть опубликованы в силу различных субъективных факторов в журналах с низким импакт фактором.

Оценка научного потенциала региона и определение его основных научных направлений может производиться усреднением указанных параметров по отдельным организациям. Таким образом возникает естественная иерархическая структура: сотрудник; творческий коллектив; организация; регион.

Определение коэффициента значимости журнала

Научные журналы могут быть разделены на тематические группы (например, в соответствии с классификацией ГРНТИ). Каждый журнал обладает индексом цитирования, что позволяет упорядочить журналы в группе в порядке убывания значений индекса и сопоставить журналу его относительную позицию в этом списке (наиболее часто цитируемые издания будут иметь значения близкие к единице, а менее цитируемые — ближе к нулю). Чем выше значение данного показателя у журналов, в которых авторы организации публикуют свои работы, тем более вероятно, что они обладают признанными результатами. График зависимости коэффициента значимости «к» научного издания от его относительной позиции представлен на рис. 3.3.

Использование коэффициента значимости научного издания позволяет более точно определить, какие научные направления развиваются в выбранной организации или регионе. Большое количество публикаций не означает, вообще говоря, что в данной организации хорошо развито конкретное научное направление. Значительное количество публикаций могут быть представлены в журналах, которые не входят в число ведущих изданий в своей предметной области. Следует однако отметить, что в случае конкретного учёного отсутствие публикаций в ведущих журналах не всегда означает отсутствие

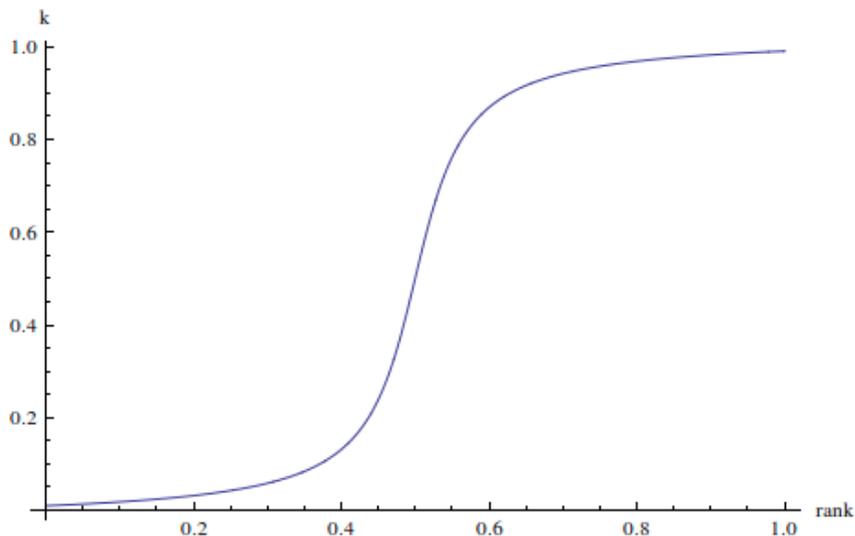


Рис. 3.3. Функция коэффициента значимости журнала.

качественных результатов. Распределение количества ссылок на публикации одного журнала очень неравномерно, а многие авторы традиционно публикуют свои результаты в определённых журналах. С другой стороны, резкое снижение относительной позиции региона при переходе от количественного к весовому методу расчёта может свидетельствовать о более низком качестве результатов.

3.4. Анализ графа соавторов

При наличии данных, позволяющих разрешить неоднозначность авторства публикаций, то есть разделить авторов с одинаковыми фамилиями, если они являются разными учёными, появляется возможность проведения автоматического анализа графа соавторства и выделения творческих коллективов и их руководителей. Отметим, что частично задачу определения авторства позволяют решить сведения о месте работы авторов. Однако, ввиду возможности смены места работы или одновременной работы одного научного сотрудника в

нескольких организациях, такое деление не может быть точным. Основными прикладными задачами анализа графа соавторов являются задача автоматического выявления руководителей творческих коллективов и определение его состава. Такие данные могут использоваться как самостоятельно, так и при решении задач оценки эффективности работы подразделений или научных групп.

3.4.1. Выделение ключевых узлов графа соавторов

Основная идея выделения руководителей коллективов состоит в выделении таких вершин в графе соавторства, которые связывают различные группы авторов, не имеющих общих публикаций. Данная модель предполагает, что при наличии нескольких коллективов, работающих в одной предметной области, наиболее вероятным является включение в авторский коллектив одной группы наиболее активных или влиятельных авторов из другого коллектива.

Сформулируем данную задачу более формально. На вход поступает граф, в котором множество вершин V соответствует множеству авторов публикаций, а ребро означает наличие по крайней мере одной совместной работы данных авторов. Для каждой пары вершин в графе G можно построить множество кратчайших путей, соединяющих эти вершины. Коэффициентом значимости $B(v)$ вершины v будем называть количество кратчайших путей между всеми парами вершин в графе, которые проходят через данную вершину

$$B(v) = \sum_{(u,w) \in V \setminus \{v\}} n_v(u, w),$$

где $n_v(u, w)$ обозначает число кратчайших путей, проходящих через v , а суммирование ведётся по всем парам вершин, отличных от v . В терминах рассматриваемой задачи, вершины с большим значением функции $B(v)$ соответствуют учёным, которые объединяют разные коллективы.

Рассмотрим в качестве примера массив данных о публикациях авторов из Московского университета, поскольку в данной коллекции разделение публикаций по отдельным учёным производится са-

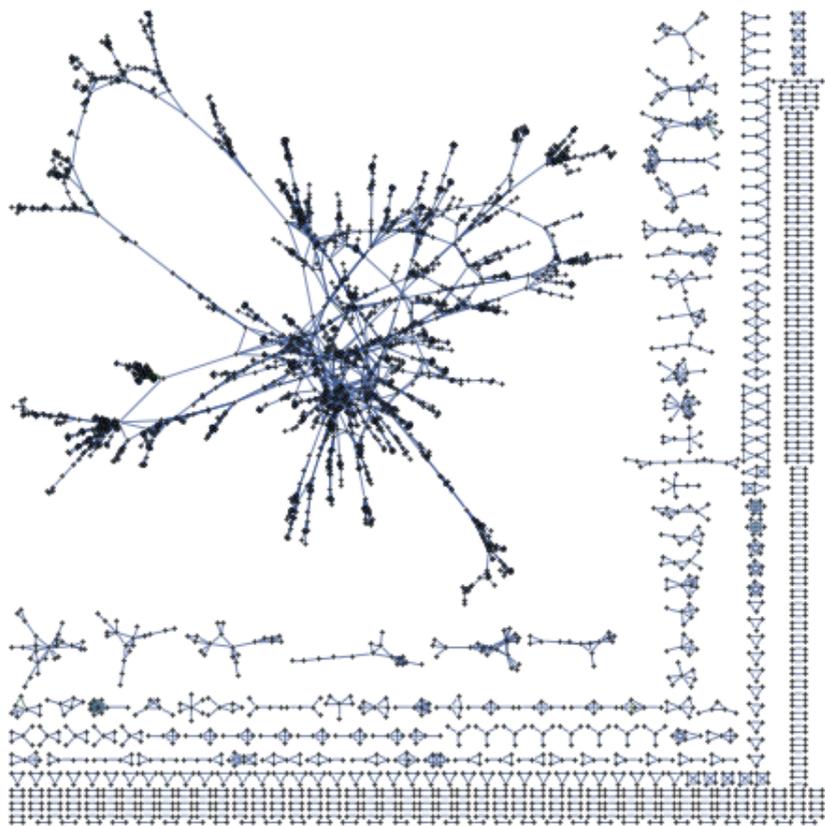


Рис. 3.4. Граф соавторства для публикаций 2010 года.

ими авторами. На рисунке 3.4 представлен граф соавторства по публикациям 2010 года. Видно, что выделяется достаточно большое число изолированных коллективов и большая компонента связности.

Вычисление коэффициентов $B(u)$ приписывает максимальные значения вершинам, которые соответствуют деканам некоторых факультетов, заведующим кафедрам, и известным учёным. Таким образом введённый параметр может косвенным образом отражать высокий уровень влияния конкретного учёного в профессиональной среде. Следует отметить, что данный метод анализа обладает всеми недостатками, связанными с неравномерным распределением среднего числа соавторов в различных областях науки. Кроме того, учёные, работающие в экспериментальных областях, обладают преимуществом перед теми, кто занимается фундаментальными исследованиями. Ниже приведены данные о среднем числе соавторов публикаций сотрудников нескольких факультетов различного профиля:

- 5.14 Факультет биоинженерии и биоинформатики;
- 4.93 НИИ ФХБ им. А.Н.Белозерского;
- 4.84 Факультет фунда. физико-химической инженерии;
- 4.54 Факультет фундаментальной медицины;
- 2.18 Механико-математический факультет;
- 1.2 Философский факультет;
- 1.2 Филологический факультет;
- 1.16 Юридический факультет.

В верхней части списка оказываются направления, которые были сформированы в последние десятилетия. При этом общее число публикаций в расчёте на одного автора (то есть общее число публикаций делённое на среднее число соавторов) оказывается сопоставимым для различных подразделений:

23504.65	Физический факультет;
21335.94	Биологический факультет;
18213.20	Химический факультет;
14665.12	Механико-математический факультет;
12656.50	Географический факультет;
11899.91	Геологический факультет;
11448.77	Исторический факультет.

Другая тенденция, которая наблюдается при изменении объёма данных, – это укрупнение компоненты связности с ростом глубины анализа. Рассмотрим граф соавторства по публикациям последних пяти лет (рис. 3.5). Наглядно видно, что при большем числе авторов количество компонент связности уменьшилось. Такая тенденция отражает естественный процесс развития научных интересов учёных.

В результате проведённого анализа можно сделать следующие выводы:

- введённый коэффициент значимости вершины отражает степень влияния конкретного учёного;
- анализ графа соавторства целесообразно ограничивать глубиной анализа в два-три года;
- необходимо разработать методы, учитывающие характерные особенности отдельных областей науки.

3.4.2. Поиск устойчивых авторских коллективов

В структуре графа соавторства выделяется одна крупная и несколько мелких компонент связности. Это означает, что существует большая группа авторов, для которой между любой парой участников существует цепочка соавторов. Например, автор А написал работу в соавторстве с Б, Б, в свою очередь, написал совместную статью с В, а В опубликовал работу с Г. К каждой компоненте связности были применены алгоритмы выделения сообществ в неориентированном графе, которые позволили выделить творческие коллективы. Распределение числа авторов в выделенных коллективах: 139 коллективов состоит менее чем из 10 членов; 10 коллективов включает от 10 до 20

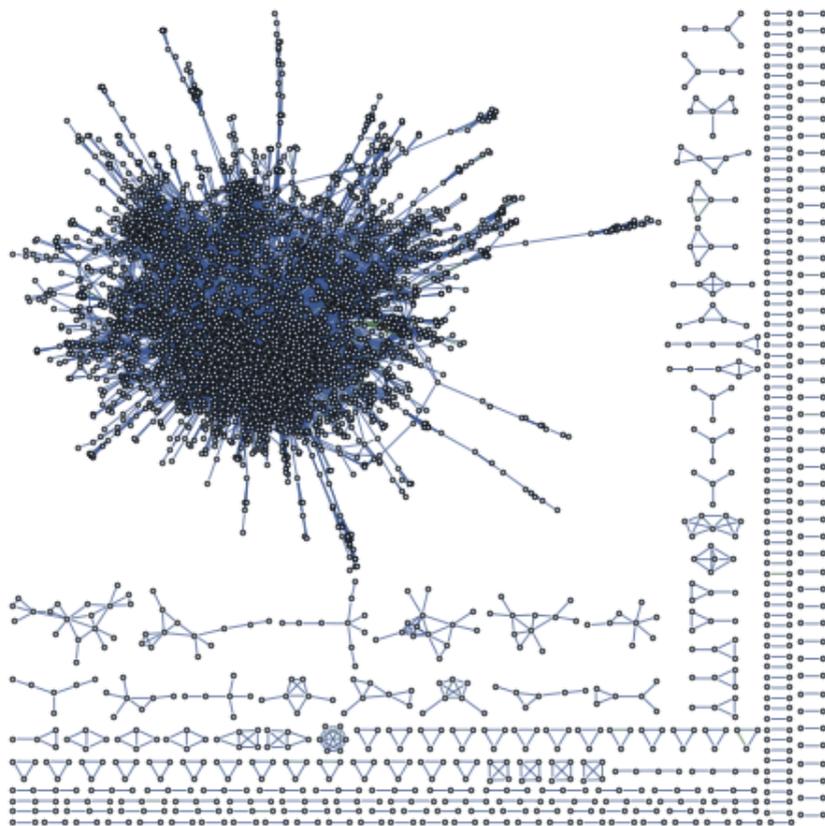


Рис. 3.5. Граф соавторства для публикаций за 5 лет.

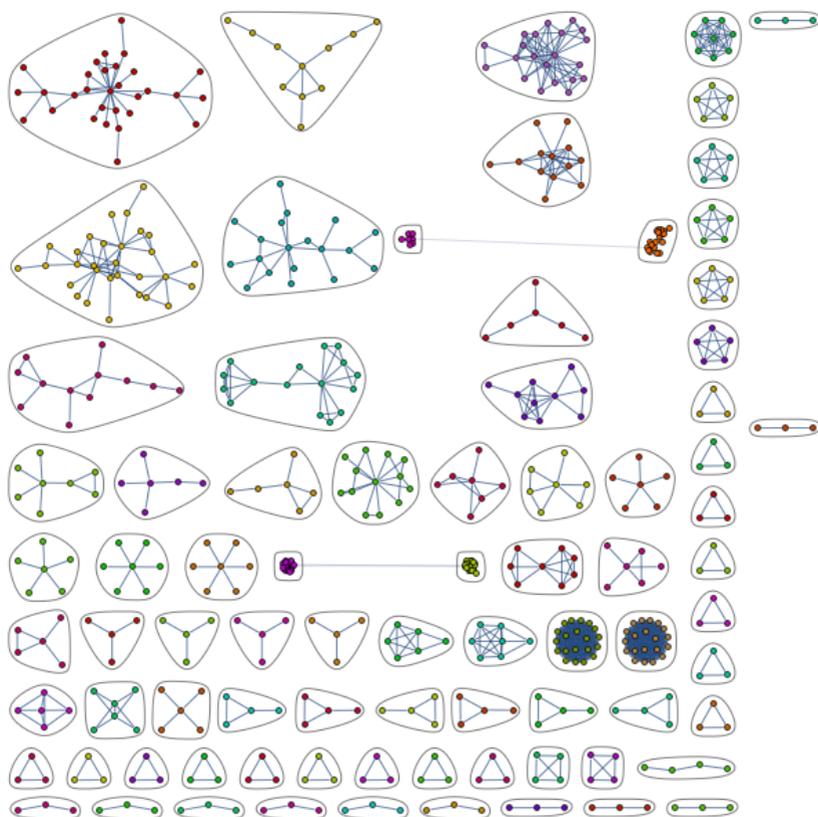


Рис. 3.6. Выделенные коллективы для компонент связности малой мощности.

членов; 6 коллективов включает от 20 до 30 членов; 10 коллективов содержат более 30 членов. На рис. 3.6 и 3.7 представлена структура соавторства для компонент связности малой и большой мощности, соответственно.

Наличие коллективов с большим числом участников может быть связано наличием ошибочных совпадений авторов (однофамильцев с совпадающими инициалами) и наличием в исходных данных сведений о соавторах, работающих в других организациях. В тоже время автор, который был выделен как основной в

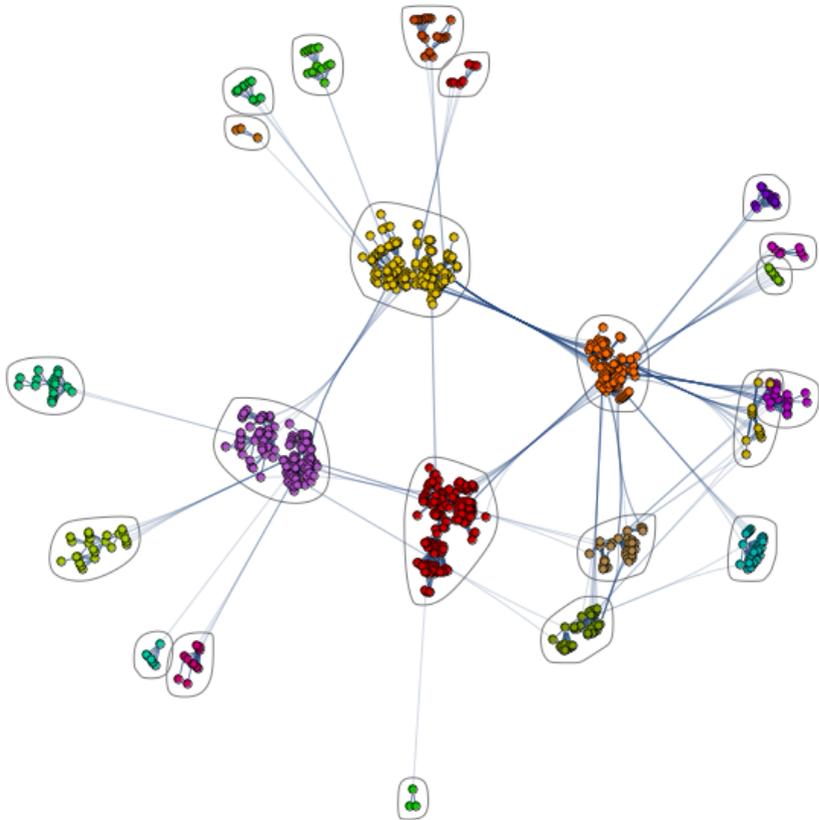


Рис. 3.7. Выделенные коллективы для компоненты связности большой мощности.

коллективе максимальной мощности, является заведующим кафедрой биологической направленности. Это хорошо согласуется с понятием авторского коллектива. Во-первых, заведующий кафедрой публикует большое число работ в соавторстве, а во-вторых, в данной предметной области традиционно публикуются работы в соавторстве. Максимальный диаметр сообщества (то есть длина кратчайшей цепочки соавторов, соединяющей участников коллектива) составляет 10. Значение 2 соответствует ситуации, когда у для каждой пары членов коллектива найдётся общий соавтор («руководитель»).

В целом можно сделать вывод о корректности работы алгоритма выделения творческих коллективов. Однако, для практического использования полученных результатов необходимо провести более детальное исследование.

Глава 4

Внедрение в МГУ

Концепции, модели и алгоритмы, используемые в системе ИСТИНА, разрабатывались коллективом авторов с 2000 года. Написание программного кода Системы началось в январе 2011 года. Уже в феврале была запущена тестовая версия в НИИ механики МГУ. Еще через год, в мае 2012 года, ИАС ИСТИНА была принята к использованию во всех подразделениях Московского университета для учета высокорейтинговых публикаций и расчета размера поощрительных надбавок авторам за эти публикации (раздел 4.1). В настоящее время Система используется в университете также и для следующих целей:

- формирование информационного листа для прохождения сотрудниками процедуры переизбрания на научно-преподавательские должности (раздел 2.4);
- формирование отчетных материалов по мониторингу сети диссертационных советов для Министерства образования и науки в сентябре 2013 года (раздел 2.5.2);
- формирование годовых отчетных материалов по подразделениям и отдельным сотрудникам МГУ (раздел 2.4);
- размещение материалов на официальном сайте диссертационных советов МГУ (раздел 4.2).

4.1. Поощрение авторов высокорейтинговых публикаций

В рамках проекта поощрения сотрудников МГУ за высокорейтинговые публикации рассматривались статьи в журналах, входящих хотя бы в один из следующих списков:

1. топ-25% по импакт-фактору 2010 года (либо 2-летнему, либо 5-летнему) в своей тематической рубрике в Thomson Reuters Science Citation Index или Social Sciences Citation Index;
2. топ-25% по импакт-фактору 2011 года (либо 2-летнему, либо 5-летнему) в своей тематической рубрике в Thomson Reuters Science Citation Index или Social Sciences Citation Index;
3. все журналы из списка Thomson Reuters Arts & Humanities Citation Index¹.

Соответствующие списки журналов (общие и с разделением по рубрикам) были вывешены на веб-сайт Системы. Каждый сотрудник, претендующий на премию, должен был внести информацию о своих высокорейтинговых публикациях 2012 года в Систему. Всего было введено более 500 таких статей. После этого сотрудниками университета вручную была проведена их верификация. На странице со списком высокорейтинговых статей (как общим, так и под авторством отдельного сотрудника) эти сотрудники могли подтвердить (или отменить подтверждение) правильность данных отдельной статьи. По результатам верификации летом 2012 года к поощрению были представлены авторы 388 статей. Для каждого автора w множества A_w высокорейтинговых статей был рассчитан коэффициент надбавки C_w по формуле:

$$C_w = \sum_{a \in A_w} \frac{1}{\sqrt{N_a}},$$

¹<http://istina.msu.ru/journals/index/113514/>

где N_a – количество авторов статьи a .

Отметим, что выбранные критерии для отбора публикаций стали причиной того факта, что подавляющее большинство статей были написаны сотрудниками естественнонаучных факультетов. На рис. 4.1 отображены суммарные коэффициенты подразделений МГУ по итогам проекта, рассчитанные как сумма коэффициентов отдельных сотрудников подразделения.

Распределение суммарных коэффициентов среди подразделений МГУ

На данной странице представлена информация по состоянию на 8 июня 2012 года. Эти данные были использованы при распределении части стимулирующей надбавки за первое полугодие 2012 года. Текущие показатели подразделений, с учетом новых введенных статей и некоторых исправленных неточностей Вы можете посмотреть на [странице статистики МГУ имени М.В. Ломоносова](#).

74.69	Научно-исследовательский институт ядерной физики имени Д.В.Скобельцына
56.25	Химический факультет
51.45	Физический факультет
25.36	Биологический факультет
18.04	НИИ ФХБ имени А.Н. Белозерского
14.70	Государственный астрономический институт имени П.К.Штернберга
6.54	Факультет вычислительной математики и кибернетики
5.88	Механико-математический факультет
4.61	Факультет наук о материалах
3.65	Факультет биоинженерии и биоинформатики
2.14	Геологический факультет
1.89	Научно-исследовательский вычислительный центр
1.88	НИИ механики МГУ
1.17	Институт комплексных исследований образования
1.00	Институт стран Азии и Африки
0.87	Географический факультет
0.37	Факультет фундаментальной медицины
0.35	Международный учебно-научный лазерный центр
0.35	Факультет почвоведения

Рис. 4.1. Распределение суммарных коэффициентов среди подразделений МГУ по итогам летнего премирования сотрудников в 2012 году.

В конце 2012 года проект поощрения авторов высокорейтинговых публикаций вновь был реализован. К поощрению были представлены авторы 820 статей. По результатам выполнения обоих проектов были сформированы следующие отчетные материалы:

- список статей сотрудников МГУ 2012 года в высокорейтинговых журналах¹ (рис. 4.2);

¹<http://istina.msu.ru/statistics/articles/top/>

- список статей сотрудников МГУ 2012 года в высокорейтинговых журналах, распределенных по рубрикам;
- список рубрик с количеством высокорейтинговых журналов в каждой;
- список сотрудников МГУ, имеющих высокорейтинговые статьи 2012 года, с указанием коэффициентов (рис. 4.3);
- список подразделений МГУ, сотрудники которых имеют высокорейтинговые статьи, с указанием суммарных коэффициентов сотрудников подразделения.

Каждый сотрудник имеет доступ к отдельной странице со списком своих высокорейтинговых публикаций.

После успешного выполнения пилотного проекта было принято решение использовать ИСТИНУ на постоянной основе для учета и анализа результатов научной и педагогической деятельности сотрудников университета. В частности, руководством университета было предложено использовать информацию, введенную в Систему, для автоматического формирования информационного листа сотрудника МГУ для аттестации на Ученом совете. Информационный лист представляет собой резюме всех результатов научной и педагогической деятельности сотрудника за 5 лет. Пример информационного листа представлен на рис. 2.16. Он используется при прохождении сотрудником переизбрания на научные и преподавательские должности в университете, которое проходит раз в 5 лет. Таким образом, каждый сотрудник университета по меньшей мере раз в 5 лет должен добавить сведения о результатах своей деятельности в систему ИСТИНА, чтобы успешно пройти процедуру переизбрания. Отметим, что после того, как пользователь добавляет информацию о своих работах в Систему, информационный лист формируется автоматически, и в том числе в формате PDF-файла, сгенерированном с помощью *LaTeX*, что облегчает его распечатку. Распечатав отчет, сотрудник должен подписать его и прикрепить к пакету документов для прохождения процедуры переизбрания.

4.1. ПООЩРЕНИЕ АВТОРОВ ПУБЛИКАЦИЙ

- 1   
2012 Coherent quantum phase slip
Astafiev O.V., Ioffe L.B., Kafanov S., Pashkin Yu A., **Arutyunov Yu K.**, Shahar D., Cohen O., Tsai J.S.
в журнале *Nature*, издательство *Nature Publishing Group (United Kingdom)*, том 484, № 7394, с. 355-358 DOI
- 2   
2012 The complex of tmRNA-SmpB and EF-G on translocating ribosome
Ramrath D.J., Yamamoto H., Rother K., Wittek D., Pech M., Mielke T., Loerke J., Scheerer P., Ivanov P., Teraoka Y., **Shpanchenko O.**, Nierhaus K.H., Spahn C.M.
в журнале *Nature*, издательство *Nature Publishing Group (United Kingdom)*, том 485, № 7399, с. 526-529 DOI
- 3  
2012 366 days: Images of the year
Daniel Cressey, Mark Peplow
в журнале *Nature*, издательство *Nature Publishing Group (United Kingdom)*, № 492, с. 328-333
- 4   
2012 The heliosphere's interstellar interaction: no bow shock
McComas D.J., Alexashov D., Bzowski M., Fahr H., Heerikhuisen J., **Izmodenov V.**, Moebius M.A., Pogorelov N.V., Schwadron N.A., Zank G.P.
в журнале *Science*, издательство *American Association for the Advancement of Science (United States)* DOI
- 5   
2012 A new boson with a mass of 125 GeV observed with the CMS experiment at the Large Hadron Collider
Belyaev A., Boos E., Bunichev V., Demianov A., Dubinin M., Dudko L., Ershov A., Gribushin A., Ilyin V., Katkov I., Kaminsky A., Kiyukhin V., Kodolova O., Korotkiikh V., Kryukov A., Lokhtin I., Markina A., Obraztsov S., Perflöv M., Petrushanko S., Popov A.A., Proskuryakov A., Sarycheva L., Savrin V., Snigirev A., Vardanyan I., CMS Collaboration
в журнале *Science*, издательство *American Association for the Advancement of Science (United States)*, том 338, № 6114, с. 1569-1575 DOI
- 6   
2012 A Massive Particle Consistent with the Standard Model Higgs Boson observed with the ATLAS Detector at the Large Hadron Collider
Giadlilin L.K., Grishkevich Y.V., Kramarenko V.A., Rud V.I., Sivoktokov S.Yu, Smirnova L.N., et al, ATLAS Collaboration
в журнале *Science*, издательство *American Association for the Advancement of Science (United States)*, том 338, № 6114, с. 1576-1582 DOI
- 7   
2012 Ultrafast X-ray pulse characterization at free-electron lasers
Girguras I., Maier A.R., Behrens C., Mazza T., Kelly T.J., Raddiffe P., Duesterer S., Kazansky A.K., **Kabachnik N.M.**, Tschtscherer Th, Costello J.T., Meyer M., Hoffmann M.C., Schlarb H., Cavalleri A.L.
в журнале *Nature Photonics*, издательство *Nature Pub. Group (United Kingdom)*, том 276, с. advanced online publication
- 8   
2012 Universal formation dynamics and noise of Kerr-frequency combs in microresonators
Herr T., Hartinger K., Riemensberger J., Wang C.Y., Gavartin E., Holzwarth R., **Gorodetsky M.L.**, Kippenberg T.J.
в журнале *Nature Photonics*, издательство *Nature Pub. Group (United Kingdom)*, том 6, с. 480-486 DOI
- 9   
2012 Micellar Interpolyelectrolyte Complexes
Pergushov D.V., Müller A.H., Schacher F.H.
в журнале *Chemical Society Reviews*, издательство *Royal Society of Chemistry (United Kingdom)*, том 41, № 21, с. 6888-6901 DOI
- 10   
2012 Quantum Chemistry Behind Bioimaging: Insights from Ab Initio Studies of Fluorescent Proteins and Their Chromophores
Bravaya Ksenia B., **Bella L. Grigorenko, Alexander V Nemukhin, Krylov Anna I.**
в журнале *Accounts of Chemical Research*, издательство *American Chemical Society (United States)*, том 45, № 2, с. 265-275 DOI

Рис. 4.2. Список статей сотрудников МГУ 2012 года в высокорейтинговых журналах (фрагмент).

Абакумов Артем Михайлович	0,61	Химический факультет
Авдеенков Александр Владимирович	1,34	Научно-исследовательский институт ядерной физики имени Д.В. Скобельцына
Аверина Елена Борисовна	0,14	Химический факультет
Агапов Игорь Иванович	0,33	Биологический факультет
Агафонов Иван Николаевич	0,45	Международный учебно-научный лазерный центр
Агол Вадим Израилевич	1,11	Научно-исследовательский институт физико-химической биологии им. А.Н.Белозерского
Азовский Андрей Игоревич	0,64	Биологический факультет
Аккуратов Евгений Евгеньевич	0,11	Международный биотехнологический центр
Аксянов Евгений Алексеевич	0,16	Научно-исследовательский институт физико-химической биологии им. А.Н.Белозерского
Алексеев Станислав Олегович	0,25	Государственный астрономический институт имени П.К.Штернберга
Алексеевский Андрей Владимирович	0,3	Научно-исследовательский институт физико-химической биологии им. А.Н.Белозерского, Факультет биоинженерии и биоинформатики
Алиев Рамиз Автандилович	0,12	Научно-исследовательский институт ядерной физики имени Д.В. Скобельцына
Алиева Ирина Борисовна	0,15	Научно-исследовательский институт физико-химической биологии им. А.Н.Белозерского
Алешин Владимир Вениаминович	0,14	Научно-исследовательский институт физико-химической биологии им. А.Н.Белозерского
Амитонов Сергей Владимирович	0,11	Физический факультет
Амочкина Татьяна Владимировна	0,14	Научно-исследовательский вычислительный центр
Андреев Дмитрий Евгеньевич	0,12	Научно-исследовательский институт физико-химической биологии им. А.Н.Белозерского
Андреева Ирина Петровна	0,11	Химический факультет
Анохин Денис Валентинович	0,39	Факультет фундаментальной физико-химической инженерии
Антипов Евгений Викторович	0,9	Химический факультет, Химический факультет
Антоненко Юрий Николаевич	0,3	Научно-исследовательский институт физико-химической биологии им. А.Н.Белозерского
Антонова Елизавета Евгеньевна	0,11	Научно-исследовательский институт ядерной физики имени Д.В. Скобельцына
Апями Владимир Владимирович	0,25	Химический факультет
Аржанцев Иван Владимирович	0,5	Механико-математический факультет
Артамонов Борис Павлович	0,07	Государственный астрономический институт имени П.К.Штернберга
Арутюнов Константин Юрьевич	0,33	Научно-исследовательский институт ядерной физики имени Д.В. Скобельцына
Арутюнян Александр Мигранович	0,4	Научно-исследовательский институт физико-химической биологии им. А.Н.Белозерского
Асмолов Евгений Савельевич	0,5	Научно-исследовательский институт механики
Аэров Артём Анатольевич	0,33	Физический факультет
Баханова Зульфия Галеевна	0,33	Научно-исследовательский вычислительный центр

Рис. 4.3. Список сотрудников МГУ, имеющих высокорейтинговые статьи 2012 года, с указанием коэффициентов (фрагмент).

4.2. Официальный сайт диссертационных советов

В соответствии с приказом №326 Министерства образования и науки РФ «Об утверждении Порядка размещения в информационно-телекоммуникационной сети «Интернет» информации, необходимой для обеспечения порядка присуждения ученых степеней» в Московском университете в рамках проекта ИСТИНА был разработан модуль – Официальный сайт диссертационных советов МГУ¹. Отметим, что основные сущности, необходимые для работы модуля («пользователь», «диссертационный совет», «диссертация», «прикрепляемый файл»), уже присутствовали и активно использовались в Системе ранее. Этот факт существенно облегчил разработку сайта под широкий перечень функциональных возможностей, которые по некоторым позициям перекрывают положения упомянутого приказа Минобрнауки РФ. Официальный сайт диссертационных советов МГУ включает:

- страницу со списком всех действующих (утвержденных ВАК) диссертационных советов (рис. 4.4);
- страницы диссертационных советов, содержащие списки их членов, список специальностей, контактную информацию, список защищенных диссертаций и диссертаций, принятых к рассмотрению и защите (рис. 4.5);
- интерфейс редактирования информации о диссертационном совете для уполномоченных на это сотрудников.

Такой подход позволяет, с одной стороны, иметь репрезентативный (полный) объем информации, которая необходима для оперативного мониторинга результатов деятельности диссертационных советов при МГУ, с другой – расширять набор функций и сервисов Официального сайта диссертационных советов МГУ.

¹http://istina.msu.ru/dissertation_councils/by_organization/214524



Диссертационные советы МГУ имени М.В. Ломоносова

1. Д 501.001.01	26. Д 501.001.31	51. Д 501.001.60	76. Д 501.001.86	101. Д 501.002.15
2. Д 501.001.02	27. Д 501.001.32	52. Д 501.001.61	77. Д 501.001.88	102. Д 501.002.16
3. Д 501.001.03	28. Д 501.001.33	53. Д 501.001.62	78. Д 501.001.89	103. Д 501.002.18
4. Д 501.001.04	29. Д 501.001.34	54. Д 501.001.63	79. Д 501.001.90	
5. Д 501.001.07	30. Д 501.001.35	55. Д 501.001.64	80. Д 501.001.91	
6. Д 501.001.08	31. Д 501.001.36	56. Д 501.001.65	81. Д 501.001.93	
7. Д 501.001.09	32. Д 501.001.37	57. Д 501.001.66	82. Д 501.001.94	
8. Д 501.001.10	33. Д 501.001.38	58. Д 501.001.67	83. Д 501.001.95	
9. Д 501.001.11	34. Д 501.001.39	59. Д 501.001.68	84. Д 501.001.96	
10. Д 501.001.13	35. Д 501.001.40	60. Д 501.001.69	85. Д 501.001.97	
11. Д 501.001.14	36. Д 501.001.41	61. Д 501.001.70	86. Д 501.001.98	
12. Д 501.001.15	37. Д 501.001.42	62. Д 501.001.71	87. Д 501.001.99	
13. Д 501.001.16	38. Д 501.001.43	63. Д 501.001.72	88. Д 501.002.01	
14. Д 501.001.17	39. Д 501.001.44	64. Д 501.001.73	89. Д 501.002.02	
15. Д 501.001.18	40. Д 501.001.45	65. Д 501.001.74	90. Д 501.002.03	
16. Д 501.001.19	41. Д 501.001.46	66. Д 501.001.76	91. Д 501.002.04	
17. Д 501.001.20	42. Д 501.001.47	67. Д 501.001.77	92. Д 501.002.05	
18. Д 501.001.21	43. Д 501.001.49	68. Д 501.001.78	93. Д 501.002.06	
19. Д 501.001.22	44. Д 501.001.50	69. Д 501.001.79	94. Д 501.002.07	
20. Д 501.001.23	45. Д 501.001.51	70. Д 501.001.80	95. Д 501.002.09	
21. Д 501.001.24	46. Д 501.001.52	71. Д 501.001.81	96. Д 501.002.10	
22. Д 501.001.25	47. Д 501.001.53	72. Д 501.001.82	97. Д 501.002.11	
23. Д 501.001.26	48. Д 501.001.55	73. Д 501.001.83	98. Д 501.002.12	
24. Д 501.001.27	49. Д 501.001.57	74. Д 501.001.84	99. Д 501.002.13	
25. Д 501.001.30	50. Д 501.001.59	75. Д 501.001.85	100. Д 501.002.14	

Рис. 4.4. Список диссертационных советов МГУ.

4.3. Механизм обратной связи

Отметим, что основной целевой аудиторией Системы являются опытные научные сотрудники, имеющие большое количество публикаций и других результатов деятельности. У таких пользователей часто возникают различные проблемы с использованием современных информационных систем, в частности, имеющих веб-интерфейс. В связи с этим обстоятельством важнейшими при внедрении Системы стали вопросы обеспечения механизма обратной связи, поддержки пользователей, а также – разработки дружелюбного и интуитивно-понятного пользовательского интерфейса. В частности, на страницах регистрации и входа в Систему указана ссылка на отдельную страницу помощи при возникающих у пользователя вопросах, связанных с регистрацией и входом в Систему.

В системе ИСТИНА присутствует функция обратной связи, которая позволяет каждому пользователю отправить сообщение ее раз-

Диссертационный совет Д 501.002.16 при МГУ имени М.В. Ломоносова

Номер диссертационного совета: Д 501.002.16

Организация: МГУ имени М.В. Ломоносова

Действующие члены совета:

Садовничий Виктор Антонович, д.ф.-м.н., проф., акад. РАН - председатель
Алексеев Валерий Борисович, д.ф.-м.н., проф. - зам. председателя
Васенин Валерий Александрович, д.ф.-м.н., проф. - зам. председателя
Кобельков Георгий Михайлович, д.ф.-м.н., проф. - зам. председателя
Корнев Андрей Алексеевич, д.ф.-м.н., проф. - ученый секретарь
Алешин Станислав Владимирович, д.ф.-м.н., проф.
Гацков Сергей Борисович, д.ф.-м.н., проф.
Денисов Александр Михайлович, д.ф.-м.н., проф.
Захаров Евгений Владимирович, д.ф.-м.н., проф.
Кумсков Михаил Иванович, д.ф.-м.н., проф.
Латышев Виктор Николаевич, д.ф.-м.н., проф.
Ложкин Сергей Андреевич, д.ф.-м.н., проф.
Михалёв Александр Васильевич, д.ф.-м.н., проф.
Нестеренко Юрий Валентинович, д.ф.-м.н., проф., член-корр. РАН
Новиков Борис Асенович, д.ф.-м.н., проф.
Ольшанский Максим Александрович, д.ф.-м.н.
Пальцев Борис Васильевич, д.ф.-м.н., проф.
Подловченко Римма Ивановна, д.ф.-м.н., проф.
Редькин Николай Петрович, д.ф.-м.н., проф.
Тимонина Елена Евгеньевна, д.т.н., проф.
Фомичев Владимир Михайлович, д.ф.-м.н., проф.
Четверушкин Борис Николаевич, д.ф.-м.н., проф., акад. РАН
Чечкин Александр Витальевич, д.ф.-м.н., проф.
Чижиков Евгений Владимирович, д.ф.-м.н., проф.

Адрес совета:

РФ, 119991, Москва, ГСП-1, Ленинские горы, д. 1, ДК МГУ, комн. 12-24

Адрес места заседаний:

Москва, Ленинские горы, д. 1, ДК МГУ, ауд. 14-08

Телефон совета:

(495) 939-14-70

Электронная почта email:

texmat.disser16@mail.ru

Список специальностей:

01.01.07 - Вычислительная математика (физ.-мат. науки)
05.13.17 - Теоретические основы информатики (физ.-мат. науки)
05.13.19 - Методы и системы защиты информации, информационная
безопасность (физ.-мат. науки)

Прикрепленные файлы [редактировать](#)

Работа с диссертационным советом

[Редактировать совет](#)
[Прикрепить файл](#)
[Удалить совет](#)

Рис. 4.5. Страница диссертационного совета в системе ИСТИНА.

работчикам со своими замечаниями, пожеланиями и предложениями. Необходимо отметить, что при внедрении Системы в МГУ оказалось, что сил разработчиков Системы недостаточно для обработки большого количества запросов пользователей и оперативных ответов на них. Тогда было принято решение создать сеть ответственных сотрудников, прикрепленных к подразделениям университета, которые будут обеспечивать ответы на вопросы сотрудников своего подразделения и, при необходимости, консультировать их по телефону, с использованием электронной почты и в индивидуальном порядке. Такая сеть была создана, и в нее вошли 93 сотрудника. Ответственным сотрудникам были разосланы инструкции по работе с пользователями, а также номера телефонов разработчиков. Взаимодействие между ответственными в подразделениях и разработчиками Системы в настоящее время осуществляется на постоянной основе.

У каждого сотрудника на странице обратной связи в Системе перечислены ответственные сотрудники по его подразделению, с указанием адреса электронной почты, телефона и иногда рабочего адреса (рис. 4.6). При необходимости, сотрудник может посмотреть и полный список ответственных в организации. Если по каким-то причинам ответственный по подразделению не отвечает на запрос сотрудника (например, он заболел или ушел в отпуск), то сотрудник может обратиться к ответственному по всему университету, а уже после него – напрямую к разработчикам. Таким образом, рассматриваемый механизм позволил снизить нагрузку по обеспечению поддержки пользователей, которая ранее ложилась только на разработчиков Системы.

Несмотря на механизм поддержки, разработчики продолжают получать большое число писем и звонков от ответственных. Для его уменьшения в Системе разрабатываются средства, позволяющие ответственным самим осуществлять различные операции по администрированию Системы, например, объединять дубликаты журналов, сотрудников, удалять лишние данные и другие действия. Предполагается, что в дальнейшем администрирование Системы будет осуществляться руками ответственных, а разработчики Системы смогут сосредоточиться на развитии математического

Обратная связь с разработчиками системы

Если у вас (или у вашего коллеги) возникает проблема с регистрацией или входом в систему, то прежде чем обращаться к ответственным, пожалуйста, следуйте инструкциям на странице [Проблемы с регистрацией или входом в систему](#).

В связи с большим количеством сообщений от пользователей механизм общения с разработчиками системы был изменен. Если при работе с системой у вас возникают какие-либо вопросы, проблемы или предложения, **свяжитесь с ответственным сотрудником по вашему подразделению**. Список ответственных, приведенный ниже, сформирован на основании приказа № 921 от 13 сентября 2012 года "Об Информационно-аналитической системе "Наука-МГУ". Что делать, если при взаимодействии с ответственными возникают проблемы, см. ниже под таблицей.

Список активных ответственных сотрудников в вашем подразделении

Ниже представлен список *активных* ответственных сотрудников по вашему подразделению. Полный список ответственных по всем подразделениям МГУ [доступен по ссылке](#).

№	Подразделение	Ответственный	Должность	Email	Телефон	Адрес	Примечание
1	Научно-исследовательский институт механики	 Рязанцева Марина Юрьевна	учёный секретарь	mapna-ryazantseva@yandex.ru	939-5654		
2	Научно-исследовательский институт механики		Научный отдел		939-5654		

Если ответственный не может решить проблему

Попросите его связаться с разработчиками, и мы ответим ему, а уже он ответит вам.

Если в списке нет ответственного по вашему подразделению

Обратитесь в научный отдел своего подразделения или к ответственному по всей организации из списка ниже. Сообщите ему ваше подразделение и отметьте, что по нему нет ответственного. Также опишите проблему, по которой вы хотели к нему обратиться.

Если ответственный не отвечает (заболел / ушел в отпуск)

Обратитесь в научный отдел своего подразделения или к ответственному по всей организации из списка ниже.

Список ответственных сотрудников по всей организации

Внимание! Пожалуйста, обращайтесь к ответственному из этого списка только если возникают проблемы при взаимодействии с ответственными по вашему подразделению (или если таких ответственных нет).

№	Подразделение	Ответственный	Должность	Email	Телефон	Адрес	Примечание
1	МГУ имени М.В. Ломоносова	 Чертовин Александр Викторович	заместитель начальника Управления инновационной политики и международных научных связей МГУ	chertov@rector.msu.ru	939-1013	ГЗ, к. 901	

Рис. 4.6. Страница обратной связи в системе ИСТИНА.

обеспечения, программного кода и исправлении ошибок и недочетов, которые объективно обнаруживаются в процессе эксплуатации Системы.

4.4. Общая статистика внедрения в МГУ

С мая 2012 года проводится внедрение системы ИСТИНА в МГУ имени М.В. Ломоносова. Веб-сайт Системы доступен по адресу <http://istina.msu.ru>. Система в настоящее время активно используется для анализа деятельности сотрудников МГУ и его структурных подразделений, а также в двух формальных процедурах в рамках университета, а именно - для распределения поощрительных надбавок сотрудникам за публикации в высокорейтинговых журналах, а также для формирования информационных листов для переизбрания сотрудников на научно-преподавательские должности. В Системе зарегистрировано более 10 тыс. сотрудников университета из 880 подразделений, которыми добавлено более 300 тыс. статей, 29 тыс. книг, 57 тыс. докладов на конференциях, 3.9 тыс. патентов, 3.1 тыс. научных отчетов, 3 тыс. членств в редколлегиях журналов, 4.4 тыс. членств в редколлегиях сборников, 6.6 тыс. членств в программных комитетах конференций, 2.7 тыс. членств в диссертационных советах, 13 тыс. диссертаций, защищенных в 2.6 тыс. диссертационных советах, 24 тыс. дипломных работ и 18 тыс. учебных курсов.

На следующих иллюстрациях отображена статистика посещений веб-сайта ИСТИНА по данным Google Analytics: количество посещений в неделю (рис. 4.7); количество уникальных посетителей в неделю (рис. 4.8); общая статистика сайта (рис. 4.9); распределение посетителей по городам (рис. 4.10), браузерам (рис. 4.11) и операционным системам (рис. 4.12).

4.4. ОБЩАЯ СТАТИСТИКА ВНЕДРЕНИЯ В МГУ



Рис. 4.7. Количество посещения веб-сайта Системы в неделю.



Рис. 4.8. Количество уникальных посетителей веб-сайта Системы в неделю.



Рис. 4.9. Общая статистика веб-сайта Системы по данным Google Analytics.

Город	Сеансы	Сеансы, %
1. Moscow	1 108 177	46,25 %
2. Saint Petersburg	97 571	4,07 %
3. (not set)	67 034	2,80 %
4. Kyiv	47 207	1,97 %
5. Novosibirsk	31 632	1,32 %
6. Yekaterinburg	29 039	1,21 %
7. Samara	23 655	0,99 %
8. Kazan	22 567	0,94 %
9. Minsk	22 210	0,93 %
10. Nizhny Novgorod	21 092	0,88 %

Рис. 4.10. Распределение посетителей веб-сайта Системы по городам.

Браузер	Сеансы	Сеансы, %
1. Chrome	926 170	38,66 %
2. Firefox	542 989	22,66 %
3. Internet Explorer	324 879	13,56 %
4. Opera	317 374	13,25 %
5. Safari	152 785	6,38 %
6. YaBrowser	60 916	2,54 %
7. Android Browser	29 151	1,22 %
8. Opera Mini	15 205	0,63 %
9. MRCHROME	6 483	0,27 %
10. Safari (in-app)	4 254	0,18 %

Рис. 4.11. Распределение посетителей веб-сайта Системы по браузерам.

Операционная система	Сеансы	Сеансы, %
1. Windows	2 061 182	86,03 %
2. iOS	114 414	4,78 %
3. Macintosh	82 789	3,46 %
4. Android	64 302	2,68 %
5. Linux	59 829	2,50 %
6. (not set)	8 792	0,37 %
7. Windows Phone	2 185	0,09 %
8. SymbianOS	615	0,03 %
9. Series40	596	0,02 %
10. FreeBSD	205	0,01 %

Рис. 4.12. Распределение посетителей веб-сайта Системы по операционным системам.

Глава 5

Информационная безопасность и разграничение доступа

5.1. Модели логического разграничения доступа

Система ИСТИНА относится к классу систем с большими базами данных научно-технического содержания. Подобные системы имеют следующие особенности:

1. большое число пользователей, зарегистрированных в системе (несколько десятков тысяч);
2. большое число записей, касающихся содержательной части системы, в том числе информация о результатах научной деятельности (несколько сотен тысяч записей), структуре организаций (несколько тысяч записей) и другие данные;
3. большое число часто изменяющихся связей между объектами системы (несколько сотен тысяч), как простых, так и сложных.

Отмеченные особенности предъявляют особые требования к моделям и методам разграничения доступа к ресурсам таких систем, а

также к их программной реализации. Эти требования отличаются от традиционно используемых и на настоящее время способы их практической реализации до конца не разработаны. Далее будет показано, как задача разграничения доступа решена на настоящее время и каким образом планируется развивать такие модели и их программные механизмы в дальнейшем, чтобы сделать Систему для пользователей более удобной и безопасной.

5.1.1. Модели, действующие в системе в настоящее время

На настоящее время в системе ИСТИНА реализована комбинация дискреционной (англ. discretionary access control, DAC) и ролевой (англ. Role Based Access Control, RBAC) моделей логического разграничения доступа (ЛРД), каждая из которых имеет свои особенности.

Дискреционная модель

Дискреционная модель ЛРД (рис. 5.1) основана на концепции «таблицы доступов» - специальной таблицы, в которой явно отмечено, какие действия может выполнять конкретный субъект (например, пользователь системы или процесс) с конкретным объектом (например, статья или материалы научной конференции). Преимуществом такой модели является возможность задания индивидуальных («точечных») правил доступа к объектам системы. Однако, в чистом виде такая модель неприменима в системах с большими базами научно технического содержания. Причина в том, что таблицы доступа в таком случае становятся исключительно громоздкими и нечитаемыми, что влечёт за собой сложности в задании новых правил, не противоречащих существующим, а также приводит к экспоненциально быстрому падению производительности Системы при обработке запросов на доступ к её ресурсам.

В системе ИСТИНА дискреционный доступ применяется, например, в случаях, когда необходимо выдать право редактирования толь-

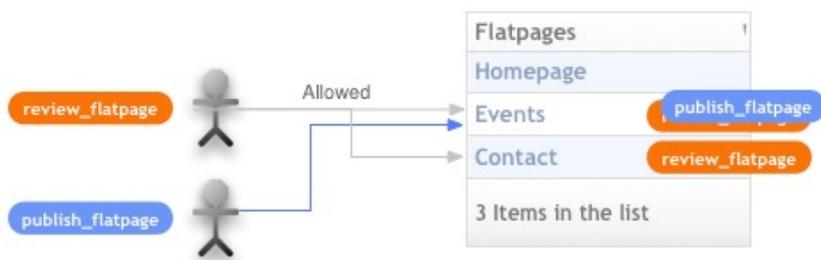


Рис. 5.1. Дискреционная модель ЛРД.

ко конкретной статьи некоторому конкретному пользователю.

Ролевая модель

Управление доступом на основе ролей (рис. 5.2) представляет собой развитие дискреционной модели ЛРД. Главная особенность ролевой модели заключается в том, что права доступа субъектов доступа к объектам группируются с учётом специфики их применения, образуя роли. Эта модель достаточно гибкая и эффективная для работы в системах с большим числом субъектов и объектов. К её преимуществам можно отнести то обстоятельство, что при таком подходе удобно создавать и сопровождать правила разграничения доступа, поскольку с помощью группировки пользователей по ролям можно свести число правил к минимуму. Однако, у неё есть существенный недостаток. Его суть в том, что в системах с большими базами данных научно-технического содержания, к которым относится ИСТИНА (сотни ролей, тысячи пользователей и миллионы разрешений), управление ролями, пользователями, разрешениями и их взаимосвязями является сложной задачей. Решить её практически невозможно малой группой администраторов безопасности.

В системе ИСТИНА на настоящее время существуют следующие основные роли:

- пользователь неавторизованный;

RBAC: Role Based Access Control

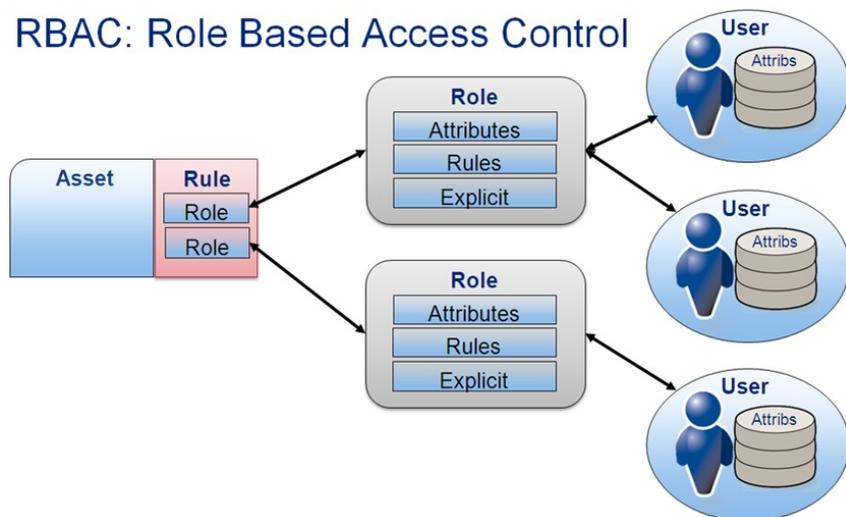


Рис. 5.2. Ролевая модель ЛРД.

- пользователь авторизованный, в том числе – «Сотрудник библиотеки», «Сотрудник ректората», «Ответственный по подразделению», «Редактор ответственных».

5.1.2. Реляционная модель

При разработке программных механизмов логического разграничения доступа субъектов к объектам системы ИСТИНА следует учитывать ряд важных свойств организации данных в этой системе. Одно из них заключается в том, что Система допускает регистрацию новых пользователей без участия администраторов. При добавлении в Систему каждого нового пользователя у её объектов появляется новый атрибут – тип взаимосвязи с новым пользователем. Этот атрибут может существенно влиять на предоставление пользователям Системы доступа к этому объекту. В связи с этим обстоятельством, в Системе нельзя выделить статическое множество типов объектов, которые бы однозначно определяли возможность доступа к объектам каждого пользователя, как это делается при использовании моделей

логического разграничения доступа (ЛРД) в режиме принудительной типизации и ролевых моделей.

Следует заметить, что информационные системы, в которых постоянно меняется очень большое число пользователей, в настоящее время достаточно широко распространены. К их числу относятся общедоступные службы электронной почты, блог-сервисы и форумы. Подобные системы объединяют с ИАС «ИСТИНА» следующие свойства:

- постоянно меняется множество пользователей системы;
- у администратора системы нет возможности при добавлении новых пользователей оперативно обновлять политику безопасности или изменять метки безопасности объектов;
- с каждым объектом системы может быть ассоциировано несколько категорий пользователей, обладающих различными правами доступа к этому объекту.

С учетом изложенного, возникает необходимость разработки модели ЛРД, ориентированной на системы с постоянно изменяющимся множеством пользователей. Такая модель ЛРД должна предусматривать механизмы автоматического, без участия администратора, назначения прав доступа к объектам для новых пользователей системы. На настоящее время разработана и проходит тестовые испытания модель ЛРД, обладающая указанными свойствами. Предлагается именовать новую модель ЛРД реляционной от английского слова Relation – отношение, так как это понятие является в разработанной модели одним из центральных.

Далее представлены основные сведения об этой модели, программной реализации её первой версии и о результатах тестовых испытаний.

5.1.3. Основные множества и сущности реляционной модели

Определение 1. Пусть в Системе заданы следующие множества:

- *Objects* – множество объектов;
- *Users* \subset *Objects* – множество пользователей;
- *Classes* – множество классов объектов;
- *Types* – множество типов объектов;
- *Actions* – множество видов доступа;
- *Relations* – множество отношений;
- *Roles* – множество ролей;
- *Privileges* – множество привилегий;
- *Sessions* – множество сеансов;

следующие далее отношения:

- *ObjectRelations* \subset *Objects* \times *Relations* \times *Objects*, определяющее, какие объекты связаны определенным отношением;
- *UserRoles* \subset *Users* \times *Roles*, определяющее роли каждого пользователя;
- *RolePrivileges* \subset *Roles* \times *Privileges*, определяющее, какие привилегии дает пользователю принадлежность к определенной роли;
- *RelationTypeAccessGranted* \subset *Relations* \times *Types* \times *Actions*, определяющее, какие действия может совершать с объектом пользователь в зависимости от отношения этого пользователя к этому объекту;
- *RelationTypeAccessDenied* \subset *Relations* \times *Types* \times *Actions*, определяющее, какие действия пользователь не может совершать с объектом в зависимости от отношения этого пользователя к этому объекту;
- *RoleTypeAccessGranted* \subset *Roles* \times *Types* \times *Actions*, определяющее, какие действия с объектами определенного типа может совершать пользователь, обладающий определенной ролью;

- $RoleTypeAccessDenied \subset Roles \times Types \times Actions$, определяющее, какие действия с объектами определенного типа пользователь, обладающий определенной ролью, совершать не может;
- $ClassAccess \subset C \times A$, определяющее, какие виды доступа разрешены для объектов каждого класса;
- $\leq C Roles \times Roles$ – иерархия ролей;
- $\leq C Relations \times Relations$ – иерархия отношений;
- $\leq C Classes \times Classes$, означающее, что один класс является подклассом другого;
- $RelationLeftClasses \subset Relations \times Classes$, определяющее, объекты каких классов могут быть связаны определенным отношением с другими объектами;
- $RelationRightClasses \subset Relations \times Classes$, используемое аналогично $RelationLeftClasses$;
- $RelationTransitions \subset Relations \times Relations \times Relations$ – правила переноса отношений;

а также следующие отображения:

- $ObjectClass : Objects \rightarrow Classes$, определяющее класс каждого объекта;
- $ObjectType : Objects \rightarrow Types$, определяющее тип каждого объекта;
- $SessionUser : Sessions \rightarrow Users$, определяющее принадлежность сеанса доступа в Системе определенному пользователю;
- $RelationType : Relations \rightarrow \{One - to - One, One - to - Many, Many - to - One, Many - to - Many\}$.

Пусть при этом выполняются следующие условия:

- множество *Classes* содержит такой выделенный элемент *user*, что $o \in Users \leftrightarrow ObjectClass(o) \leq user$, что означает, что все пользователи Системы являются объектами класса *user* или одного из его подклассов;
- множество *Relations* содержит выделенный элемент *user_created_object*, означающий, что данный пользователь является создателем данного объекта.

Тогда будем считать, что в Системе задана реляционная модель ЛРД.

Здесь множество *Objects* – это множество всех объектов Системы, *Users* – множество ее пользователей.

Множество *Classes* содержит все возможные виды «сущности» объектов. Например, в Системе можно выделить следующие классы объектов:

1. пользователь;
2. сотрудник;
3. подразделение;
4. результат научно-исследовательской деятельности;
5. публикация;
6. доклад на конференции;
7. НИР;
8. патент;
9. свидетельство о регистрации прав на программное обеспечение;
10. отчет;
11. привлечение инвестиции;

12. членство в редколлегии журнала;
13. членство в редколлегии сборника;
14. членство в программном комитете конференции;
15. членство в диссертационном совете;
16. диссертация;
17. руководство диссертацией;
18. руководство дипломной работой;
19. авторство учебного курса;
20. преподавание учебного курса;
21. награда;
22. почетное членство в организации;
23. членство в научном коллективе;
24. стажировка;
25. журнал;
26. конференция;
27. диссертационный совет.

При этом множество классов имеет иерархическую структуру, то есть одни классы могут являться потомками других. В представленном перечне классы 5 – 24 являются потомками класса «результат научно-исследовательской деятельности».

Множество типов *Ttypes* содержит возможные дополнительные метки объекта, влияющие на права доступа пользователей к нему. В системе ИСТИНА можно предложить несколько возможных способов назначения объектам таких дополнительных меток. Один из

них – это присваивание каждому зарегистрированному в Системе сотруднику такого атрибута, как его разрешение на публикацию в открытом доступе его персональных данных. Подробно использование этого атрибута рассматривается в 5.1.4. В других информационных системах, для которых может применяться реляционная модель ЛРД, множество типов объектов возможно использовать в других целях. Например, в блог-сервисах каждой записи может быть дополнительно приписан флаг полного запрета комментариев.

Множество *Actions* содержит все возможные действия с объектами. Все виды доступа к объектам Системы перечислены в 5.1.4.

Множество *Relations* содержит возможные виды взаимосвязи между объектами Системы. Например, в Системе можно выделить следующие отношения:

1. пользователь связан с данным сотрудником;
2. сотрудник работает в данном подразделении;
3. сотрудник предположительно является автором данной работы;
4. сотрудник подтвердил, что является автором данной работы;
5. статья опубликована данном журнале;
6. статья опубликована данном сборнике;
7. доклад сделан на данной конференции;
8. диссертация защищена на данном совете;
9. одно подразделение входит в другое подразделение;
10. пользователь является ответственным по данному подразделению;
11. пользователь является редактором ответственных по данному подразделению;

5.1. МОДЕЛИ ЛОГИЧЕСКОГО РАЗГРАНИЧЕНИЯ ДОСТУПА

12. пользователь является ответственным по подразделению, в котором числится данный сотрудник;
13. пользователь является редактором ответственных по подразделению, в котором числится данный сотрудник;
14. пользователь является ответственным по подразделению, в которое входит в качестве дочернего данное подразделение;
15. пользователь является редактором ответственных по подразделению, в которое входит в качестве дочернего данное подразделение;
16. пользователь является ответственным по подразделению, в котором числится хотя бы один из авторов данной работы;
17. пользователь является редактором ответственных по подразделению, в котором числится хотя бы один из авторов данной работы;
18. сотрудник является членом диссертационного совета, в котором защищена данная диссертация;
19. сотрудник является членом редколлегии журнала, в котором опубликована данная статья.

Ответственные по подразделению – это пользователи, имеющие право изменять информацию о сотрудниках соответствующего подразделения. Редакторы ответственных по подразделению – это пользователи, которые имеют право изменять список ответственных, однако при этом сами могут и не иметь полномочий, которые имеют ответственные.

Множество *Roles* – это множество ролей пользователя. В системе ИСТИНА в настоящее время предусмотрены следующие роли:

- незарегистрированные пользователи,

- зарегистрированные пользователи,
- суперпользователи.

Множество *Privileges* – это множество возможных действий в Системе, которые не относятся к конкретному объекту. Примером привилегии является возможность создать объект определенного класса, а также осуществлять другие операции с Системой, не относящиеся к конкретному объекту – например, просмотр списка всех пользователей Системы. В целях увеличения быстродействия операций проверки доступа, к привилегиям целесообразно также отнести возможность совершения операций, затрагивающих все объекты определенного класса, например, отправки сообщения всем пользователям Системы.

Множество *Sessions* – содержит сеансы связи пользователя с Системой. Поскольку пользователи работают с Системой удаленно, необходимо отслеживать такие события, как вход в неё с необычного IP-адреса. Например, в случае, если пользователь одновременно зарегистрировался в Системе с двух разных IP-адресов, возможно, следует отправить на его электронный адрес предупреждение об этом с предложением изменить пароль. Более подробно методы отслеживания подозрительной активности пользователей Системы рассмотрены далее.

Множество *ObjectRelations* определяет, какие объекты связаны определенным отношением. Если $o_1, o_2 \in Objects$, $r \in Relations$ и $(o_1, r, o_2 \in ObjectRelations$, то будем считать, что объекты o_1 и o_2 связаны отношением r .

Множество *UserRoles* определяет, какими ролями обладает каждый из пользователей.

Отношение *RolePrivileges* связывает роли Системы с привилегиями.

Множество *RelationTypeAccessGranted* определяет, какие действия над объектом может выполнять пользователь, обладающий определенным отношением к этому объекту. Если объект o имеет тип t , u – пользователь, $(u, r, o) \in ObjectRelations$ и

$(r, t, a) \in RelationTypeAccessGranted$, то пользователь u имеет право осуществлять доступ a к объекту o .

Множество $RelationTypeAccessDenied$ определяет, какие действия над объектом пользователь, обладающий определенным отношением к этому объекту, выполнять не может даже в случае, если они ему разрешены одним из правил множества $RelationTypeAccessGranted$.

Отношение $RoleTypeAccessGranted$ используется аналогично $RelationTypeAccessGranted$, но определяет доступ пользователя к объекту в зависимости от его роли в Системе, а не отношения к объекту. Отношение $RoleTypeAccessDenied$ определяет, какие доступы к объектам запрещены пользователям, обладающим определенной ролью, даже если они разрешены другим ролям этого пользователя. Один из способов использования этого отношения рассматривается в настоящем подразделе далее.

Отношение $ClassAccess$ связывает каждый из классов с видами доступа, которые допускаются (разрешены) для объектов этого класса.

Отношения $RelationLeftClasses$ и $RelationRightClasses$ определяют, объекты каких классов могут находиться в заданном отношении с другими объектами. Если $(o_1, r, o_2) \in ObjectRelations$, объект o_1 имеет класс c_1 , а объект o_2 имеет класс c_2 , то $(r, c_1) \in RelationLeftClasses$ и $(r, c_2) \in RelationRightClasses$

Множество $RelationTransitions$ используется для автоматического переноса отношений пользователей к объектам на другие объекты по определенным правилам. Например, чтобы пользователь, ответственный по подразделению, в котором работает один из авторов определенной публикации, автоматически получал определенные права по отношению к этой публикации. Если o_1, o_2 и o_3 – объекты (при этом, как правило, u_1 – пользователь), $(r_{12}, r_{23}, r_{13}) \in RelationTransitions$ и при этом $(o_1, o_2, r_{12}) \in ObjectRelations$, $(o_2, o_3, r_{23}) \in ObjectRelations$, то автоматически считается, что $(o_1, o_3, r_{13}) \in ObjectRelations$. В частности, если o_1 – пользователь, то он автоматически получает права доступа к объекту o_3 , соответ-

ствующие отношению r_{13} . При этом по отношению к более ранней версии реляционной модели ЛРД отменяется требование, что объект o_1 является пользователем, чтобы данное правило разрешения доступа можно было применять по цепочке объектов большей длины.

Определение 2. В условиях определения 1 элементы множества *RelationTransitions* называются правилами переноса отношений. Если $(r_1, r_2, r_3) \in \text{RelationTransitions}$, то отношения r_1 и r_2 называются непосредственно порождающими r_3 . Отношение r_0 называется порождающим отношение r' , если существуют такие отношения r_1, r_2, \dots, r_n , что $r_n = r'$ и для любого $i = 1, \dots, n$ отношение r_{i-1} непосредственно порождает r_i .

Отображение *RelationType* определяет тип отношения между объектами и принимает одно из четырех значений: *One – to – One*, *One – to – Many*, *Many – to – One*, *Many – to – Many*. Если *RelationType*(r) = *One – to – One* и $(o_1, r, o_2) \in \text{ObjectRelations}$, то не существует такого объекта o_3 , отличного от o_1 , что $(o_3, r, o_2) \in \text{ObjectRelations}$; а также не существует такого объекта o_4 , отличного от o_2 , что $(o_1, r, o_4) \in \text{ObjectRelations}$. Примером такого отношения является отношение «пользователь ассоциирован с определенным сотрудником». Если *RelationType*(r) = *One – to – Many* и $(o_1, r, o_2) \in \text{ObjectRelations}$, то не существует такого объекта o_3 , отличного от o_1 , что $(o_3, r, o_2) \in \text{ObjectRelations}$. Примером такого отношения является отношение «подразделение o_1 содержит в себе подразделение o_2 ». Если *RelationType*(r) = *Many – to – One* и $(o_1, r, o_2) \in \text{ObjectRelations}$, то не существует такого объекта o_3 , отличного от o_2 , что $(o_1, r, o_3) \in \text{ObjectRelations}$. Примером такого отношения является отношение «данная статья опубликована в данном журнале». Если *RelationType*(r) = *Many – to – Many*, то любой объект может быть связан этим отношением с любым количеством других. Примером является отношение «данный сотрудник является автором данной работы».

Пользователь u имеет право доступа a к объекту o класса c и типа t в том и только том случае, если a является разрешенным действием для объектов класса c и выполнено одно из следующих условий:

- $\exists r \in Roles : (u, r) \in UserRoles, (r, t, a) \in RoleTypeAccessGranted;$
- $\exists r_1, r_2 \in Roles : r_1 \leq r_2, (u, r_2) \in UserRoles, (r_1, t, a) \in RoleTypeAccessGranted;$
- $\exists r \in Relations : (u, r, o) \in ObjectRelations, (r, t, a) \in RelationTypeAccessGranted;$
- $\exists r_1, r_2 \in Relations : r_1 \leq r_2, (u, r_2, o) \in ObjectRelations, (r_1, t, a) \in RelationTypeAccessGranted;$
- $\exists r_{12}, r_{23}, r_{13} \in Relations, \exists o_2 \in O : (u, r_{12}, o_2) \in ObjectRelations, (o_2, r_{23}, o) \in ObjectRelations, (r_{13}, t, a) \in RelationTypeAccessGranted, (r_{12}, r_{23}, r_{13}) \in RelationTransitions;$

и при этом не выполнено ни одно из следующих условий:

- $\exists r \in Roles : (u, r) \in UserRoles, (r, t, a) \in RoleTypeAccessDenied;$
- $\exists r_1, r_2 \in Roles : r_1 \leq r_2, (u, r_2) \in UserRoles, (r_1, t, a) \in RoleTypeAccessDenied;$
- $\exists r \in Relations : (u, r, o) \in ObjectRelations, (r, t, a) \in RelationTypeAccessDenied;$
- $\exists r_1, r_2 \in Relations : r_1 \leq r_2, (u, r_2, o) \in ObjectRelations, (r_1, t, a) \in RelationTypeAccessDenied;$
- $\exists r_{12}, r_{23}, r_{13} \in Relations, \exists o_2 \in O : (u, r_{12}, o_2) \in ObjectRelations, (o_2, r_{23}, o) \in ObjectRelations, (r_{13}, t, a) \in RelationTypeAccessDenied, (r_{12}, r_{23}, r_{13}) \in RelationTransitions;$

В обоих случаях применение правил, связанных с множеством *RelationTransitions* также может быть возможно по цепочке. Данный механизм подробно рассмотрен далее. Заметим, что применение

этих правил является трудоемким процессом, так как в ряде случаев проверка доступа конкретного пользователя к конкретному объекту требует проверки свойств множества других объектов. Так, например, для проверки отношения «данный пользователь является ответственным по подразделению, в котором работает данный сотрудник», необходимо перебрать все подразделения, ответственным по которым является пользователь (их может быть несколько) на принадлежность к ним сотрудника (один сотрудник также может числиться в нескольких подразделениях).

Следует также заметить, что некоторые операции при работе с Системой требуют одновременного осуществления доступа к нескольким объектам. Так, при связывании автора с результатом научно-исследовательской деятельности необходимо осуществить доступ «включить в число авторов результата» к объекту «сотрудник» и доступ «включить сотрудника в число авторов» к объекту «результат научно-исследовательской деятельности». Операция завершается успешно в том и только том случае, если пользователь имеет право на оба этих доступа.

Алгоритмы анализа реляционной модели

Как следует из изложенного выше, одной из особенностей реляционной модели ЛРД является высокая трудоемкость проверки разрешения конкретного пользователя на конкретный доступ к конкретному объекту. При этом следует заметить, что некоторые из вычислений, которые необходимо проделать для принятия решения о предоставлении пользователю доступа к объекту, зависят только от множеств, «жёстко» заданных в модели ЛРД подконтрольной Системы и не изменяющихся в процессе ее функционирования. Эти вычисления можно проводить только при изменении модели ЛРД Системы, а не при каждой проверке доступа пользователя к объекту.

Определение 3. В определении 1 множества *Classes*, *Types*, *Actions*, *Relations* и *Roles*; отношения *RelationTypeAccessGranted*, *RoleTypeAccessGranted*, *ClassAccess*, *RelationLeftClasses*, *RelationRightClasses* и

RelationTransitions; отображение *RelationType*; а также отношения порядка на множествах *Roles*, *Relations* и *Classes* называются статическими.

Определение 4. В определении 1 множества *Objects*, *Users*, *Groups* и *Sessions*; отношения *ObjectRelations* и *UserRoles*; отображения *ObjectClass*, *ObjectTypes* и *SessionUser*; а также отношение порядка на множестве групп пользователей называются динамическими.

Смысл представленных определений в том, что статические множества и отношения определяются администратором Системы при задании политики безопасности и не изменяются при функционировании Системы до загрузки новой политики безопасности. Динамические множества и отношения могут изменяться при функционировании Системы без вмешательства администратора, в рамках определенных правил. Разделение базовых множеств модели на статические и динамические необходимо для оптимизации трудоемкости проверки доступа субъектов к объектам по той причине, что вычисления над элементами статических множеств можно выполнять однократно при загрузке политики безопасности, в то время как вычисления, затрагивающие динамические множества, необходимо проводить при каждой попытке доступа пользователя к объекту.

Алгоритмы проверки прав доступа, предоставляемых с использованием цепочек отношений

Из всех условий, которые необходимо проверить для определения разрешенных видов доступа пользователя к объекту, наиболее трудоемкой задачей является проверка условий, связанных с использованием отношения *RelationTransitions*. Так, например, в настоящее время одним из условий предоставления пользователю доступа на изменение информации о результатах его деятельности является: «пользователь выполняет обязанности ответственного по подразделению, в котором работает сотрудник, являющийся автором статьи». Наиболее очевидный алгоритм проверки этого условия требует про-

верки всех авторов работы на принадлежность ко всем подразделениям, ответственным по которым является пользователь. Ввиду потенциально высокой трудоемкости операции проверки доступа с использованием переноса отношений, возникает необходимость выработки точной оценки трудоемкости подобных операций и построения алгоритмов проверки доступа, оптимальных по трудоемкости. Для этой цели потребуется определить ряд вспомогательных объектов.

Определение 5. В условиях определения 1 обозначим:

- $ObjectRelations_0 = ObjectRelations$;
- $ObjectRelations_{n+1} = ObjectRelations_n \cup \{(o_1, r_{13}, o_3) : \exists o_2 \in Objects \exists r_{12}, r_{23} \in Relations : (o_1, r_{12}, o_2) \in ObjectRelations, (o_2, r_{23}, o_3) \in ObjectRelations, (r_{12}, r_{23}, r_{13}) \in RelationTransitions\}$ для $n \geq 0$;
- $ExtendedObjectRoles = \bigcup_{n \geq 0} ObjectRelations_n$.

Определение 6. В условиях определения 1 цепочкой отношений называется конечное упорядоченное множество отношений.

Определение 7. В условиях определения 1 будем считать, что отношение r порождается цепочкой отношений (r_1, \dots, r_n) , если:

- $n = 1$ и $r_1 = r$;
- $n = 2$ и $(r_1, r_2, r) \in RelationTransitions$;
- $n > 2$ и $\exists r' \in Relations \exists i \in (1, \dots, n - 1) : (r_i, r_{i+1}, r') \in RelationTransitions$ и отношение r порождается цепочкой отношений $(r_1, \dots, r_{i-1}, r', r_{i+2}, \dots, r_n)$.

Цепочка отношений длины 1 называется вырожденной.

Теорема 1. Пусть $R = (r_1, \dots, r_n)$ — невырожденная цепочка отношений, порождающая отношение r . Тогда существуют

такие отношения r' и r'' и число $m < n$, что $(r', r'', r) \in RelationTransitions$, цепочка отношений $R' = (r_1, \dots, r_m)$ порождает отношение r' , а цепочка $R'' = (r_{m+1}, \dots, r_n)$ порождает отношение r'' .

Доказательство. Если $n = 2$ то утверждение теоремы верно, так как $r' = r_1, r'' = r_2$ и $m = 1$.

Пусть утверждение теоремы верно для всех цепочек длины $n \leq k$. Пусть $n = k + 1$, R – цепочка длины n , порождающая отношение r . Тогда $\exists r''' \in Relations \exists i \in (1, \dots, n - 1): (r_i, r_{i+1}, r') \in RelationTransitions$ и отношение r порождается цепочкой отношений $(r_1, \dots, r_{i-1}, r''', r_{i+2}, \dots, r_n)$. Эта цепочка имеет длину $n - 1$, поэтому для нее отношения r' и r'' и число m , соответствующие условиям теоремы, существуют. Если $m \leq i - 1$, то цепочка отношений $R' = (r_1, \dots, r_m)$ порождает отношение r' , а цепочка $R''_{n-1} = (r_{m+1}, \dots, r''', r_{i+2}, \dots, r_n)$ порождает отношение r'' . Тогда и цепочка $R'' = (r_{m+1}, \dots, r_i, r_{i+1}, r_{i+2}, \dots, r_n)$ также порождает отношение r'' и таким образом утверждение теоремы доказано для цепочки R . Случай $m \geq i$ рассматривается аналогично.

Таким образом, утверждение теоремы доказано методом математической индукции по длине цепочки R . \square

В примере, рассмотренном выше, пользователь получает доступ на изменение данных о результате научно-исследовательской деятельности благодаря двойному переносу отношений:

1. $(user, is_responsible_for, subdivision) \in ObjectRoles,$
 $(subdivision, works_in, worker) \in ObjectRoles,$
 $(is_responsible_for, worker_in_subdivision,$
 $is_responsible_for) \in RelationTransitions,$
 $\rightarrow (user, is_responsible_for, worker) \in ExtObjectRelations;$
2. $(user, is_responsible_for, worker) \in ExtObjectRelations,$
 $(worker, is_author_of, work) \in ObjectRelations,$
 $(is_responsible_for, is_author_of,$

$is_responsible_for) \in RelationTransitions,$
 $\rightarrow (user, is_responsible_for, work) \in ExtObjectRelations.$

По мере дальнейшего развития Системы может появиться необходимость в предоставлении пользователю доступа к объекту и по более длинной цепочке отношений.

Определение 8. В условиях определения 1 множество $RelationTransitions$ называется ациклическим, если для любого отношения r и порождающей его цепочки (r_1, \dots, r_n) $r \neq r_i$ для $i \in (1, \dots, n)$.

Теорема 2. В условиях определения 1, если множество $RelationTransitions$ является ациклическим, то любое отношение, которое может быть порождено цепочкой отношений, и при этом любая порождающая его цепочка имеет длину, не превосходящую 2^{R-1} , где R – количество рассматриваемых отношений в Системе.

Доказательство. Назовем отношение r отношением нулевого уровня, если не существует таких отношений r_1 и r_2 , что $(r_1, r_2, r) \in RelationTransitions$. Если множество $RelationTransitions$ является ациклическим, то хотя бы одно отношение нулевого уровня существует.

В противном случае, возьмем произвольное отношение, обозначим его r_0 . Далее, для любого $n \geq 0$ найдем такое отношение r_{n+1} , что для некоторого отношения r' $(r_{n+1}, r', r_n) \in RelationTransitions$. Для любого n существует отношение r_{n+1} , так как в противном случае отношение r_n было бы нулевого уровня. Таким образом, построена бесконечная последовательность отношений, в которой каждое отношение r_n порождается некоторой цепочкой, содержащей следующее отношение r_{n+1} . Поскольку множество отношений конечно, хотя бы одно отношение r встретится в этой последовательности более одного раза, то есть r порождается цепочкой, содержащей r , что противоречит ацикличности множества $RelationTransitions$.

Таким образом, предположение о том, что в Системе не существует отношений нулевого уровня, привело к противоречию, то есть в Системе существует хотя бы одно отношение нулевого уровня.

Будем считать, что отношение r имеет уровень $n + 1$, где $n \geq 0$, если существуют такие отношения r_1, r_2 , что $(r_1, r_2, r) \in RelationTransitions$, и хотя бы одно из отношений r_3 и r_4 имеет уровень n , и не существует таких отношений существуют такие отношения r_1, r_2 , что $(r_1, r_2, r) \in RelationTransitions$, и хотя бы одно из отношений r_3 и r_4 имеет больший уровень, чем n .

Таким образом каждое отношение в Системе получает некоторый уровень, и если существует отношение уровня n , то существуют отношения всех уровней, меньших, чем n . Это значит, что уровень любого отношения в Системе меньше общего количества отношений.

Если отношение r имеет уровень $n > 0$, то любая порождающая его цепочка имеет длину не более 2_n . Докажем данное утверждение методом математической индукции по n . Для $n = 1$ утверждение верно, так как если r порождается цепочкой (r_1, r_2) , то отношения r_1 и r_2 имеют нулевой уровень и порождены другими цепочками быть не могут.

Если отношение r уровня $n + 1$ порождается некоторой цепочкой, то в соответствии с определением 7 существуют такие отношения r_1 и r_2 , что $(r_1, r_2, r) \in RelationTransitions$ и цепочка, порождающая r либо имеет вид (r_1, r_2) , либо является конкатенацией цепочек, порождающих r_1 и r_2 . Поскольку уровни отношений r_1 и r_2 не превосходят n , длины порождающих их цепочек не превосходят 2^n , а длина их конкатенации не превосходит 2^{n+1} . Таким образом, отношение любого уровня $n > 0$ порождается цепочками длины не более 2^n . Поскольку максимальный уровень отношения меньше общего количества отношений, из этого следует утверждение теоремы. \square

Заметим, что в реальных системах максимальный уровень отношения может иметь длину, существенно меньшую, чем общее количество отношений. Несмотря на то, что к настоящему времени полного математического анализа иерархии порождения отношений в системе ИСТИНА не проводилось, есть основания ориентировать-

ся на максимальный уровень отношений в ней равный 2 – 3.

Наиболее простой алгоритм нахождения уровней всех отношений в Системе имеет следующий вид.

для каждого отношения r из *Relations*:

для каждой тройки отношений (r_1 , r_2 , r_3)

из *RelationTransitions*:

если $r = r_3$, то установить для r флаг "пропустить"

если для r не установлен флаг "пропустить",

то присвоить r уровень 0

для N от 1 до общего количества отношений:

установить количество пропущенных отношений в 0

для каждого отношения r из *Relations*:

сбросить для r флаг "пропустить"

если r уже присвоен уровень,

то закончить обработку этого r

для каждой тройки отношений (r_1 , r_2 , r_3)

из *RelationTransitions*:

если r не равно r_3 ,

то закончить обработку этой тройки

если уровень r_1 не присвоен

или уровень r_2 не присвоен,

то установить для r флаг "пропустить"

если r_1 или r_2 имеет уровень N ,

то установить для r флаг "пропустить"

если для r не установлен флаг "пропустить",

то присвоить r уровень N

иначе увеличить количество пропущенных отношений

если количество пропущенных отношений равно 0,

то закончить работу.

Трудоёмкость первого этапа данного алгоритма не превосходит $\|Relations\| * \|RelationTransitions\|$. Трудоёмкость второго этапа не превосходит $\|Relations\| * \|Relations\| * \|RelationTransitions\|$.

Следующий алгоритм строит для каждого отношения список цепочек, порождающих его.

Отсортировать множество отношений по уровням

для каждого отношения r

(сначала обрабатываются отношения меньших уровней):

5.1. МОДЕЛИ ЛОГИЧЕСКОГО РАЗГРАНИЧЕНИЯ ДОСТУПА

список цепочек, порождающих $r = \{\text{цепочка из одного } r\}$
если r имеет уровень 0, то закончить обработку r .
для каждой тройки (r_1, r_2, r_3) из RelationTransitions:
если r не равно r_3 ,
то закончить обработку этой тройки
для каждой цепочки, $chain1$, порождающей r_1
для каждой цепочки, $chain2$, порождающей r_2
добавить конкатенацию цепочек $chain1$ и $chain2$
к списку цепочек, порождающих r_3

Трудоёмкость сортировки множества отношений по уровням при использовании алгоритма пирамидальной сортировки не превосходит $\|Relations\| * \log_2 \|Relations\|$. Трудоёмкость основной части алгоритма равна $\sum_{n=0}^N LevelTime(n)$, где N – число уровней отношений (оно не превосходит общего количества отношений в Системе), а $LevelTime(n)$ – трудоёмкость обработки всех отношений уровня n при условии, что для всех отношений уровня менее n все порождающие цепочки уже построены.

Трудоёмкость обработки нулевого уровня пропорциональна количеству отношений нулевого уровня. Заметим, что алгоритм перечисления всех цепочек отношений, порождающих данное отношение, использует только статические множества, то есть выполнять его необходимо только при перезагрузке модели логического разграничения доступа Системы.

Одновременно с этим, наиболее существенным для удобства работы с Системой фактором является трудоёмкость операций, над динамическими множествами, которые выполняются во время функционирования Системы, непосредственно при проверке прав доступа пользователя к объекту.

Рассмотрим пример. Пусть необходимо определить, имеет ли пользователь u право изменить объект o , относящийся к классу «конференция». Пусть модель логического разграничения доступа предполагает, что пользователь имеет такое право в том случае, если является ответственным по подразделению, в которое (или в одно из его дочерних) входит один из авторов одного из докладов на рассматриваемой конференции. Чтобы проверить это условие, необходимо:

- перебрать все подразделения, по которым пользователь является ответственным;
- для каждого из них перебрать все дочерние подразделения;
- для каждого из них перебрать всех сотрудников;
- для каждого из сотрудников проверить, все доклады на конференциях, проверив, есть ли среди них доклад на данной конференции.

Очевидно, что выполнять данную проверку при каждой попытке пользователя изменить атрибуты конференции будет крайне трудоемко. Предлагается два варианта сокращения трудоемкости этой операции.

- При реализации механизмов безопасности, для проверки возможность доступа пользователя к объекту, использовать не функцию, возвращающую значение типа `Boolean` на основании идентификаторов пользователя, объекта и вида доступа, а функцию, возвращающую список возможных видов доступа на основании идентификаторов субъекта и объекта. В ряде случаев это поможет избежать многократных проверок цепочек отношений. Так, например, при отображении некоторых страниц в Системе желательно отображать ссылки на страницы, соответствующие разрешенным действиям пользователя, в виде гиперссылок, однако при этом не отображать ссылки на страницы, соответствующие запрещенным для этого пользователя действиям. В этом случае при загрузки страницы потребуется многократно проверять, разрешен ли пользователю тот или иной доступ к объекту, и время загрузки страницы можно существенно сократить, если проанализировать цепочки отношений между пользователем и объектом многократно.
- Постоянно хранить в памяти расширенное множество объектов отношений *ExtObjectRelations*, а не только множество *ObjectRelations*. При этом необходимо при каждом изменении множества *ObjectRelations* приводить множество

ExtObjectRelations в соответствии с ним. Далее представлены оценки объема памяти, которая необходима для хранения расширенного множества отношений и трудоемкости его поддержки в актуальном состоянии.

Трудоемкость поддержки расширенного множества отношений в актуальном состоянии

Заметим, что любое изменение множества *ObjectRelations* требует изменения состояния как минимум двух объектов в Системе – при включении или исключении из множества *ObjectRelations* тройки (o_1, r, o_2) изменяется состояние объектов o_1 и o_2 . Этот факт означает, что для осуществления пользователем такого изменения потребуется не менее двух проверок прав доступа – к объекту o_1 и к объекту o_2 . Таким образом, проверка прав доступа в Системе выполняется не менее чем в два раза чаще, чем изменение множества *ObjectRelations*. Заметим, что в действительности проверка прав доступа пользователей к объектам Системы выполняется намного чаще по причине, отмеченной в предыдущем разделе, а именно – при отображении некоторых страниц в Системе желательно отображать ссылки на страницы, соответствующие разрешенным действиям пользователя, в виде гиперссылок, но при этом не отображать ссылки на страницы, соответствующие запрещенным для этого пользователя действиям.

С учетом изложенного, одним из способов уменьшить трудоемкость проверки прав доступа пользователей к объектам Системы является постоянное хранение в памяти и поддержка в актуальном состоянии расширенного множества отношений между объектами. Следует, однако, учесть, что при этом возрастает не только трудоемкость операции связывания отношением двух объектов, но и объем памяти, необходимой для функционирования Системы. В связи с этим обстоятельством, необходимо получить оценки объема памяти, необходимой для хранения расширенного множества отношений, а также затрат времени на поддержание этого множества в актуальном состоянии при изменении основного множества отношений.

Определение 9. Пусть в информационной системе задана ре-

ляционная модель ЛРД. Графом переноса отношений называется ориентированный граф $RelationTransitionsGraph = (Relations, \{(r_1, r_2) : \exists r_3(r_2, r_3, r_1) \in RelationTransitions\} \cup \{(r_1, r_2) : \exists r_4(r_4, r_2, r_1) \in RelationTransitions\})$.

Данное определение означает, что вершинами графа переноса отношений являются отношения между объектами Системы, и ребра графа соединяют каждое отношение с отношениями, непосредственно порождающими его.

Теорема 3. Множество $RelationTransitions$ является ациклическим в том и только том случае, если соответствующий ему граф переноса отношений не содержит ориентированных циклов.

Лемма 3.1. В графе переноса отношений существует путь от отношения r к отношению r' в том и только том случае, если r порождается некоторой цепочкой, содержащей r' .

Доказательство. Пусть отношение r порождается цепочкой (r_1, \dots, r_n) и $\exists i \in \{1, \dots, n\} r_i = r'$. Докажем, что в этом случае в графе переноса отношений существует путь от r к r' . При $n = 2$ цепочка отношений, порождающая r , имеет вид (r', r'') , либо (r'', r') . Это значит, что либо $(r', r'', r) \in RelationTransitions$, либо $(r'', r', r) \in RelationTransitions$, и, следовательно, в графе переноса отношений есть ребро, соединяющее r и r' .

Пусть для $n = k \geq 2$ верно, что если отношение r порождается цепочкой отношений длины n содержащей r' , то в графе переноса отношений есть путь от r к r' . Пусть $n = k + 1$. Согласно определению 7, если отношение r порождается цепочкой отношений длины n содержащей r' , то $\exists r'' \in Relations \exists i \in (1, \dots, \dots n - 1) : (r_i, r_{i+1}, r'') \in RelationTransitions$ и отношение r порождается цепочкой отношений $(r_1, \dots, r_{i-1}, r'', r_{i+2}, \dots, r_n)$. В этом случае цепочка отношений $(r_1, \dots, r_{i-1}, r'', r_{i+2}, \dots, r_n)$ имеет длину k , а значит, в графе переноса отношений есть путь, соединяющий отношение r с каждым из отношений этой цепочки. Если r' входит в эту цепочку, то утверждение доказано. В противном случае $r' = r_i$, либо

$r' = r_{i+1}$. Тогда в графе переноса отношений есть ребро, соединяющее отношение r'' с отношением r' . Вместе с тем, в графе переноса отношений также есть путь от отношения r к отношению r'' , поскольку r'' входит в цепочку длины r , порождающую r . Добавив к этому пути ребро (r'', r) можно построить путь, соединяющий r и r' .

Таким образом, одно из утверждение леммы доказано методом математической индукции по длине цепочки, порождающей r . Докажем обратное утверждение.

Пусть граф переноса отношений содержит путь, соединяющий отношения r и r' , пусть m – длина этого пути. Если $m = 1$, то это означает, что в графе переноса отношений есть ребро (r, r') , то есть существует такое отношение r'' , что $(r', r'', r) \in RelationTransitions$, либо $(r'', r', r) \in RelationTransitions$, таким образом отношение r порождается цепочкой отношений, содержащей r' .

Пусть для $m = k$ верно. Если отношения r и r' соединены в графе переноса отношений путем длины не более m , то отношение r порождается цепочкой отношений, содержащей r' . Пусть $m = k + 1$. Пусть r'' – это такое отношение, что в графе переноса отношений есть ребро (r'', r') , входящее в путь от r к r' . Тогда отношение r соединено с r'' путем длины не более k , а значит, отношение r порождается цепочкой, содержащей отношение r'' . При этом, поскольку в графе переноса отношений есть ребро (r'', r') , существует такое отношение r''' , что либо $(r''', r', r'') \in RelationTransitions$, либо $(r', r''', r'') \in RelationTransitions$. Если в цепочке отношений, содержащей r'' и порождающей r , заменить вхождение отношения r'' соответственно на (r', r''') , либо (r''', r') , то можно получить цепочку отношений, содержащую r' и порождающую r .

Таким образом, и второе утверждение леммы доказано методом математической индукции по длине пути, соединяющего отношения r и r' . \square

Доказательство теоремы. Пусть граф переноса отношений $RelationTransitionGraph$ содержит хотя бы один ориентированный цикл $(r_0, r_1, r_2, \dots, r_n = r_0)$. Это значит, что

для любого $i = 0, 1, \dots, n - 1$ существует такое отношение r'_i , что $(r_i, r'_i, r_{i+1}) \in RelationTransitions$, либо r'_i , что $(r_i, r'_i, r_{i+1}) \in RelationTransitions$.

Отмеченный выше факт означает, что отношение для любого $i = 0, 1, \dots, n$ отношение r_{i+1} порождается цепочкой, содержащей r_i – либо цепочкой (r_i, r'_i) , либо (r'_i, r_i) . В частности, существует цепочка отношений, содаржащая r_{n-1} и порождающая r_n . Если в этой цепочке заменить r_{n-1} на цепочку, содержащую r_{n-2} и порождающую r_{n-1} , можно получить цепочку, содержащую r_{n-2} и порождающую r_n . Повторяя этот шаг, можно построить цепочку отношений, содержащую отношение r_0 и порождающую отношение r_n . Поскольку изначально $r_0 = r_n$, это значит, что r_0 порождается цепочкой отношений, содержащей r_0 . Таким образом, если граф переноса отношений содержит ориентированный цикл, то множество правил переноса отношений не является ациклическим.

Докажем обратное утверждение. Пусть множество правил переноса отношений не является ациклическим, то есть существует такое отношение r и порождающая его цепочка отношений (r_1, \dots, r_n) , что $n \geq 2$ и $r = r_i$ для некоторого $i \in \{1, \dots, n\}$. В этом случае, согласно лемме, в графе переноса отношений есть путь, соединяющий вершину r с ней самой, то есть в графе переноса отношений существует ориентированный цикл. \square

Определение 10. Пусть в информационной системе задана реляционная модель ЛРД, r – одно из отношений. Обозначим как $RelationTransitionsSubgraph(r)$ подграф графа $RelationTransitionsGraph$, состоящий из всех его вершин, к которым существует путь от вершины r , и всех ребер, встречающихся в этих путях.

Определение 11. Пусть в информационной системе задана реляционная модель ЛРД. Множество $RelationTransition$ называется древовидным, если оно является ациклическим, то есть для любого отношения r существует не более одной такой пары отношений (r_1, r_2) , что $(r_1, r_2, r) \in RelationTransitions$, и множество $RelationTransitions$ не содержит таких элементов (r_1, r_2, r_3) ,

что $r_1 = r_2$.

В этом случае для любого отношения r подграф $RelationTransitionsSubgraph(r)$ является бинарным деревом.

Определение 12. Отношение r называется неприводимым, если в множестве $RelationTransitions$ нет элементов вида (r_1, r_2, r) ни для каких r_1 и r_2 . Цепочка, состоящая только из неприводимых отношений, также называется неприводимой.

Теорема 4. Пусть в информационной системе задана реляционная модель ЛРД с древовидным множеством правил переноса отношений. Тогда для любого отношения r существует единственная порождающая его цепочка отношений, все отношения в которой неприводимы.

Доказательство. Пусть $Level(r)$ – это уровень отношения, определенный в доказательстве теоремы 2. Заметим, что отношение r неприводимо в том и только том случае, если его уровень равен 0. Если отношение r является неприводимым, то утверждение верно, поскольку единственной цепочкой, порождающей r , является цепочка из одного r . Пусть утверждение теоремы верно для всех отношений уровня не более n . Пусть отношение r имеет уровень $n + 1$. Тогда существует единственная такая пара отношений (r_1, r_2) , что $(r_1, r_2, r) \in RelationTransitions$. Отношения r_1 и r_2 имеют уровень не более n , поэтому утверждение теоремы для них верно. Пусть $R_1 = (r_{1,1}, r_{1,2}, \dots)$ и $R_2 = (r_{2,1}, r_{2,2}, \dots)$ соответственно порождающие их неприводимые цепочки отношений, цепочка R – их конкатенация. Тогда цепочка R порождает отношение r . Пусть есть еще одна неприводимая цепочка отношений R' , также порождающая r .

Согласно теореме 1 цепочка отношений R' также является конкатенацией двух цепочек отношений R'_1 и R'_2 , порождающих соответственно отношения r_1 и r_2 . Однако для каждого из этих отношений существует единственная порождающая его неприводимая цепочка, поэтому $R'_1 = R_1$ и $R'_2 = R_2$. Следовательно, $R' = R$, что и требовалось доказать. \square

Теорема 5. Пусть в информационной системе задана реляционная модель ЛРД с древовидным множеством правил переноса отношений, r – одно из отношений, $R = (r_1, \dots, r_n)$ – порождающая его неприводимая цепочка, причем все отношения r_j различны. Пусть C_0 – множество всех объектов, входящих в левые классы для r_1 , C_n – множество всех объектов, входящих в левые классы для r_n , и для $j \in \{1, \dots, n - 1\}$ C_j – это пересечение множеств объектов, входящих в левые классы для r_{j+1} и в правые классы для r_j . Пусть для некоторого i и некоторых объектов $o_{i-1} \in C_{i-1}$ и $o_i \in C_i$ в множество *ObjectRelations* добавляется тройка (o_{i-1}, r_i, o_i) . Тогда трудоемкость обновления расширенного множества отношений между объектами при изменении множества *ObjectRelations* не превосходит произведения всех $\|C_j\|$, кроме $j = i$ и $j = i - 1$.

Доказательство. Пусть $n = 2$. Тогда, при связывании объектов o_1 и o_2 отношением r_2 необходимо также связать объект o_2 отношением r со всеми объектами o_0 для которых $(o_0, r_1, o_1) \in \text{ObjectRelations}$. В худшем случае для этого необходимо проверить условие $(o_0, r_1, o_1) \in \text{ObjectRelations}$ для каждого $o_0 \in C_0$. Трудоемкость такой проверки равна $\|C_0\|$. Аналогично, при добавлении в множество *ObjectRelations* элемента (o_0, r_1, o_1) для приведения множества *ExtObjectRelations* в соответствие с изменением множества *ObjectRelations* потребуется $\|C_2\|$ операций. Таким образом, для $n = 2$ утверждение теоремы верно.

Пусть $n > 2$. Пусть r_l и r_r – такие отношения, что $(r_l, r_r, r) \in \text{RelationTransitions}$, R_l и R_r – порождающие их неприводимые цепочки отношений. Пусть отношение r_i входит в цепочку R_l . Пусть k – длина цепочки отношений R_l . Тогда для обновления отношения r_l потребуется не более $\prod_{j \leq k, j \neq i, j \neq i-1} \|C_j\|$ операций, каждая из которых связывает объект o_0 класса C_0 отношением r_l с объектом класса C_k . При выполнении каждой из них может также потребоваться связать объект o_0 со всеми объектами o_n класса C_n , которые связаны отношением r_r с объектом o_k . Для этого требуется не более $\|C_n\|$ операций. Таким образом, общее количество пар объектов, которые необходимо связать отношением r при связывании двух объектов от-

ношением r_j , входящим в цепочку, порождающую r , не превосходит $\|C_n\| * \prod_{j \leq k, j \neq i, j \neq i-1} \|C_j\| \leq \prod_{j \leq n, j \neq i, j \neq i-1} \|C_j\|$, что и требовалось доказать. \square

Следует отметить, что данная оценка трудоемкости является крайне грубой, и в настоящее время активно ведется работа над ее улучшением.

Заметим, что при хранении в памяти Системы расширенного множества отношений операция разрыва отношения между объектами является намного более трудоемкой, чем операция связывания объектов отношениями. Рассмотрим пример. Пусть $(r_{12}, r_{23}, r_{13}) \in RelationTransitions$, $(o_1, r_{12}, o_2) \in ObjectRelations$, $(o_2, r_{23}, o_3) \in ObjectRelations$, $(o_1, r_{13}, o_3) \in ExtObjectRelations$. Пусть из множества $ObjectRelations$ удаляется тройка (o_1, r_{12}, o_2) . При этом из множества $ExtObjectRelations$ следует удалить тройку (o_1, r_{13}, o_3) в том и только том случае, если не выполнено ни одно из следующих условий:

- $(o_1, r_{13}, o_3) \in ObjectRelations$;
- $\exists o'_2 : (o_1, r_{12}, o'_2) \in ObjectRelations, (o_2, r_{23}, o_3) \in ObjectRelations$;
- $\exists r'_{12}, r'_{23} \in Relations \exists o'_2 \in Objects : (r'_{12}, r'_{23}, r_{13}) \in RelationTransitions, (o_1, r'_{12}, o'_2) \in ObjectRelations, (o'_2, r'_{23}, o_3) \in ObjectRelations$.

Прямая проверка данных условий является крайне трудоемкой операцией.

Для уменьшения трудоемкости операции разрыва отношения между объектами предлагается следующий подход.

Определение 13. Для элемента $(o_s, r, o_f) \in ExtObjectRelations$ его мощностью называется $ObjectRelationPower(o_s, r, o_f) =$

- ∞ , если $(o_s, r, o_f) \in ObjectRelations$;

- n , если $(o_s, r, o_f) \notin ObjectRelations$ и существует ровно n таких различных последовательностей объектов $(o_s = o_0, o_1, \dots, o_m = o_s)$, что для любого $i \in (1, \dots, m)$ существует такое отношение r_i , что $(o_{i-1}, r_i, o_i) \in ObjectRelations$ и цепочка (r_1, \dots, r_m) порождает отношение r .

Если тройка $(o_s, r, o_f) \notin ExtObjectRelations$ будем считать, что ее мощность $ObjectRelationPower(o_s, r, o_f) = 0$.

Если вместе с каждым элементом множества $ExtObjectRelations$ хранить в памяти Системы его мощность, можно сделать трудоемкость операции разрыва отношения между объектами такой же, как и трудоемкость связывания двух объектов отношением.

Приведение «почти ациклических» правил переноса отношений к ациклическому виду

Представленные в предыдущем пункте рассуждения верны для случая, когда множество правил переноса отношений Системы является ациклическим, то есть ни одно из отношений в Системе не может быть порождено цепочкой, содержащей это отношение. Однако, данное правило выполнено не во всех системах, для которых целесообразно использование реляционной модели ЛРД. Так, в системе ИСТИНА множество правил переноса отношений не является ациклическим из-за неограниченной глубины вложенности подразделений МГУ. Поскольку отношение «подразделение A входит в подразделение B » является транзитивным, в Системе присутствует порождающее его правило вида $(subdivision_in_subdivision, subdivision_in_subdivision, subdivision_in_subdivision) \in RelationTransitions$.

Определение 14. В условиях определения 1 правило переноса отношений (r_1, r_2, r_3) называется правилом транзитивности, если $r_1 = r_2 = r_3$.

Рассмотрим информационную систему, использующую реляционную модель ЛРД со следующими свойствами:

- множество правил переноса ролей Системы содержит правила транзитивности;
- ни одно из отношений, встречающихся в правилах транзитивности, не порождается никакими отношениями, кроме самого себя;
- множество $RelationTransitions'$, полученное из $RelationTransitions$ удалением всех правил транзитивности, является ациклическим.

В этом случае модель ЛРД Системы можно привести к ациклическому виду, следующим преобразованием:

- исключить из правил переноса отношений все правила транзитивности;
- при связывании пары объектов Системы o_1 и o_2 отношением транзитивным отношением r включать в множество $ObjectRelations$ не только элемент (o_1, r, o_2) , но и все элементы, которые должны быть включены в него вместе с ним для сохранения транзитивности отношения r .

При таком подходе, однако, становится крайне трудоемкой задача разрыва отношения r между двумя объектами. В самом деле, если в множество $ObjectRelations$ добавляется элемент (o_1, r, o_2) то вместе с ним должны быть добавлены элементы (o_3, r, o_4) для любых o_3 и o_4 , для которых $(o_3, r, o_1) \in ObjectRelations$ и $(o_2, r, o_4) \in ObjectRelations$. Одновременно с этим, при удалении элемента (o_1, r, o_2) из множества $ObjectRelations$ вместе с ним должны быть удалены только те элементы (o_3, r, o_4) , которые не связаны «альтернативной» цепочкой объектов, связанных отношением r и не содержащей объектов o_1 и o_2 .

Заметим, что в случае отношения вложенности подразделений в системе ИСТИНА изменение правил его хранения рассмотренным способом не приведет к резкому увеличению трудоемкости операции

разрыва отношения в силу того, что отношение вложенности подразделений имеет древовидную структуру, то есть каждое из подразделений может входить не более чем в одно другое подразделение. По этой причине при исключении подразделения *A* из подразделения *B* все дочерние подразделения *A* перестают быть дочерними для всех родительских подразделений *B*. Трудоемкость операции включения одного подразделения в другое и исключения одного подразделения из другого при этом равна произведению количества дочерних подразделений включаемого подразделения и количества родительских подразделений для подразделения, в которое его необходимо включить. Следует также принимать во внимание следующее:

- в силу древовидной структуры отношения включения подразделений одно в другое, количество родительских подразделений для заданного подразделения обычно незначительное;
- включение одного подразделения в другое производится в Системе достаточно редко, в то время как операция проверки принадлежности одного подразделения другому выполняется намного чаще и приведение правил переноса отношений к ациклическому виду позволяет значительно ускорить эту проверку.

Таким образом, затраты времени на включение одного подразделения в другое и исключение одного подразделения из другого оказываются приемлемыми.

Заметим, что в ряде других информационных систем, для которых целесообразно использовать реляционную модель ЛРД, транзитивные отношения между объектами в большинстве случаев также имеют древовидную структуру. Например, ряд свободно распространяемых программ интернет-форумов допускают неограниченное число уровней вложенности подфорумов, однако, каждый форум может являться непосредственным подфорумом не более чем одного другого форума.

Направления дальнейшего развития модели разграничения доступа

К потенциально возможным направлениям дальнейшего развития модели разграничения доступа в Системе можно отнести перечисленные далее.

- Назначение сеансам отдельного списка прав доступа, возможно, не совпадающего со списком прав доступа пользователя, владеющего сеансом. Эта возможность может быть использована для того, чтобы при наличии оснований, считать вход пользователя в Систему подозрительным и ограничить его в праве на действия, которые могут представлять наибольшую угрозу информационной безопасности Системы.
- Анализ и оценка выразительных возможностей реляционной модели ЛРД, которая реализуется механизмами разграничения доступа свободно распространяемого программного обеспечения систем управления содержимым сайтов, ориентированных на изменяющееся множество пользователей и возможность назначать каждому пользователю свои права доступа для разных объектов – например, свободно распространяемыми движками интернет-форумов.
- Минимизация числа суперпользователей, обладающих неограниченным доступом к Системе.
- Оценка трудоемкости различных операций с Системой при изменении режимов её функционирования следующим образом – все виды разрешенных доступов к объектам системы для пользователя вычисляются при входе этого пользователя в Систему.
- Разработка программы для визуального редактирования правил переноса отношений.

Работа с сеансами доступа

Одним из вариантов дальнейшего развития реляционной модели ЛРД является включение в нее механизмов, позволяющих ограничить привилегии пользователя при наличии обоснованных подозрений, что от имени этого пользователя в Системе действует злоумышленник. Основанием для таких подозрений может быть, например, работа пользователя в системе одновременно с двух разных IP-адресов. Для пользователей, обладающих в Системе максимальными правами доступа (суперпользователь), возможно, имеет смысл запретить использование привилегий, злоупотребление которыми может привести к наибольшему ущербу для Системы, если этот пользователь вошел в неё с IP-адреса, не входящего в специальный список доверенных IP-адресов.

Для пользователя, от имени которого в Системе зафиксирована подозрительная активность, вводится специальная роль *suspected_role*, для которой запрещается использование ряда привилегий. Например, для такого пользователя могут временно отменяться полномочия ответственного по подразделению и редактора ответственных по подразделению. На e-mail пользователя при этом отправляется сообщение с просьбой подтвердить свою ответственность за действия, которые Система считает подозрительными, либо изменить свой пароль. После выполнения одного из указанных действий, роль *suspected_role* для этого пользователя отменяется.

Возможно, для более эффективного отслеживания подозрительной активности пользователей в Систему следует добавить новый класс объектов – IP-адрес. Одновременно с ним в Систему следует добавить отношение *session_from_ip*, связывающее его с сеансом. В рамках класса IP-адресов при этом возможно выделить следующие три типа:

- «черный» IP-адрес, с которого была зафиксирована попытка несанкционированного доступа к информации, и авторизованный вход в Систему с этого IP-адреса запрещается;
- «белый» IP-адрес, с которого возможен вход в Систему супер-

пользователей, обладающих в Системе неограниченными правами доступа;

- остальные IP-адреса, с которых возможен вход в Систему любых пользователей, однако суперпользователям запрещено использование некоторых из их привилегий.

Разработка программы для визуального редактирования правил переноса отношений

Визуализация множества правил переноса отношений позволит разработчикам Системы лучше понять их структуру и легко найти возможные нежелательные отношения, порождаемые другими отношениями. Для визуализации взаимосвязи между отношениями Системы предлагается разработать специальную программу, выполняющую следующие функции:

- визуализировать структуру отношений между объектами в виде графа переноса отношений или его подграфов;
- визуализировать структуру отношений между объектами в виде графа, вершинами которого являются классы объектов, а ребрами – отношения, которые могут связывать объекты соответствующих классов;
- визуализировать иерархию классов объектов;
- изменять правила переноса отношений между объектами;
- работать с файлами описания модели в формате, совместимом с форматом файлов других программ для работы с реляционной моделью ЛРД.

В настоящее время планируется, что основная часть исходного кода такой программы будет написана на языке C++ с использованием библиотеки Qt для графического интерфейса пользователя.

5.1.4. Требования к подсистеме логического разграничения доступа к данным на основе реляционной модели и её основные атрибуты

С учётом представленных выше основных положений реляционной модели ЛРД далее сформулированы общие требования к её программной реализации, а также описаны основные её атрибуты.

Общие требования

1. Система должна проверять возможность выполнения пользователем указанной операции или множества операций с конкретным объектом.
2. В Системе должна присутствовать возможность получения списка всех операций, которые может выполнять заданный пользователь с заданным объектом.
3. Изменение прав доступа не должно требовать перекомпиляции кода.
4. Пользователь может обладать различными ролями в различных организациях или подразделениях.
5. Принадлежность пользователя роли может быть рекурсивной по структуре организации.
6. Правила доступа к индивидуальным объектам могут выдаваться как роли, так и отдельным пользователям.
7. Для задания принадлежности пользователя роли должен быть разработан простой веб-интерфейс.
8. Права на выполнение операции зависят от данных, которые хранятся в базе.
9. В случае, когда производится модификация данных, необходимо иметь возможность определять допустимость операции в зависимости от новых значений.

10. Список объектов и операций, которые можно выполнять, должен быть расширяемым.
11. Каждая организация должна иметь возможность формирования собственной политики для доступа к «своим» объектам.
12. Должна присутствовать возможность проверки непротиворечивости и корректности заданного набора правил доступа.
13. Права на выдачу прав должны прописываться в правилах.
14. Система должна работать при загрузке около 100 запросов в секунду.

Доступ к персональным данным пользователя

Одной из важных подзадач защиты информационных систем является защита персональных данных пользователей этих систем. С каждым пользователем в ИАС «ИСТИНА» ассоциированы следующие данные:

- ФИО;
- история мест работы;
- информация об ученых степенях и званиях;
- фотография;
- адрес электронной почты;
- пароль;
- информация о деятельности (конкретный человек является автором некоторой публикации, руководителем работы и так далее).

Заметим, что вся перечисленная информация, кроме адреса электронной почты и пароля, ассоциирована не с объектом «пользователь», а с объектом «сотрудник», с которым не обязательно должен

быть связан пользователь Системы. Таким образом, ситуация усложняется тем фактом, что информация о деятельности сотрудника фактически может быть опубликована без его ведома при добавлении информации о его публикациях его соавторами. Одним из возможных вариантов ограничения доступа к подобной информации может быть использование предусмотренного реляционной моделью ЛРД поля типа объекта, а именно – введение следующих типов для объекта класса «сотрудник»:

- сотрудник не связан с пользователем;
- сотрудник связан с пользователем, но не сообщил об отношении к публикации своих персональных данных;
- сотрудник связан с пользователем и подтвердил согласие на обработку всех результатов своей деятельности;
- сотрудник связан с пользователем и потребовал ограничить доступ к некоторым категориям своих данных.

Списки возможных видов доступа для каждого класса

К объектам класса «пользователь» допускаются следующие виды доступа:

- связать с определенным сотрудником;
- просмотреть адрес электронной почты;
- изменить адрес электронной почты;
- просмотреть пароль;
- изменить пароль;
- сделать ответственным по подразделению;
- отозвать полномочия ответственного по подразделению;
- сделать редактором ответственных по подразделению;
- отозвать полномочия редактора ответственных по подразделению;

5.1. МОДЕЛИ ЛОГИЧЕСКОГО РАЗГРАНИЧЕНИЯ ДОСТУПА

- отправить сообщение на электронный адрес пользователя;
- разрешить некоторые виды доступа к данным определенного сотрудника, не разрешенные для других пользователей.

Последний вид доступа из представленного списка можно также назвать «связать отношением доверия с определенным сотрудником».

Виды доступа к объектам класса «сотрудник»:

- связать с определенным пользователем;
- включить в число предполагаемых авторов определенного результата научно-исследовательской деятельности;
- исключить из числа авторов предполагаемых определенного результата научно-исследовательской деятельности;
- включить в число подтвержденных авторов определенного результата научно-исследовательской деятельности;
- исключить из числа авторов подтвержденных определенного результата научно-исследовательской деятельности;
- просмотреть информацию о степенях и званиях;
- просмотреть полный список работ;
- изменить информацию о степенях и званиях;
- включить в определенное подразделение;
- исключить из определенного подразделения;
- установить/снять ограничения на чтение данных о деятельности и местах работы;
- объединить с другим сотрудником;
- разрешить какому-либо пользователю некоторые виды доступа к данным сотрудника, запрещенные для других пользователей.

Виды доступа к объектам класса «подразделение»:

- включить сотрудника в подразделение;
- исключить сотрудника из подразделения;
- включить дочернее подразделение;
- включить подразделение в другое в качестве дочернего;
- включить пользователя в число ответственных по подразделению;
- исключить пользователя из числа ответственных по подразделению;
- включить пользователя в число редакторов ответственных по подразделению;
- исключить пользователя из числа редакторов ответственных по подразделению;
- просмотреть список всех сотрудников подразделения;
- просмотреть список всех публикаций сотрудников подразделения.

Виды доступа к объектам класса «результат научно-исследовательской деятельности»:

- включить сотрудника в число предполагаемых авторов;
- исключить сотрудника из числа предполагаемых авторов;
- включить сотрудника в число подтвержденных авторов;
- исключить сотрудника из числа подтвержденных авторов;
- удалить результат.

Виды доступа к объектам класса «публикация»:

- скачать полный текст;
- загрузить в Систему полный текст;
- изменить привязку к журналу или сборнику;

- изменить название.

Отметим, что класс «публикация» является подклассом класса «результат научно-исследовательской деятельности», и все виды доступа, допустимые для результатов научно-исследовательской деятельности, допустимы и для публикаций.

Виды доступа к объектам класса «журнал»:

- изменить название;
- просмотреть список всех публикаций;
- включить публикацию;
- исключить публикацию;
- добавить связанный объект «членство в редколлегии журнала»;
- удалить связанный объект «членство в редколлегии журнала».

Разрешение конфликтов между политиками информационной безопасности разных организаций

Одним из требований, предъявляемых к модели ЛРД ИАС «ИСТИНА» является возможность каждой из зарегистрированных в Системе организаций формировать собственную политику безопасности для доступа к «своим» объектам. Следует заметить, однако, что некоторые из объектов Системы нельзя однозначно отнести к определенной организации. Один сотрудник может работать в нескольких организациях, а также один результат научно-исследовательской деятельности может иметь соавторов из разных организаций. Одним из требований, предъявляемых к модели ЛРД ИАС «ИСТИНА» является автоматическое разрешение конфликтов при наличии у одного результата научно-исследовательской деятельности соавторов из разных организаций.

Как отмечено выше, некоторые из операций в Системе изменяют атрибуты одновременно нескольких объектов, и такие операции требуют наличия у пользователя соответствующих прав доступа ко всем

этим объектам. Данное свойство модели в некоторой степени обеспечивает разрешение конфликтов при наличии у результата научно-исследовательской деятельности соавторов в разных организациях. Рассмотрим пример. Пусть некоторая публикация имеет несколько соавторов из двух разных организаций *A* и *B*. Пусть необходимо добавить к числу авторов этой публикации одного из сотрудников подразделения *A*, либо исключить из числа авторов публикации одного из сотрудников подразделения *A*. В обоих случаях данная операция требует наличия прав доступа не только к объекту «публикация», но и к объекту «сотрудник». Поскольку сотрудник числится в организации *A*, такие права доступа есть только у пользователей, имеющих административные полномочия в организации *A*. Таким образом, если публикация имеет соавторов в двух разных организациях, то пользователь, имеющий полномочия по администрированию данных одной из этих организаций, может изменять атрибуты публикации, определяющие ее привязку к сотрудникам своей организации, но не может изменять ее привязку к сотрудникам другой организации.

Разграничение прав доступа на изменение других атрибутов публикации, например, ее названия, является более сложной задачей. Принимая во внимание эволюционное (поэтапное) развитие механизмов и подсистемы разграничения доступа к ресурсам, которая используется в Системе, представляет интерес:

- сравнение отдельных моделей ЛРД, оценка полноты покрытия ими перечисленных выше требований и выразительных свойств;
- оценка вычислительной производительности Системы при выполнении правил разграничения доступа в условиях близких к реальным.

5.1.5. Выразительные свойства реляционной модели

Рассмотрим выразительные свойства реляционной модели ЛРД - самой «критичной» по вычислительной производительности при вы-

полнении правил разграничения доступа к ресурсам Системы.

Сравнение реляционной модели с дискреционной моделью UNIX-подобных ОС

Определение 15. Пусть в информационной системе заданы следующие множества:

- $Users$ – множество файлов;
- $Groups$ – множество файлов;
- $Files$ – множество файлов;
- $Processes$ – множество процессов;

следующие далее отображения:

- $Owner : Files \sqcup Processes \rightarrow Users$, ставящее в соответствие каждому файлу или процессу пользователя-владельца;
- $OwnerGroup : Files \sqcup Processes \rightarrow Users$, ставящее в соответствие каждому файлу или процессу некоторую выделенную группу пользователей;
- $AccessMode : Files \rightarrow Z_2^9$, определяющее права доступа на чтение, запись или исполнение файла для его владельца, пользователей, входящих в выделенную группу и остальных пользователей;

а также отношение

- $UserInGroup \subset Users \times Groups$, определяющее принадлежность пользователя группе;

и выделенный пользователь $u_0 \in Users$, называемый суперпользователем.

Тогда будем считать, что в системе задана дискреционная модель ЛРД UNIX-подобных ОС.

Данная модель ЛРД используется, в частности, в операционной системе Linux. Для файлов возможно три вида доступа: чтение, запись и исполнение. Для файла f отображение $AccessMode(f)$ ставит в соответствие каждому файлу набор из девяти бит, первые три из которых определяют права доступа к файлу для пользователя $Owner(f)$. Следующие три бита определяют разрешенные права доступа к файлу для пользователей, являющихся членами группы-владельца файла, то есть таких пользователей u , что $u \neq Owner(f)$ и $(u, OwnerGroup(f)) \in UserInGroup$. Последние три бита определяют права доступа к файлу всех остальных пользователей. При этом выделенный пользователь u_0 имеет все права доступа ко всем объектам вне зависимости от их меток безопасности.

Теорема 6. *Любая дискреционная модель ЛРД UNIX-подобных ОС может быть выражена в терминах реляционной модели ЛРД.*

Доказательство. Определим базовые множества реляционной модели ЛРД следующим образом:

- $Objects = Users \sqcup Groups \sqcup Files \sqcup Processes,$
- $Classes = \{user, group, file, process\},$
- $Types = Z_2^9,$
- $Actions = \{read, write, execute, kill, change_user_group, delete, chown, chmod\},$
- $Relations = \{owner_user, owner_group, user_in_group, user_in_owner_group\},$
- $Roles = \{root, normal_user\},$
- $Privileges = \{create_user, create_group\}.$

Виды доступа $read$ и $write$ соответствуют чтению или записи файла, $execute$ соответствует запуску программы в исполнимом файле. Доступ $kill$ означает возможность послать сигнал процессу, которой обладает пользователь-владелец процесса, либо суперпользователь. Доступ $change_user_group$ означает включение или исклю-

чение пользователя из группы. Доступ *delete* – это удаление пользователей или групп, которое разрешено только суперпользователю. Заметим, что удаление файлов, как и их создание, регламентируется как доступ на запись к каталогу, содержащему этот файл. Доступ *chown* означает изменение владельца файла, доступ *chmod* – изменение его типа.

Определим базовые отношения реляционной модели ЛРД:

- *ObjectRelations* =
 $\{u, user_in_group, g : u \in Users, g \in Groups, (u, g) \in UserInGroup\}$
 $\sqcup \{(u, owner_user, f) : u \in Users, f \in Files, u = Owner(f)\}$
 $\sqcup \{(g, owner_group, f) : u \in Users, f \in Files, g = OwnerGroup(f)\}$
 $\sqcup \{(u, owner_user, p) : u \in Users, p \in Processes, u = Owner(p)\}$
 $\sqcup \{(g, owner_group, p) : u \in Users, p \in Processes, g = OwnerGroup(p)\},$
- *UserRoles* = $\{(u_0, root)\} \sqcup \{(u, normal_user) : u \in Users\},$
- *RolePrivileges* = $\{(root, create_user), (root, create_group)\},$
- *ClassAccess* =
 $\{(file, read), (file, write), (file, execute), (process, kill)\}$
 $\cup \{(user, change_user_group), (group, change_user_group), (user, delete)\}$
 $\cup \{(group, delete), (file, chmod), (file, chown)\},$
- *RelationLeftClasses* = $\{(owner_user, user), (owner_group, group), (user_in_group, user)\} \cup \{(user_in_owner_group, user)\},$
- *RelationRightClasses* =
 $\{(owner_user, file), (owner_group, file), (user_in_group, group)\}$
 $\cup \{(owner_group, process), (owner_user, process)\}$
 $\cup \{(user_in_owner_group, file), (user_in_owner_group, process)\},$

- $RelationTransitions = \{(user_in_group, owner_group, user_in_owner_group)\}$.

Множество $RelationTypeAccessGranted$ является объединением следующих подмножеств:

- $\{(owner_user, t, read)\}$ для всех $t \in Z_2^9$, первая координата которых равна 1,
- $\{(owner_user, t, write)\}$ для всех $t \in Z_2^9$, вторая координата которых равна 1,
- $\{(owner_user, t, execute)\}$ для всех $t \in Z_2^9$, третья координата которых равна 1,
- $\{(user_in_owner_group, t, read)\}$ для всех $t \in Z_2^9$, четвертая координата которых равна 1,
- $\{(user_in_owner_group, t, write)\}$ для всех $t \in Z_2^9$, пятая координата которых равна 1,
- $\{(user_in_owner_group, t, execute)\}$ для всех $t \in Z_2^9$, шестая координата которых равна 1,
- $\{(owner_user, t, kill)\}$ для всех типов t .
- $\{(owner_user, t, chmod)\}$ для всех типов t .
- $\{(owner_user, t, chown)\}$ для всех типов t .

Множество $RelationTypeAccessDenied$ является объединением следующих подмножеств:

- $\{(owner_user, t, read)\}$ для всех $t \in Z_2^9$, первая координата которых равна 0,
- $\{(owner_user, t, write)\}$ для всех $t \in Z_2^9$, вторая координата которых равна 0,
- $\{(owner_user, t, execute)\}$ для всех $t \in Z_2^9$, третья координата которых равна 0,

- $\{(user_in_owner_group, t, read)\}$ для всех $t \in Z_2^9$, четвертая координата которых равна 0,
- $\{(user_in_owner_group, t, write)\}$ для всех $t \in Z_2^9$, пятая координата которых равна 0,
- $\{(user_in_owner_group, t, execute)\}$ для всех $t \in Z_2^9$, шестая координата которых равна 0.

Множество *RoleTypeAccessGranted* является объединением следующих подмножеств:

- $\{(normal_user, t, read)\}$ для всех $t \in Z_2^9$, седьмая координата которых равна 1,
- $\{(normal_user, t, write)\}$ для всех $t \in Z_2^9$, восьмая координата которых равна 1,
- $\{(normal_user, t, execute)\}$ для всех $t \in Z_2^9$, девятая координата которых равна 1,
- $\bigcup_{t \in Types, a \in Actions} \{(root, t, a)\}$.

Таким образом, в Системе построена реляционная модель ЛРД, дающая каждому пользователю такие же права доступа ко всем объектам, как и исходная дискреционная модель ЛРД.

□

5.1.6. Алгоритмы, необходимые для работы макета

Как отмечено в определении реляционной модели ЛРД, пользователь u имеет право доступа a к объекту o класса c и типа t в том и только том случае, если a является разрешенным действием для объектов класса c и выполнено одно из следующих условий:

- $\exists r \in Roles : (u, r) \in UserRoles, (r, t, a) \in RoleTypeAccessGranted;$
- $\exists r_1, r_2 \in Roles : r_1 \leq r_2, (u, r_2) \in UserRoles, (r_1, t, a) \in RoleTypeAccessGranted;$

- $\exists r \in Relations : (u, r, o) \in ObjectRelations, (r, t, a) \in RelationTypeAccessGranted;$
- $\exists r_1, r_2 \in Relations : r_1 \leq r_2, (u, r_2, o) \in ObjectRelations, (r_1, t, a) \in RelationTypeAccessGranted;$
- $\exists r_{12}, r_{23}, r_{13} \in Relations, \exists o_2 \in O : (u, r_{12}, o_2) \in ObjectRelations, (o_2, r_{23}, o) \in ObjectRelations, (r_{13}, t, a) \in RelationTypeAccessGranted, (r_{12}, r_{23}, r_{13}) \in RelationTransitions;$

и при этом не выполнено ни одно из следующих условий:

- $\exists r \in Roles : (u, r) \in UserRoles, (r, t, a) \in RoleTypeAccessDenied;$
- $\exists r_1, r_2 \in Roles : r_1 \leq r_2, (u, r_2) \in UserRoles, (r_1, t, a) \in RoleTypeAccessDenied;$
- $\exists r \in Relations : (u, r, o) \in ObjectRelations, (r, t, a) \in RelationTypeAccessDenied;$
- $\exists r_1, r_2 \in Relations : r_1 \leq r_2, (u, r_2, o) \in ObjectRelations, (r_1, t, a) \in RelationTypeAccessDenied;$
- $\exists r_{12}, r_{23}, r_{13} \in Relations, \exists o_2 \in O : (u, r_{12}, o_2) \in ObjectRelations, (o_2, r_{23}, o) \in ObjectRelations, (r_{13}, t, a) \in RelationTypeAccessDenied, (r_{12}, r_{23}, r_{13}) \in RelationTransitions;$

и при этом правила, связанные с использованием множества *RelationTransitions*, могут применяться по цепочке. С помощью определений, представленных в 5.1.3, эти правила можно переформулировать следующим образом. Пользователь *u* имеет право доступа *a* к объекту *o* класса *c* и типа *t* в том и только том случае, если *a* является разрешенным действием для объектов класса *c* и выполнено одно из следующих условий:

- $\exists r \in Roles : (u, r) \in UserRoles, (r, t, a) \in RoleTypeAccessGranted;$

- $\exists r_1, r_2 \in Roles : r_1 \leq r_2, (u, r_2) \in UserRoles, (r_1, t, a) \in RoleTypeAccessGranted;$
- $\exists r \in Relations : (u, r, o) \in ObjectRelations, (r, t, a) \in RelationTypeAccessGranted;$
- $\exists r, r_1, r_2 \in Relations : r_1 \leq r_2, (u, r_2, o) \in ObjectRelations, (r_1, t, a) \in RelationTypeAccessGranted;$
- $\exists r_{1,2}, r_{2,3}, \dots, r_{n-1,n} \in Relations, \exists u = o_1, o_2, \dots, o_{n-1}, o_n = o \in O : (u, r_{12}, o_2) \in ObjectRelations : \forall j (o_j, r_{j,j+1}, o_{j+1}) \in ObjectRelations, (r, t, a) \in RelationTypeAccessGranted$
и отношение r порождается цепочкой отношений $r_{1,2}, \dots, r_{n-1,n}$.

и при этом не выполнено ни одно из следующих условий:

- $\exists r \in Roles : (u, r) \in UserRoles, (r, t, a) \in RoleTypeAccessDenied;$
- $\exists r_1, r_2 \in Roles : r_1 \leq r_2, (u, r_2) \in UserRoles, (r_1, t, a) \in RoleTypeAccessDenied;$
- $\exists r \in Relations : (u, r, o) \in ObjectRelations, (r, t, a) \in RelationTypeAccessDenied;$
- $\exists r_1, r_2 \in Relations : r_1 \leq r_2, (u, r_2, o) \in ObjectRelations, (r_1, t, a) \in RelationTypeAccessDenied;$
- $\exists r_{12}, r_{23}, r_{13} \in Relations, \exists o_2 \in O : (u, r_{12}, o_2) \in ObjectRelations, (o_2, r_{23}, o) \in ObjectRelations, (r_{13}, t, a) \in RelationTypeAccessDenied, (r_{12}, r_{23}, r_{13}) \in RelationTransitions;$

В первой версии предлагается временно предположить отсутствие запрещающих правил. Тогда приблизительный алгоритм получения всех видов доступа, разрешенных для пользователя u к объекту o , (такой алгоритм, согласно требованию 2, представленному в 5.1.4, должен быть реализован) будет выглядеть следующим образом:

```
список_доступов = ()
```

Для всех цепочек отношений R , которые могут соединять объект u с объектом o :

```
    проверить, соединяет ли цепочка  $R$  объекты  $u$  и  $o$ .
```

```
    если ответ отрицательный, перейти к следующей цепочке
```

```
     $r$  = отношение, порождаемое цепочкой  $R$ 
```

```
    для все отношений  $r' \leq r$ 
```

```
        список_доступов += список_доступов_отн_к_объекту ( $r', o$ )
```

```
    вернуть список_доступов
```

Замечание: цепочка отношений называется корректной, если она порождает какое-либо отношение.

Представленный алгоритм включает в себя два других алгоритма, один из которых должен получать список всех корректных цепочек отношений, которые могут соединять объекты u и o , исходя из их классов, а другой – проверять, соединяет ли данная цепочка конкретно объекты u и o .

Предположим, что множество *Relation_Transitions* является ациклическим. Алгоритм построения множества всех цепочек, которые могут соединять объект u с объектом o , исходя из их классов, выглядит следующим образом:

```
Функция все_цепочки (исходные_классы, конечные_классы,
наибольшая_длина)
```

```
    результат = пустое_множество_цепочек
```

```
    Для каждого отношения  $r$ :
```

```
        Если левые_классы( $r$ ) не пересекаются
        с исходными_классами
```

```
            перейти к следующему  $r$ 
```

```
         $R$  = все_цепочки (правые_классы( $r$ ),
                        конечные_классы,
                        наибольшая_длина - 1)
```

```
        для каждой цепочки ( $r_1, \dots, r_n$ ) из  $R$ 
```

```
            результат += ( $r, r_1, \dots, r_n$ )
```

```
    вернуть результат
```

Соответственно, при первом вызове алгоритма множество исходных классов должно состоять из класса «пользователь» и его потомков, множество конечных классов – из класса целевого объекта и его предков.

Заметим, что этот алгоритм оперирует только статическими множествами и его можно вычислять при загрузке политики безопасности Системы, тем или иным способом сохраняя результат. Нет необходимости повторять эти вычисления при каждом получении списка разрешенных доступов определенного пользователя к определенному объекту. Для упрощения алгоритма в первой версии следует потребовать, чтобы все множества левых и правых классов каждого отношения состояли из одного главного класса и его классов-потомков.

5.1.7. Результаты тестовых испытаний

Программная реализация реляционной модели ЛРД прошла экспериментальную проверку на быстроедействие при выполнении правил разграничения доступа к данным. Как следует из гистограмм, представленных на следующих далее рис. 5.3 – 5.5, получены хорошие результаты в условиях максимально близких к реальным. В настоящее время программная реализация новой модели готовится к внедрению в системе ИСТИНА.

Были проведены эксперименты для трёх различных СУБД: Oracle, PostgreSQL и MySQL. В каждом случае в базе содержится следующее число записей:

- 22 тыс. пользователей, зарегистрированных в системе («User»);
- 1.5 тыс. подразделений («Subdivision»);
- 300 тыс. научных достижений («Achievement»);
- 15 тыс. авторов («Author»);
- 1.5 тыс. записей в таблице отношения «Пользователь управляет подразделением» («Management»)
- 7.5 тыс. записей в таблице отношения «Пользователь работает в некотором подразделении» («Employment»)
- 600 тыс. записей в таблице отношения «Пользователь является автором научного достижения» («Authorship», «Author of Achievement»)

Результаты экспериментов для каждой СУБД представлены ниже.

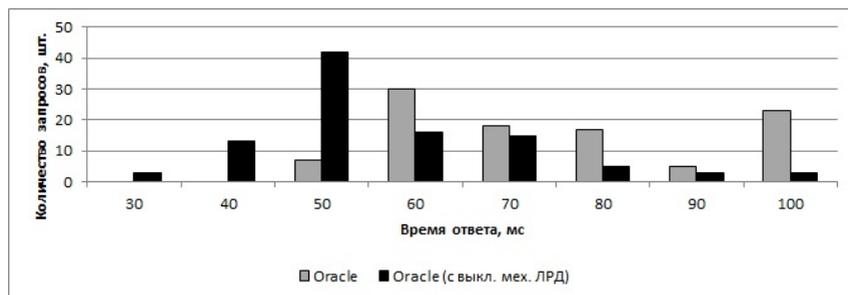


Рис. 5.3. Распределение времени ответа для СУБД Oracle Database Express Edition 11g Release 2.

Среднее время ответа для страницы, на которой осуществляется ЛРД, для СУБД Oracle:

- с включенными механизмами ЛРД – 76.9 мс;
- с выключенными механизмами ЛРД – 56.6 мс.

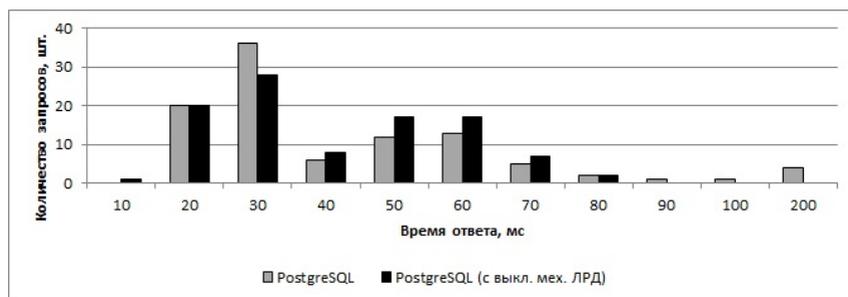


Рис. 5.4. Распределение времени ответа для PostgreSQL 9.3.

Среднее время ответа для страницы, на которой осуществляется ЛРД, для СУБД PostgreSQL:

- с включенными механизмами ЛРД – 46 мс;
- с выключенными механизмами ЛРД – 40.7 мс.

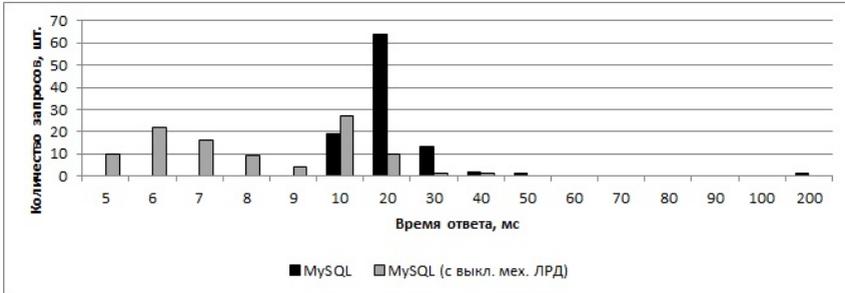


Рис. 5.5. Распределение времени ответа для MySQL 5.6.

Среднее время ответа для страницы, на которой осуществляется ЛРД, для СУБД MySQL:

- с включенными механизмами ЛРД – 21.4 мс;
- с выключенными механизмами ЛРД – 9.5 мс.

На основе полученных значений можно сделать вывод, что проверка правил для реляционной модели требует небольших затрат времени (разница между временем обработки страницы с включенными и выключенными механизмами ЛРД составляет от 6 до 20 мс). Следовательно, модель способна выполнять задачи по разграничению доступа в реальных системах, не вызывая их «зависания» при обработке запросов на доступ.

5.2. Механизмы защиты, встроенные в Django

В настоящем разделе представлены механизмы защиты, встроенные в инструментарий Django, которые в настоящее время используются в системе ИСТИНА. Они позволяют защитить данные от реализации достаточно широкого спектра известных угроз. В качестве источника используется официальная документация Django¹.

¹<https://docs.djangoproject.com/en/1.4/topics/security/>

5.2.1. Защита от межсайтового скриптинга

Атаки типа XSS (Cross Site Scripting) позволяют пользователю вставить собственные JS-скрипты в браузеры других пользователей. Эта цель достигается с помощью сохранения вредоносных скриптов в базе данных. Такие скрипты затем запрашиваются и отображаются браузерами других пользователей, или путём косвенного (опосредованного) принуждения пользователя нажать на ссылку, которая позволит вредоносному скрипту выполниться в браузере пользователя. Более того, XSS-атаки могут исходить из любого недоверенного источника данных, такого как cookie или web-сервисы, в том случае, когда данные были недостаточно очищены перед их размещением на странице.

Использование в системе ИСТИНА шаблонов Django защищает её от большинства XSS-атак. Далее, покажем на простых примерах, как именно работает механизм защиты.

При создании HTML-файла с использованием шаблонов, появляется риск того, что переменная может содержать символы, которые повлияют на структуру полученного файла. Например, пусть задан такой фрагмент:

```
Hello, {{ name }}.
```

На первый взгляд это кажется безобидным способом отображения имени пользователя. Однако заметим, что может произойти, если пользователь выбрал такое имя:

```
<script>alert('hello')</script>
```

Шаблон с этим именем вернет:

```
Hello, <script>alert('hello')</script>
```

что приведет к отображению alert-окна JavaScript.

Аналогично, если имя содержит символ '<':

```
<b>username,
```

то шаблон вернет содержимое:

```
Hello, <b>username
```

В результате оставшееся содержимое страницы будет выделено полужирным шрифтом. Очевидно, пользовательскими данными нельзя «слепо» доверять и вставлять непосредственно в содержимое страницы, так как злоумышленники могут этим воспользоваться.

Чтобы избежать отмеченных выше вопросов, в системе ИСТИНА используются два варианта действий.

- Ко всем сомнительным переменным применяется фильтр `escape`, который преобразует потенциально опасные HTML-символы в безопасные.
- Содержимое HTML-файлов экранируется автоматически.

В большинстве случаев в системе ИСТИНА используется вариант, при котором применяется автоматическое экранирование, поэтому рассмотрим его более подробно.

По умолчанию в Django каждый шаблон экранирует все переменные. В частности выполняются такие замены:

- `<` заменяется на `<`;
- `>` заменяется на `>`;
- `'` (одинарная кавычка) заменяется на `'`;
- `”`(двойная кавычка) заменяется на `"`;
- `&` заменяется на `&`.

Тем не менее, хотя экранирование и защищает пользователя от большинства видов вредоносного ввода, оно не является панацеей. Например, оно не защитит от инъекции такого кода:

```
<style class={{ var }}>...</style>
```

Если `var` содержит `'class1 onmouseover=javascript:func()'`, то это может привести к неавторизованному запуску JavaScript. Здесь всё зависит от того, как браузер интерпретирует несовершенный HTML-файл.

5.2.2. Защита от подделки межсайтового запроса

Атаки типа CSRF (Cross Site Request Forgery) позволяют недобросовестному пользователю выполнять действия от имени другого пользователя, без ведома последнего или без его согласия. Инструментарий Django обладает встроенной защитой против большинства типов CSRF-атак. В системе ИСТИНА эти механизмы также активированы и используются. Однако, как это обычно бывает, существуют ограничения. Например, есть возможность отключить защиту CSRF глобально или на уровне отдельного представления. Такое можно делать только в том случае, если существует полная уверенность в выполняемых действиях.

Защита от CSRF-атак работает, проверяя метку с текущим временем в каждом POST-запросе. Она не позволяет злоумышленнику просто сформировать POST-запрос формы к сайту системы ИСТИНА и создавать, тем самым, возможность другому авторизованному пользователю возможно без злого умысла отправить эту форму.

5.2.3. Защита от внедрения SQL

Внедрение SQL — это тип атаки, при которой недобросовестный пользователь имеет возможность выполнить в базе данных определённый SQL-запрос. Результатом выполнения такого запроса может быть удаление или даже утечка данных.

Поскольку в системе ИСТИНА используется Django ORM, то созданный SQL-запрос будет правильно экранирован соответствующим драйвером базы данных. Однако Django предоставляет разработчикам возможность писать запросы напрямую или выполнять собственные запросы. Эти возможности также используются

в Системе там, где это необходимо, причём экранируются все параметры, которые предоставлены пользователем.

5.2.4. Защита от скликивания

Скликивание (clickjacking) представляет собой тип атаки, при которой злоумышленник может получить доступ к конфиденциальной информации или даже получить доступ к компьютеру пользователя. Её принцип основан на том, что поверх видимой страницы располагается невидимый слой, в который и загружается нужная злоумышленнику страница. При этом элемент управления (кнопка, ссылка), необходимый для осуществления требуемого действия, совмещается с видимой ссылкой или кнопкой, нажатие на которую ожидается от пользователя.

В системе ИСТИНА используется предоставляемая Django защита от этой атаки в виде модуля X-Frame-Options, который, при использовании соответствующего браузера, предотвращает отображение сайта внутри фрейма.

5.2.5. Криптографическая подпись

Программное средство Django предоставляет низкоуровневое API для подписывания значения и высокоуровневое API для установки и чтения подписанных cookie. В системе ИСТИНА эта подпись может быть полезна для выполнения следующих действий:

- генерация URL для восстановления аккаунта пользователя, которые будут отправлены пользователю, потерявшему свой пароль;
- проверка целостности данных, скрытых в скрытом поле формы;
- генерация одноразового секретного URL для того, чтобы обеспечить временный доступ к защищённому ресурсу, например, на скачивание файла.

Защита SECRET_KEY

При создании в Django проекта системы ИСТИНА с помощью startproject автоматически был сгенерирован файл settings.py и определено случайное значение SECRET_KEY, которое является ключевым аспектом защиты подписанных данных.

Глава 6

Перспективы развития

В настоящее время система ИСТИНА используется в МГУ имени М.В. Ломоносова для сбора и систематизации, контроля и анализа результатов научной деятельности отдельных сотрудников, подразделений и университета в целом. Отметим, что она может применяться и для мониторинга состояния научных исследований на уровне большого числа организаций, например, в отдельном регионе или во всей стране.

Дальнейшее развитие Системы связано со следующими направлениями:

- учет новых видов результатов научной и педагогической деятельности, уточнение характеристик публикаций (например, географических карт, газетных публикаций);
- внедрение системы учета эффективности использования научного оборудования, в частности, установление соответствия между результатами научно-педагогической деятельности и научным оборудованием;
- дальнейшее расширение полномочий ответственных сотрудников от подразделений или организаций (методы выявления и исправления некорректно введенных данных, возможность создания отчетных форм и другие функции);

- внедрение методов интеллектуального анализа данных.

Наибольший научный интерес представляет задача интеллектуального анализа научно-технической информации, поскольку, в отличие от других перечисленных направлений развития Системы, она связана с решением сложных проблем в области обработки текстов на естественном языке. На основе методов онтологического моделирования и анализа текстов на естественном языке предполагается решение ряда практических задач, включая создание средств персонафицированного поиска, тематического анализа и выявления актуальных научных направлений. Эти задачи более подробно рассматриваются далее. Следует отметить, что при решении подобных задач в рамках системы ИСТИНА можно использовать большой объем данных, которые прошли ручную проверку и описывают взаимосвязанные объекты разных типов. Наличие такой информации должно способствовать повышению точности алгоритмов автоматического анализа.

6.1. Персонафицированный поиск

Отдельным большим направлением развития и сферой применения Системы является выполнение персонафицированного поиска с помощью поисковой автоматизированной системы тематического анализа информации (АСТАИ) [18], которая разрабатывается коллективом с участием авторов данной книги. В данном разделе представлен подход, позволяющий улучшать результаты обработки запросов поисковыми системами путем их интеграции со специализированными базами знаний, которые могут предоставлять развернутую информацию о персональных интересах пользователя.

Поисковая система АСТАИ позволяет осуществлять:

- разовый поиск информации;
- непрерывный сбор информации;
- группировку и предварительную обработку данных для последующего анализа;

- классификацию информации по древовидным классификаторам;
- контекстный поиск;
- персонафицированный поиск;
- индексацию документов из Интернет;
- индексацию аннотаций документов из внутреннего хранилища.

Далее кратко представлены основные особенности этой системы [19].

- Автоматическая рубрикация текстов с использованием неограниченного количества рубрикаторов позволяет определять тематику текстов и лучше описывать информационную потребность пользователя при построении поискового запроса.
- Наличие персональных рубрикаторов и архивов позволяет пользователю держать под рукой все необходимые ему документы, содержащиеся в хранилище и полученные путем сканирования Интернет, удобно располагать такие документы на экране и быстро получать к ним доступ. Кроме того, система АСТАИ может на основании персональных архивов корректировать результаты поиска в соответствии с реальной информационной потребностью пользователя.
- Сервисно-ориентированная архитектура (SOAP) позволяет проводить интеграцию поисковой системы с другими информационными системами, использующимися различными участниками работ. Это обстоятельство дает возможность проводить индексацию специализированных хранилищ (например, банка моделей, материалов или проектов), а также встраивать интерфейс поиска в привычную для участника среду разработки.
- Открытый код АСТАИ позволяет ее модифицировать, совершенствовать и подстраивать под задачи, отражающие специ-

фику деятельности отдельного коллектива. Использование такой системы при распределенной работе большого научного коллектива значительно упрощает процессы распределения и использования информации всеми участниками совместно выполняемых работ.

Традиционные полнотекстовые поисковые системы нацелены на реализацию универсального поиска, результаты которого не зависят от конкретной предметной области и не учитывают ее специфику. В первую очередь это объясняется отсутствием возможности быстро и адекватно определить интересы конкретного пользователя. Одной из возможных альтернатив традиционным решениям является интеграция поисковых систем со специализированными базами данных и знаний, которые могут предоставлять развернутую информацию о персональных интересах пользователя. Такой подход перспективен при построении специализированных поисковых систем, которые создаются как для управления данными, распределенными в хранилищах отдельных организаций и корпораций, так и для сбора и анализа данных по определенным предметным областям в сети Интернет. Острая потребность в таких системах стимулируется также и активным развитием технических средств хранения и обмена информации, которое наблюдается в последние годы.

Разработчики многих современных поисковых систем в настоящее время стараются поддерживать и развивать механизмы персонализированного поиска, позволяя системе подстраиваться под пользователя и учитывать его персональные интересы. Одним из таких механизмов, который, например, активно используется в Google, является фиксация переходов по ссылкам и анализ истории запросов, как конкретного пользователя, так и всех пользователей системы в целом. Однако, такой подход не позволяет в полной мере получать представление о профессиональных интересах пользователя и о наиболее значимых для него темах, поскольку содержит очень много «шума». В качестве основных проблемных вопросов на этом направлении можно назвать следующие.

- Переход пользователя на найденный документ («клик») озна-

чает только тот факт, что пользователь не отсеял этот документ по краткой аннотации на странице результатов поиска. Поисковая система не может оценить, представляет ли найденный документ интерес для пользователя. Практика показывает, что более 80% документов, на которые пользователь переходит при поиске, оказываются для него неинтересными. Как следствие, учет контекста этих документов не может значительно улучшить результаты последующего поиска.

- Использование истории запросов лучше отражает персональную заинтересованность пользователя, однако медленно адаптируется к его информационным потребностям, поскольку поисковая система не может оценить смысловую связь между запросами.

Одним из эффективных методов, который способен дать ответ на отмеченные вопросы, является интеграция поисковой системы с другими существующими в организации (корпорации) информационными приложениями, базами знаний и базами данных, описывающими интересы (предпочтения) отдельного пользователя. В качестве решения может рассматриваться отдельный модуль системы ИСТИНА, который на основе анализа научной деятельности пользователя строит его информационный портрет и передает это описание поисковой системе.

Использование описания пользователя для уточнения запросов

Обеспечение политики безопасности в АСТАИ и ИСТИНА осуществляется на основе совместного репозитория пользователей. Такой подход позволяет этим системам обмениваться данными, имеющими отношение к конкретному пользователю, в том числе – характеризующими его профессиональные интересы. Используя онтологию описания предметной области, система ИСТИНА способна обеспечить построение иерархии направлений, задач, алгоритмов и ключевых понятий, которые описывают интересы

конкретного пользователя. Первичной информацией для такого построения являются его научные работы, зарегистрированные в Системе, а также данные из Интернет. Описание передается в систему поиска в виде связанного графа классифицированных понятий. При осуществлении поиска пользователь может отметить пункт «Учитывать мои работы». После отметки этого пункта поисковая система производит расширение и уточнение запроса пользователя на основе полученного графа понятий. На первом этапе обработки поискового запроса оценивается близость этого запроса к одному из основных направлений работы пользователя. Такая оценка необходима, поскольку пользователь может являться автором работ сразу по нескольким научным направлениям, каждое из которых характеризуется собственными ключевыми понятиями. Далее запрос уточняется и расширяется на основе подграфа ключевых понятий направления, связанного с запросом.

Постановка задачи

Автоматическое уточнение запроса пользователя применяется для выделения из всех результатов выполнения запроса тех документов, которые представляют для него интерес. Ручное составление дополнительных подзапросов является очень трудоемким процессом и требует от пользователя высокой квалификации не только в его профессиональной области, но и в области теории информационного поиска. Автоматизация построения множества уточняющих и расширяющих слов возможна как на основе онтологий, так и на основе кластеризации [20], [21] и тезаурусов [22]. При поиске специалист ищет информацию по конкретному понятию в своей предметной области и, как правило, не хочет задумываться над его возможным значением в других областях. Например, если авиаконструктор ищет информацию по запросу «тормозные диски» в своей области интересов, то его запрос необходимо автоматически уточнять ключевыми понятиями предметной области «тормозные диски» AND («самолет» OR «авиация» OR ...).

После проведения поиска по ключевым словам должна

осуществляться классификация полученного массива документов и оценка их близости классу интересов пользователя. В результате на первой же странице результатов поиска пользователю будут показаны нужные документы, и ему не придется просматривать десятки не представляющих для него интереса страниц с описанием автомобильных и прочих дисков.

Расширение запроса

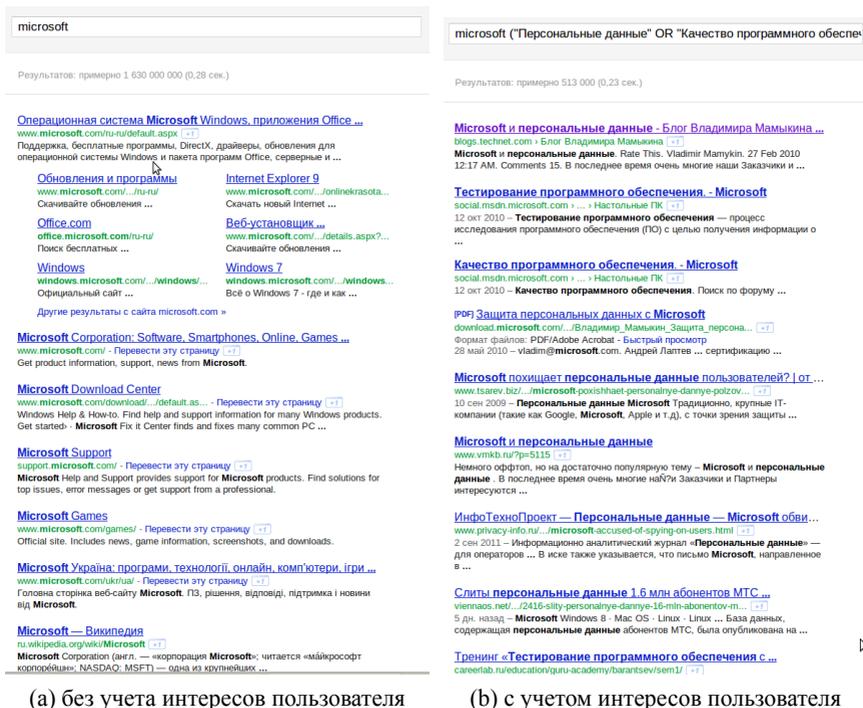
Подход, основанный на расширении запросов, используется для устранения трудностей, обусловленных наличием синонимов и близких понятий в языке, с целью улучшения полноты поиска. В АСТАИ для расширения запроса предлагается использовать информацию о связях слов запроса с ключевыми понятиями, которые могут быть получены от системы ИСТИНА.

Процесс расширения подобен процессу расширения по тезаурусам. Каждое слово дополняется списком связанных с ним дочерних понятий с учетом типа, степени и направления связи. Например, запрос «алгоритм RSA» будет дополнен понятиями «теорема Эйлера», «открытый ключ» и другими. Такое дополнение позволит найти документы, которые посвящены обсуждению связанных с алгоритмом RSA вопросов, однако не содержат его явного упоминания.

Для апробации предложенного подхода использовалось описание пользователя, построенное в системе ИСТИНА по конкретному пользователю по его научным статьям. Информационный портрет пользователя описывался понятиями: «Персональные данные»; «Качество программного обеспечения»; «Метрика программного обеспечения»; «Тестирование программного обеспечения». Сравнение результатов выдачи по запросам «microsoft» и «microsoft AND («Персональные данные» OR «Качество программного обеспечения» OR «Метрика программного обеспечения» OR «Тестирование программного обеспечения»)» показывает, что в результатах первого запроса представляющий интерес для пользователя документ находился только на семнадцатой позиции, а в результатах второго запроса только один документ из первых

ГЛАВА 6. ПЕРСПЕКТИВЫ РАЗВИТИЯ

двадцати не соответствовал интересам пользователя (рис. 6.1а, 6.1б).



(а) без учета интересов пользователя

(б) с учетом интересов пользователя

Рис. 6.1. Результаты запроса «microsoft» в Google.

На рис. 6.2 представлена общая схема интеграции системы ИСТИНА и поисковой системы в целях расширения и уточнения запросов.

Таким образом, представленный подход на основе интеграции поисковой системы с системами описания области интересов пользователя для автоматической или автоматизированной коррекции запроса пользователя может применяться для улучшения точности и полноты результатов поиска.

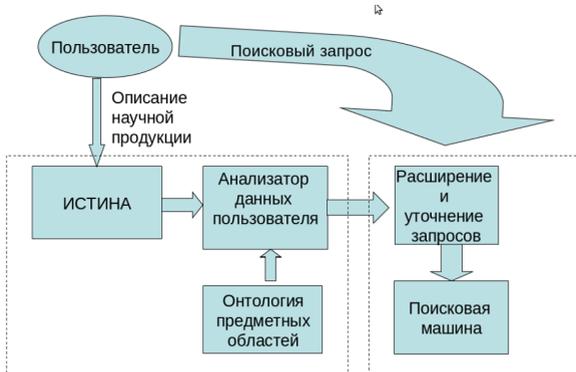


Рис. 6.2. Схема интеграции системы ИСТИНА и поисковой системы в целях расширения и уточнения запросов.

6.2. ОНТОЛОГИИ

Наиболее перспективным направлением развития системы ИСТИНА является дальнейшее использование онтологий предметных областей для различных целей. Основные сферы применения онтологий в Системе перечислены далее.

Описание типов результатов деятельности

Онтологии могут использоваться для описания типов результатов деятельности, которые позволяет учитывать Система. Это описание может использоваться не только для иллюстрации возможностей Системы. Представители подразделений научно-образовательной организации могут формализовать обладающие спецификой результаты деятельности, которые получают их сотрудниками. Формализованное в онтологии описание нового типа результатов деятельности может быть добавлено в Систему. Таким образом значительно облегчается взаимодействие между специалистами в различных предметных областях и разработчиками Системы, так как они могут изъясняться на одном языке онтологий. Отметим, что для задания онтологии не нужны навыки программирования, и

каркас онтологии может быть задан разработчиками Системы.

Формальное описание аналитических запросов

Для формального описания аналитических запросов, выполняемых к данным Системы, также можно применять онтологии. Первым шагом в этом направлении может быть процесс, аналогичный описанному ранее. Представители факультетов специфицируют запросы, которые им необходимо выполнить, и на основе их формального описания разработчики добавляют необходимые функции в Систему. Второй шаг состоит в полной автоматизации процесса. Систему можно дополнить модулем, который будет позволять выполнять произвольные (в рамках установленной политики безопасности) запросы к ее данным, специфицированные на языке онтологий. В этом случае этап взаимодействия с разработчиками опускается, и представители подразделений сами могут формировать произвольные отчетные данные, используя это средство.

Совершенствование тематического анализа

Онтологии могут использоваться для более гибкого и полного тематического анализа данных Системы следующим образом. В Систему поступают, в частности, данные о научных статьях сотрудников организации. Часто они включают в себя ключевые слова, аннотацию и полный текст статьи. В этом случае к ним можно применить более сложные алгоритмы тематического анализа. Они позволяют выделить основные термины, встречающиеся в аннотации и в полном тексте, а затем сопоставить их с понятиями онтологии соответствующей предметной области. После этого будет возможен более сложный тематический анализ работ в определенной предметной области, а также междисциплинарный анализ. Эти виды анализа могут сопоставлять отдельной статье не одну тематическую рубрику (совпадающую с тематикой журнала, в котором она опубликована, как это реализовано в настоящее время), а различные тематические груп-

пы, выявляемые динамически из общего графа понятий предметной области.

Для реализации описанного подхода необходимо создание онтологий предметных областей экспертами. Отметим, что разработчиками Системы созданы алгоритмы, облегчающие создание таких онтологий. С помощью этих алгоритмов можно в автоматизированном режиме получить начальную онтологию предметной области, составленную из терминов, встречающихся в анонсах научных конференций по заданной тематике.

Возможным вариантом реализации данного подхода без привлечения экспертов для составления онтологии является разработка методов анализа ключевых слов. В настоящее время в Системе имеется множество взаимосвязанных информационных объектов (статьи, учебные курсы, научные проекты, авторы, структурные подразделения). Некоторые из них содержат наборы ключевых слов. Анализ взаимосвязей между объектами, аналогичный современным методам анализа социальных сетей, может использоваться для установления семантических соотношений между объектами. Основным отличием этого метода от классических алгоритмов анализа ключевых слов является использование в процессе анализа развитой структуры взаимосвязей между объектами, а не только количественные характеристики совместной встречаемости ключевых слов.

Объективный анализ показателей результативности ученых в различных предметных областях

Одним из основных недостатков информационных систем, целью которых является анализ научной деятельности организации, является тот факт, что эти системы при анализе не могут учитывать специфику различных предметных областей. Эта особенность приводит к тому, что статистические данные имеют значительное отклонение (акцент), например, в сторону естественнонаучных специальностей, по которым российские ученые представлены в западных журналах наиболее полно. Кроме этого, во многих областях статьи не являются главным показателем результативности ученого.

Для преодоления отмеченного недостатка можно использовать онтологии. В онтологии на едином языке могут быть формализованы особенности разных предметных областей, что позволяет единообразно учитывать их при анализе.

6.3. Оценка актуальности тематических направлений

В современной науке постоянно появляются новые области и направления исследования, а некоторые, напротив, отмирают. Однако определение как наиболее актуальных, так и зарождающихся направлений научных исследований— непростая задача. Использование информации о публикуемых статьях в журналах и книгах не позволяет добиться должной оперативности. Более того, журналы в большинстве своём консервативны в выборе тематики, а монографии аккумулируют имеющиеся знания. Возможным объектом для анализа являются сообщения о предстоящих научных мероприятиях. Информация, полученная путём анализа сообщений о конференциях, отражает актуальные научные направления. Это обуславливается тем, что временной интервал от окончания рассмотрения заявок до начала конференции составляет всего несколько месяцев. Кроме того, хотя конференции как и журналы посвящены некоторому общему направлению, конкретный список ключевых тем конференций (так называемые, *topics of interest*) меняется от года к году и представляет непосредственный интерес.

Информация о прошедших и планируемых конференциях существенна при решении широкого круга задач. Наличие базы знаний о датах, тематике и месте проведения конференций позволит создать систему с удобными механизмами поиска и навигации по предстоящим событиям, которая будет учитывать месторасположения пользователя и область его интересов. Данные о программном комитете конференции в совокупности с внешними данными об индексе цитируемости его участников могут служить основой для автоматического построения рейтинга конференции. Такая информация, в свою

очередь, может использоваться при оценки научной работы сотрудников специализированными аналитическими системами. Анализ динамики тематических направлений конференций позволит определять тенденции в современной науке, выявлять актуальные тематические направления и определять основные организации и страны, которые работают на этих направлениях.

Для анализа сообщений о конференциях необходимо представить информацию в виде структурированных данных. В простейшем случае можно представлять, что для каждой конференции имеется её карточка, таблица пар ключ-значение, где набор ключей задан заранее и зависит от задачи, а значения суть объекты различного типа, как-то, страны, города, организации, даты, люди. Использование объектного, онтологического подхода к представлению данных позволяет выполнять онтологический вывод и получать данные, не заложенные в тексте отдельного сообщения. Примером полезного для приложений отношения является отношение университет-страна. Оно позволяет дополнить карточку конференции странами членов программного комитета, если они не указаны явно. Далее, зная координаты университетов и организаций, можно находить географические области, в которых наиболее активно развиваются те или иные направления исследований. Кроме того, связь с онтологией уменьшает разреженность данных, так как разные написания одного объекта не создают двух разных сущностей.

В простейшем случае построение описанной структуры данных может осуществляться в два этапа. На первом происходит выделение в тексте участков, которые содержат тот или иной атрибут. А на втором этапе выделенным участкам текста сопоставляется объект из онтологии. Однако во многих случаях решение на первом этапе зависит от информации на следующем. Например, дата проведения конференции не может предшествовать дате подачи работ на конференцию. Однако для сравнения дат необходимо преобразовать из используемого в сообщении текстового формата в объект, представляющий дату. Для разрешения этого конфликта можно учитывать все данные онтологии на этапе поиска объектов, но это существен-

но усложняет модель для рассматриваемой задачи, когда извлечению подлежат различные поля, использующие существенно различные алгоритмы. Как следствие, в данной работе используется итеративное построение карточки объекты, при котором построенные на первом этапе значения-кандидаты, далее отбраковываются и привязываются к онтологии.

Задача извлечения информации из текстов привлекла внимание исследователей достаточно давно, и с увеличением количества доступных в электронном виде текстовых документов рос и интерес к ее решению. Первая серия конференций, посвящённых извлечению данных, MUC (Message Understanding Conference), была организована в 1987 году, и проходила по 1997 год. Каждая конференция представляла собой соревнование по извлечению определённых полей или отношений из фиксированного набора текстов. Первые системы представляли набор просто набор правил и регулярных выражений, однако в последствии их заменили более эффективные вероятностные модели, такие как модель максимальной энтропии или модель условных случайных полей.

Однако классические методы извлечения информации не переносятся напрямую на случай выделения данных из сообщений о конференциях. Это обусловлено тем фактом, что сообщения используют множество графических элементов для обозначения таких полей как название конференции или список тематик, в то время как классические методы ожидают в этих местах лексического контекста, например: «Конференция ... состоится в ...»

Методы выделения структурных элементов долгое время изучались в области оптического распознавания символов (optical character recognition, OCR). В рамках задачи извлечения информации из CFP выделение распознавание структурных элементов, таких как список ключевых тем, также играет важную роль. Однако в отличие оптического распознавания единицами измерения не растровые, а символьный, и многие графические элементы эмулируются с помощью символьной псевдографики. Более того, некоторые типы источники (подробный анализ источников приведён в следующем разделе) могут содержать

HTML-разметку с той же семантической нагрузкой (таблицы, списки). Таким образом, специфика задачи извлечения информации состоит в большем разнообразии методов выделения тех и иных полей, причём преимущественно невербальных.

Большинство информационных ресурсов в сети Интернет, предоставляющие списка предстоящих научных мероприятий, основаны на ручном вводе данных. Другими словами, работы данной категории представляют информационные сайты в сети Интернет, данные которых пополняются с помощью специальных форм либо узких кругом ответственных, либо всеми пользователями ресурсами с последующей модерацией. Работы второй группы, напротив, используют различные алгоритмы для поиска и извлечения данных, представляющий интерес для приложений.

К числу наиболее известных ресурсов можно отнести календари конференций профессиональных сообществ, например, ACM¹ или IEEE² или проект WikiCFP³, который позволяет пользователям добавлять информацию о событии с помощью специальной формы. Пользователю предоставляется возможность полнотекстового поиска по всему тексту объявления, категориям, месту проведения и году. К недостаткам данного проекта отсутствия семантического анализа вводимых данных. Это делает невозможных реализацию таких удобных для пользователя как поиск по неточному месту проведения, выбор конференций по данной и смежной с ней темами. Другим недостатком данного проекта является нерепрезентативный набор конференция. Большинство данных связано с компьютерной тематикой.

Проект кафедры английского языка Пенсильванского университета⁴, напротив, специализируется на конференциях гуманитарного характера. Для создания объявления помимо текста самого объявления требуется минимальное количество данных, а именно контактные данные заявителя и список связанных категорий. Простая процедура внесения объявлений приводит к увеличению их количества,

¹<http://www.acm.org/calendar-of-events>

²http://www.ieee.org/conferences_events/index.html

³<http://wikicfp.com>

⁴<http://call-for-papers.sas.upenn.edu/>

но сводит возможности поиска и анализа к минимуму.

В сети Интернет можно найти и другие ресурсы, собирающие сообщения о предстоящих конференциях¹. Они представляют разную степень детализации имеющихся данных: на некоторых ресурсах известно только название и текст объявления о конференции (так называемое CFP, Call For Papers), другие — предоставляют дополнительную информацию о времени проведения и важных датах. Однако для всех баз данных о планируемых конференциях, заполняемых вручную, свойственно одно очень важное ограничение, а именно — в них не может быть одновременно достигнута высокая полнота покрытия и высокая точность и подробность метаданных. Более того, выделение некоторых полей требует определённых знаний о предметной области, в следствии чего базы данных, содержащих полезные для дальнейшей аналитики метаданные, имеют ограниченный и узкоспециализированный охват.

По описанным причинам такие системы могут быть использованы либо в качестве источника данных для автоматической системы извлечения метаинформации, либо в качестве эталона, с которым стоит сравнивать результаты автоматического извлечения. Однако работы данного класса не решают задач, связанных с репрезентативным анализом развивающихся научных направлений.

В последние несколько лет стало активно развиваться направление, связанное с автоматическим анализом сообщений о конференциях. Задача извлечения данных состоит в том, чтобы в сообщении о конференции, представленном в некотором текстовом формате или в виде совокупности Интернет страниц, найти представляющие интерес участки, и преобразовать их в машинночитаемый формат.

Первые работы по данному направлению были выполнены для конкурса, проходившего в рамках семинара по использованию машинного обучения для извлечения информации PASCAL, которое

¹<http://grid.hust.edu.cn/call/index.jsp>

<https://research.cs.wisc.edu/dbworld/browse.html>

<https://www.h-net.org/announce/group.cgi?type=CFPs>

<http://prorch.com/>

<http://eventseer.net/>

в 2005 году было посвящено извлечению данных из сообщений о конференциях. Для проведения конкурса были специально отобраны сообщения о предстоящих семинарах при конференциях преимущественно по компьютерной тематике, для которых группа ассессоров выделила участки текста, содержащие определённую информацию, а именно: 8 полей, относящихся к семинару (название, аббревиатура, домашняя страница, место проведения, дата проведения, срок подачи работ, срок уведомления о решении и срок подачи финальной версии) и 3 поля, относящиеся к конференции (название, аббревиатура и домашняя страница). Несложно видеть, что извлекаются только поля, которые могут иметь одно значение, в то время как критически важная для последующего анализа информация о тематике конференции обычно представлены несколькими значениями. Такой выбор извлекаемых полей обусловлен тем фактом, что ключевым объектом исследования для данного соревнования являлись методы машинного обучения, а также методы их корректного сравнения. По этой причине входные данные были предоставлены участникам в виде предобработанных текстов (токенизация, выделение частей речи и именованных сущностей), и использования других данных было запрещено, таким образом выделение тематических дескрипторов, в большинстве случаев представленной списка с псевдографикой, было затруднительно. Отметим также, что в задачу участников не входило «понимание» выделяемой информации (например, сопоставления места проведения со списком стран из некоторой онтологии).

Анализ существующих решений по автоматическому анализу сообщений о конференциях показал, что в настоящее время не существует систем, способных обрабатывать сообщения о конференциях в различных предметных областях. Дополнительные сложности могут быть связаны с необходимостью анализа данных на нескольких языках. В рамках системы ИСТИНА наибольшее значение имеют русский и английский языки и разработка надежных методов автоматического извлечения структурированной информации из анонсов научных мероприятий является важным направлением развития Системы.

Заключение

Информационная система ИСТИНА является актуальным на сегодняшний день программным продуктом. Она доказала свою эффективность при внедрении в Московском университете. Большое число запросов из других научно-образовательных организаций демонстрируют востребованность Системы в этих организациях. При этом потенциал, заложенный при проектировании Системы, раскрыт еще далеко не полностью. В настоящее время ИСТИНА преодолела первый этап своего развития, а именно – использование её, в первую очередь, для ввода и накопления данных о результатах деятельности сотрудников организации. Тем не менее, Система уже доказала свою применимость для эффективного анализа и отображения этих данных, а также для формирования отчетов различного уровня.

Система ИСТИНА задумывалась для использования, в том числе, как полигон для применения самых современных средств интеллектуального анализа данных. Наиболее перспективной функцией на этом направлении является полноценное использование онтологий при анализе данных Системы и при использовании других процедур (сервисов). Разработка этой функции требует большого объема труда экспертов, однако при этом предоставляет возможность глубоко анализировать отдельные тематические группы внутри области науки, и оценивать динамику их развития в организации.

В дальнейшем предполагается расширить спектр применения Системы в МГУ имени М.В. Ломоносова (например, использовать ее в качестве программной платформы для создания веб-сайтов отдельных структурных подразделений), а также предоставить к ней

доступ другим научным и образовательным организациям Российской Федерации. Когда в системе ИСТИНА будут зарегистрированы крупнейшие российские научные и образовательные учреждения, у руководителей межведомственных структур появятся следующие возможности:

- построить информационную карту российской науки по отраслям с детализацией по регионам, организациям и вплоть до конкретных ученых;
- формировать аналитические отчеты, например, «самые успешные организации / ученые в заданной научной отрасли», «научные отрасли, в которых успешно работает заданная организация» и другие, аналогичные им;
- выявлять и анализировать тренды развития организаций, регионов и страны в целом по отдельным разделам науки во времени.

Список литературы

- [1] *Mika, Peter*. Application of semantic technology for social network analysis in the sciences / Peter Mika, Tom Elfring, Peter Groenewegen // *Scientometrics*. — 2006. — Vol. 68, no. 1. — Pp. 3–27.
- [2] *Ponomareva, Natalia*. AIR: A Semi-Automatic System for Archiving Institutional Repositories / Natalia Ponomareva, Jose Gomez, Viktor Pekar // *Natural Language Processing and Information Systems* / Ed. by Helmut Horacek, Elisabeth Métais, Rafael Muñoz, Magdalena Wolska. — Springer Berlin / Heidelberg, 2010. — Vol. 5723 of *Lecture Notes in Computer Science*. — Pp. 169–181.
- [3] *Syed, Zareen*. Wikitology: Wikipedia as an ontology / Zareen Syed, Tim Finin, Anupam Joshi // *Proceedings of the Grace Hopper Celebration of Women in Computing Conference*. — ACM Press, 2008.
- [4] Возможности технологий ИСИР в поддержке Единого Научного Информационного Пространства РАН / А.А. Бездушный, А.К. Нестеренко, Т.М. Сысоев и др. // *Электронные библиотеки*. — 2004. — Т. 7, № 6. — С. 15–30.
- [5] *Боровикова, О.И.* Онтологический подход к построению систем информационной поддержки научной и производственной деятельности / О.И. Боровикова, Ю.А. Загорулько, Е.А. Сидорова // *Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–09)*. — Т. 2. — Новосибирск: Институт математики им. С.Л. Соболева СО РАН, 2009. — С. 93–102.

- [6] *Загорулько, Ю.А.* Портал знаний по компьютерной лингвистике: содержательный доступ к лингвистическим информационным ресурсам / Ю.А. Загорулько, О.И. Боровикова, Г.Б. Загорулько // Компьютерная лингвистика и интеллектуальные технологии. Электронные публикации Международной конференции «Диалог-2008». [Электронный ресурс]. Режим доступа: <http://www.dialog-21.ru/digests/dialog2008/materials/html/Zagorulko.htm>, свободный. — 2008.
- [7] *Левенштейн, В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов / В.И. Левенштейн // *ДАН СССР*. — 1965. — Т. 163, № 4. — С. 845–848.
- [8] *Ratcliff, J W.* Pattern matching: The gestalt approach / J W Ratcliff, D Metzener, Others // *Dr. Dobb's Journal*. — 1988. — Vol. 7. — P. 46.
- [9] *Васенин, В.А.* Архитектура, методы и средства базовой составляющей системы управления научной информацией «ИСТИНА – Наука МГУ» / В.А. Васенин, Д.Д. Голомазов, Г.М. Ганкин // *Программная инженерия*. — 2014. — № 9. — С. 3–12.
- [10] *Percival, Harry J.W.* Test-Driven Web Development with Python / Harry J.W. Percival. — O'Reilly Media, 2014. — P. 478.
- [11] The swrc ontology - semantic web for research communities / York Sure, Stephan Bloehdorn, Peter Haase et al. // Proceedings of the 12th Portuguese Conference on Artificial Intelligence - Progress in Artificial Intelligence (EPIA 2005), volume 3803 of LNCS. — Covilha: Springer, 2005. — Pp. 218–231.
- [12] *Васенин, В.А.* К созданию системы управления научной информацией на основе семантических технологий / В.А. Васенин, С.А. Афонин, Д.Д. Голомазов // *Материалы Всероссийской конференции с международным участием "Знания - Онтологии - Теории"*(ЗОНТ-2011), 3-5 октября 2011 г., том 1 Новосибирск.

- Институт математики им. С.Л. Соболева СО РАН, Новосибирск, 2011. — С. 78–87.
- [13] Голомазов, Д.Д. Выделение терминов из коллекции текстов с заданным тематическим делением / Д.Д. Голомазов // *Информационные технологии*. — 2010. — № 2. — С. 8–13.
- [14] Hearst, M.A. Automatic acquisition of hyponyms from large text corpora / M.A. Hearst // *Proceedings of the 14th conference on Computational linguistics-Volume 2*. — Association for Computational Linguistics, 1992. — Pp. 539–545.
- [15] Pudovkin, Alexander I. Rank-normalized impact factor: A way to compare journal performance across subject categories / Alexander I. Pudovkin, Eugene Garfield // *Proceedings of the American Society for Information Science and Technology*. — 2004. — Vol. 41, no. 1. — Pp. 507–515. <http://dx.doi.org/10.1002/meet.1450410159>.
- [16] Hirsch, J. E. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship / J. E. Hirsch // *Scientometrics*. — 2010. — . — Vol. 85, no. 3. — Pp. 741–754. <http://dx.doi.org/10.1007/s11192-010-0193-9>.
- [17] Интеллектуальная Система Тематического Исследования Научно-технической информации (ИСТИНА) / В.А. Васенин, С.А. Афонин, Д.Д. Голомазов, А.С. Козицын // *Информационное общество*. — 2013. — № 1-2. — С. 21–36.
- [18] Васенин, В.А. Автоматизированная система тематического анализа информации / В.А. Васенин, А.С. Козицын, С.А. Афонин // *Информационные технологии - приложение*. — 2009. — № 4. — С. 1–32.
- [19] Афонин, С.А. Использование онтологий в поисковых системах / С.А. Афонин, А.С. Козицын // *Материалы Всероссийской конференции с международным участием "Знания-Онтологии-Теории г. Новосибирск, т. 2*. — 2009. — Pp. 47–52.

- [20] Scatter/gather: A cluster-based approach to browsing large document collections / Douglass R Cutting, David R Karger, Jan O Pedersen, John W Tukey // Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval / ACM. — 1992. — Pp. 318–329.
- [21] *Титов, А.С.* Кластеризация на основе гравитационного метода / А.С. Титов // *Межвузовский сборник статей*. — 2003. — Vol. 2, no. 7. — Pp. 51–60.
- [22] *Браславский, П.И.* Тезаурус для расширения запросов к машинам поиска интернета: структура и функции / П.И. Браславский // *Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. Диалог'2003, Протвино*. — М.: Наука, 2003. — Pp. 95–100.
- [23] *Афонин, С.А.* Разрешение неоднозначности авторства публикаций при автоматической обработке библиографических данных / С.А. Афонин, А.Э. Гаспарянц // *Программная инженерия*. — 2014. — № 1. — С. 25–29.
- [24] *Sergey, Afonin.* SQLReports: Yet Another Relational Database Reporting System / Afonin Sergey, Kozitsyn Alexander, Astapov Ivan // Proceedings of the 9th International Conference on Software Engineering and Applications. — 2014. — Pp. 529–534.
- [25] *Sergey, Afonin.* On automated hyperonym hierarchy construction using an internet search engine / Afonin Sergey // Proc. of the 24th Pacific Asia Conf. on Language, Information and Computation (PACLIC). — 2010. — Pp. 341–348.
- [26] *Афонин, С.А.* Построение иерархии понятий на основе лексических шаблонов / С.А. Афонин, А.В. Бахтин // *Информационные технологии*. — 2012. — № 3. — С. 2–7.