МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ОРЕНБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

На правах рукописи

БАРАНОВ ДМИТРИЙ АЛЕКСАНДРОВИЧ

РЕШЕНИЕ ЗАДАЧ ПОСТРОЕНИЯ ОПТИМАЛЬНЫХ ИССЛЕДОВАТЕЛЬСКИХ ТРАЕКТОРИЙ АГЕНТОВ НАУЧНОГО ПРОИЗВОДСТВА

Диссертация на соискание ученой степени
кандидата технических наук

Специальность 05.13.10 — Управление в социальных и экономических системах

Научный руководитель:

к.т.н., доцент, Влацкая И.В.

Оренбург 2016

Содержание

Введение

Актуальность работы. Современная научная среда представляет собой сложную систему взаимодействующих объектов, так или иначе связанных с производством научного знания – агентов научного производства (АНП). К ним относятся отдельные учёные, научные коллективы, научные организации, журналы и т.д. Каждый агент научного производства в своей деятельности стремится не только к производству научного знания, но и к ряду дополнительных целей: признанию со стороны научного сообщества, получению государственного финансирования и т.д. Для достижения частных целей АНП необходимо повышать показатели результативности своей научной деятельности. На данный момент в качестве основного метода оценки результатов научной деятельности используются наукометрические показатели, такие как импакт-фактор, индекс цитирования и экспертные оценки.

Активное развитие науки в последнее время привело к повышенной конкуренции среди агентов научного производства. В этих условиях АНП всё чаще прибегают к различным способам повышения эффективности своей деятельности, т.е. увеличению показателей её результативности. В частности, применяются различные методы управления научной деятельностью. Проблемы управления научной деятельностью, рассматривают в своих работах Ю. Л. Качанов, Н. А. Шматко, Д. А. Новиков, А. Л. Суханов, А. А. Першин, C. S. Wagner, D. J. Roessner, J. T. Klein, J. Y. Tsao, M. E. Coltrin, J. G. Turnley, W. B. Gauster.

Большой интерес представляют работы, использующие математическое моделирование предметной области научных исследований (ПрО). Большой вклад в этом направлении внесли Н. В. Максимов, О. В. Окропишина, А. Е. Окропишин, И. Г. Передеряев, С. Н. Ларин, Л. И. Герасимова, Л. В. Найханова, H. Small, K. W. Boyack, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars.

Однако, в существующих работах недостаточно полно освещены методы управления научной деятельностью, учитывающие изменение ПрО в процессе этой деятельности. Процесс деятельности АНП можно описать как исследовательскую траекторию (ИТ) – последовательное изменение изучаемой ПрО по

мере развития исследования, выражаемое в публикации новых работ (объектов публикационной активности). Выбирая оптимальную исследовательскую траекторию, агент научного производства может существенно повысить свою эффективность. Например, научный журнал может достичь более высокого импакт-фактора за счёт выбора наиболее актуального материала для каждого номера, в целью научного коллектива может быть повышение среднего индекса цитируемости за счёт выбора наиболее перспективных направлений научной деятельности.

Следовательно, актуальна разработка математических моделей, алгоритмического и программного обеспечения для решения задач построения оптимальных исследовательских траекторий агентов научного производства.

Объектом исследования являются исследовательские траектории агентов научного производства.

Предметом исследования является моделирование и оптимизация исследовательских траекторий агентов научного производства.

Цель и задачи исследования. Целью диссертационной работы является повышение эффективности деятельности агентов научного производства за счёт разработки математических моделей, алгоритмического и программного обеспечения для построения оптимальных исследовательских траекторий.

Для достижения поставленной цели сформулированы следующие задачи:

1. Провести анализ существующих подходов, методов и математических моделей, используемых при решении задач управления научной деятельностью.

2. Разработать математическую модель предметной области агента научного производства.

3. Разработать математическую модель исследовательской траектории агента научного производства.

4. Разработать методику построения оптимальных исследовательских траекторий агентов научного производства.

5. Разработать информационную систему для моделирования, оценки и оптимизации исследовательских траекторий агентов научного производства.

6. Провести апробацию полученных результатов на задачах построения оптимальных исследовательских траекторий для различных агентов научного производства.

Методология и методы исследования. В основе диссертационной работы лежат методы теории множеств, линейной алгебры, теории графов, теории вероятностей, математической статистики, марковских процессов, численных методов, нечёткой логики, теории оптимального управления, теории принятия решений, теории оптимизации, динамического программирования, параллельных вычислений.

Область исследования. Содержание диссертации соответствует пунктам 4, 5 и 12 паспорта специальности 05.13.10 «Управление в социальных и экономических системах».

Научная новизна работы заключается в**:**

1. Разработанной графосемантической модели предметной области, исследуемой агентом научного производства. Полученная модель отличается представлением структуры предметной области в виде графа, вершинами которого являются терминополя, формируемые экспертами из набора ключевых слов, описывающих объекты публикационной активности агента научного производства.

2. Разработанной математической модели исследовательской траектории агента научного производства, в основе которой лежит графосемантическая модель предметной области и вероятностная графосемантическая модель.

3. Предложенной методике повышения эффективности деятельности агентов научного производства за счёт построения оптимальной исследовательской траектории агента научного производства, основанной на модификации метода динамического программирования Беллмана с применением генетического алгоритма.

Теоретическая значимость результатов исследования заключается в развитии и совершенствовании метода графосемантического моделирования, в частности в разработке вероятностной графосемантической модели, а также в разработанной методике моделирования и оптимизации предметных областей агентов научного производства.

Практическую ценность представляют разработанные алгоритмы выделения ведущих научных направлений на основе кластерного анализа, вычисления семантической карты с применением динамического программирования, алгоритм решения дискретной задачи оптимального управления на основе генетического алгоритма, а также разработанная информационная система для моделирования, оценки и оптимизации исследовательских траекторий агентов научного производства.

Основные положения, выносимые на защиту:

1. Математическая модель предметной области на основе графосемантического моделирования.

2. Математическая модель исследовательской траектории, оптимизируя которую можно добиться повышения эффективности деятельности агента научного производства.

3. Методика построения оптимальной исследовательской траектории агента научного производства на основе полученных математических моделей.

4. Решение задач построения оптимальной исследовательской траектории для научного журнала и научного коллектива на основе разработанного программного обеспечения.

Достоверность результатов исследования подтверждается внедрением разработанных моделей, методов, алгоритмов и информационной системы «Семограф» в ЗАО «Прогноз», в также в процесс научной деятельности лаборатории прикладных и экспериментальных лингвистических исследований Пермской социопсихолингвистической школы.

Апробация результатов. Основные теоретические и практические результаты работы докладывались и обсуждались на следующих научных и

научно-практических конференциях: XII Всероссийская конференция «Высокопроизводительные параллельные вычисления на кластерных системах», Нижний Новгород, 2012; VIII Международная научно-практическая конференция «Техника и технология: новые перспективы развития»; I-я Международная научная конференция «Формирование основных направлений развития современной статистики и эконометрики»; XIV-я Международная научно-техническая конференция «Проблемы техники и технологий телекоммуникаций»; Международная научная конференция «Наука и образование: фундаментальные основы, технологии, инновации».

Работа выполнялась в рамках НИР по государственному заданию ОГУ, проект № 8.2714.2011 «Когнитивно-информационные модели научной картины мира»; проекта № 12-04-12034в при поддержке РГНФ «Информационная система графосемантического моделирования (Семограф)»; НИР по контракту № 14.B37.21.0176 при поддержке Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009-2013 гг; НИР «Разработка методического обеспечения поиска, оценки и прогнозирования изменений потребностей в информационных ресурсах в отраслях экономики» по ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технического комплекса России на 2014-2020 годы».

Публикации. Основные положения и результаты диссертационной работы опубликованы в 14 работах, в том числе 5 работ в изданиях, входящих в перечень рецензируемых научных журналов ВАК, и 3 в журналах, индексируемых Scopus; 2 публикации выполнены без соавторства. Разработанное программное обеспечение защищено свидетельством о государственной регистрации программы для ЭВМ.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, библиографического списка из 105 источников, приложений. Общий объем работы составляет 145 страниц, включая 37 рисунков и 22 таблицы.

1 Актуальные задачи управление научной деятельностью и оценки её результатов

1.1 Анализ существующих подходов к управлению научной деятельностью

Наука – сфера деятельности, ориентированная на выработку знаний о мире, их систематизацию, построение образа мира и определение способов взаимодействия с миром [67]. В современном обществе наука приобретает всё большее значение. Сложно переоценить её влияние на все сферы жизни человека. Благосостояние общества напрямую зависит от уровня его технологического развития. Благодаря этому, наука является стратегическим направлением развития большинства современных государств.

Научная (научно-исследовательская) деятельность – деятельность, направленная на получение и применение новых знаний [67]. В результате повсеместной информатизации и компьютеризации быстро увеличиваются объёмы информации, обрабатываемые учёными, а также производимого ими научного знания. Вместе с тем растут и объёмы средств, выделяемых крупными государствами на развитие и стимулирование науки. При этом, как правило, распределение значительной части этих средств происходит не равномерно, а в виде грантов, премий и т.п.

Ещё одним важным аспектом развития науки явились понятия интеллектуальной собственности и авторского права. В частности, сложившаяся на данный момент система патентования изобретений стимулирует крупные коммерческие компании заниматься научной деятельностью и патентовать полученные результаты с целью извлечения дополнительной прибыли, например посредством лицензионных отчислений. При этом для некоторых компаний значительную часть активов составляет именно интеллектуальная собственность [73; 91; 100].

Следствием данных процессов является возрастающая конкуренция в научной среде. Эта конкуренция обуславливает стремление учёных к повышению собственной эффективности, заключающейся в достижении более вы-

соких результатов в меньшие сроки. Стремление к повышению эффективности ведёт к необходимости решения ряда задач, в т.ч. задач управления научной деятельностью. Следует отметить, что актуальность данного типа задач привела к появлению понятий «научного менеджмента» [78] и «менеджмента знаний» [67]. Одну из первых попыток описать научную работу как процесс управления предпринял P. Freeman в 1905 г. [19].

Актуальность данного направления на данный момент подчёркивается повышенным внимание на глобальном уровне. Так, можно выделить Федеральные целевые программы мероприятия 1.1 «Проведение исследований, направленных на формирование системы научно-технологических приоритетов и прогнозирование развития научно-технологической сферы»: 2015-02-573-0010 на тему «Разработка предложений по совершенствованию системы статистического учета в области научных исследований и разработок» и 2015-14-573-0011 «Разработка и практическая апробация системы комплексного мониторинга направлений развития науки и технологий гражданского характера».

Научную среду можно рассматривать как самоорганизующуюся систему [74; 98; 104], тогда управление научной деятельностью можно описать моделью, предложенной в [50]. Схема данной модели приведена на рисунке 1.

На сегодняшний день существует множество работ, посвящённых проблемам управления научной деятельностью на разных уровнях и наукой в целом. Можно выделить 3 основных типа решаемых задач по уровню объекта управления:

1. Глобальный уровень – уровень страны, региона и т.д.; основные задачи управления сводятся к распределению ресурсов между научными направлениями.

2. Локальный уровень – уровень научной организации, ВУЗа, предприятия в наукоёмкой отрасли; основными задачами являются достижение конкретных результатов (например, выполнение контракта или решение определённой задачи) а также распределение ресурсов между научными подразделениями, стимулирование исполнителей, оперативное управление.

3. Уровень агента научного производства – к данному уровню чаще все-

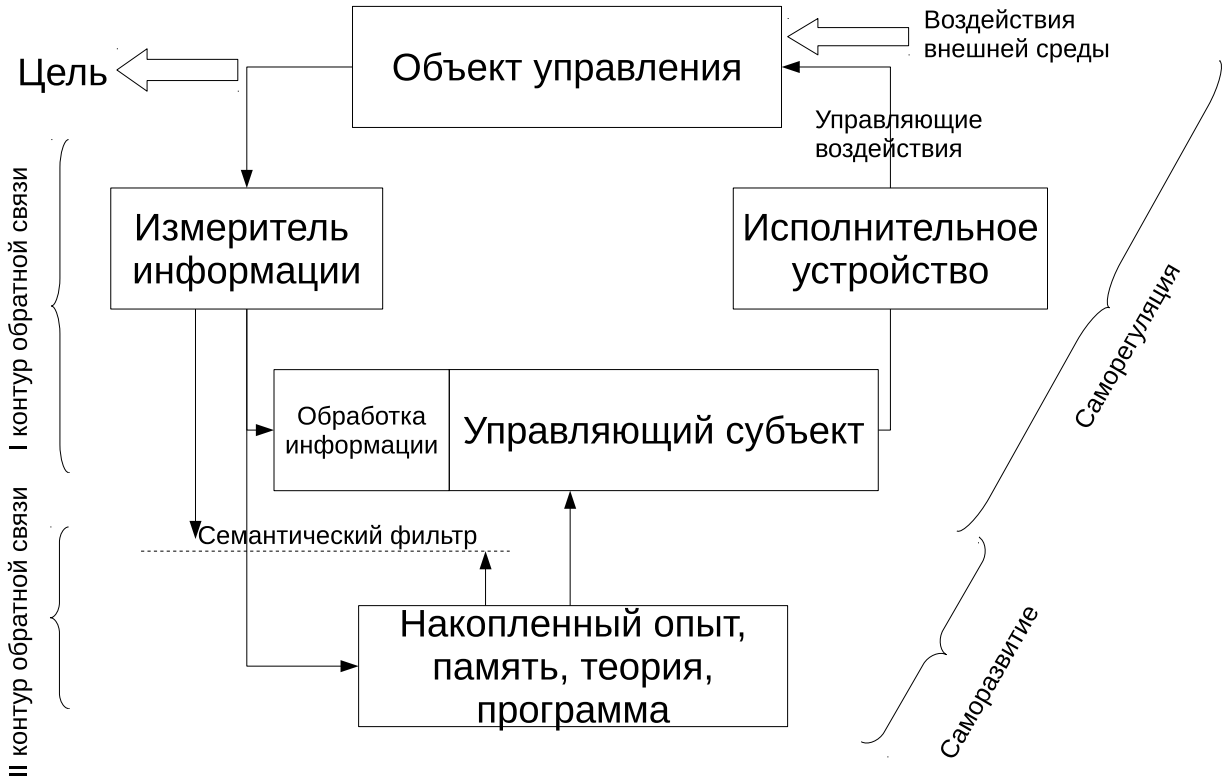

Рисунок 1 — Обобщённая модель механизма управления для самоорганизующихся систем

го относятся научные коллективы и отдельные учёные, однако с проблемами, описываемыми на данном уровне, могут сталкиваться и АНП более высоких уровней; на данном уровне чаще решаются такие задачи, как управление структурой коллектива, выбор направления научной деятельности и т.д.

Большой класс существующих работ 1 уровня посвящён изучению проблем взаимосвязи государства и науки, в частности государственному регулированию науки. Значительный вклад в изучение вопросов взаимодействия институтов государства и науки в России внесли А.Н. Авдулов, А.М. Кулькин, В.В. Глухов, С.Б. Коробко, Т.В. Маринина, А.Г. Аллахвердян, Ю.С. Афанасьев, Б.С. Гершунский, Н.А. Гордеев, Э.Д. Днепров, Н.И. Загузов, Г.И. Ильин, В.В. Краевский, Т.Е. Кузнецова, Л.П. Кураков, С.В. Куров, В.В. Лапаева, В.С. Леднёв, Н.Д. Никандров и др.

Так, в работе [51] выделены следующие шесть аспектов взаимоотношений государства и науки:

1. государство как законодатель;

2. государство как источник средств для научных исследований;
3. государство как потребитель научной и наукоёмкой продукции;
4. государство как крупный субъект научно-технической деятельности;
5. государство как координатор действий всех секторов экономики по развитию национального научно-технического потенциала;
6. государство как политическая сила, во многом определяющая позицию общества по вопросам развития науки и техники.

В работе [83] представлена схема структуры государственного управления научной деятельностью (рисунок 2). Данная схема наглядно демонстрирует взаимодействие социальных институтов науки и государства а также их составляющих.

Значительную роль в вопросах управления научной деятельностью и наукой играет науковедение – отрасль, занимающаяся исследованием науки, её структуры, динамики и взаимодействия с другими социальными институтами. Большой вклад в изучение современного состояния российской науки и науковедения внёс А.Г. Аллахвердян. В своих работах [52; 53] он анализирует новые тенденции в российской науке, рассматривает развитие отечественного науковедения, исследует социально-технологические проблемы и экономические аспекты глобализации науки, а также освещает основные тенденции и направления международного научного сотрудничества ученых.

Не меньшее внимание учёные уделяют проблемам управления научной деятельностью в научных организациях, в частности в высших учебных заведениях и научно-исследовательских институтах. В данном направлении можно выделить работы таких учёных как Н.И. Леонов, А.Н. Авдулов, Ю.Б. Татаринов, Д.А. Новиков, А.Л. Суханов, В.М. Баутин, В.И. Воропаев, Л.Д. Гитман, Л. Гохберг, М.Д. Джонк, И.И. Мазур, В.Д. Шапиро, В. А. Сластёнин, Л.С. Подымова, В.В. Кузнецов, Э. Мэнсфилд, Р. Солоу, Ф. Махлуп, Дж. Эрроу.

В работе [67] приводится определение понятия «научная организация» а также их классификация (таблица 1): научной организацией признается юридическое лицо (независимо от организационно-правовой формы и формы собственности), а также общественное объединение научных работников,

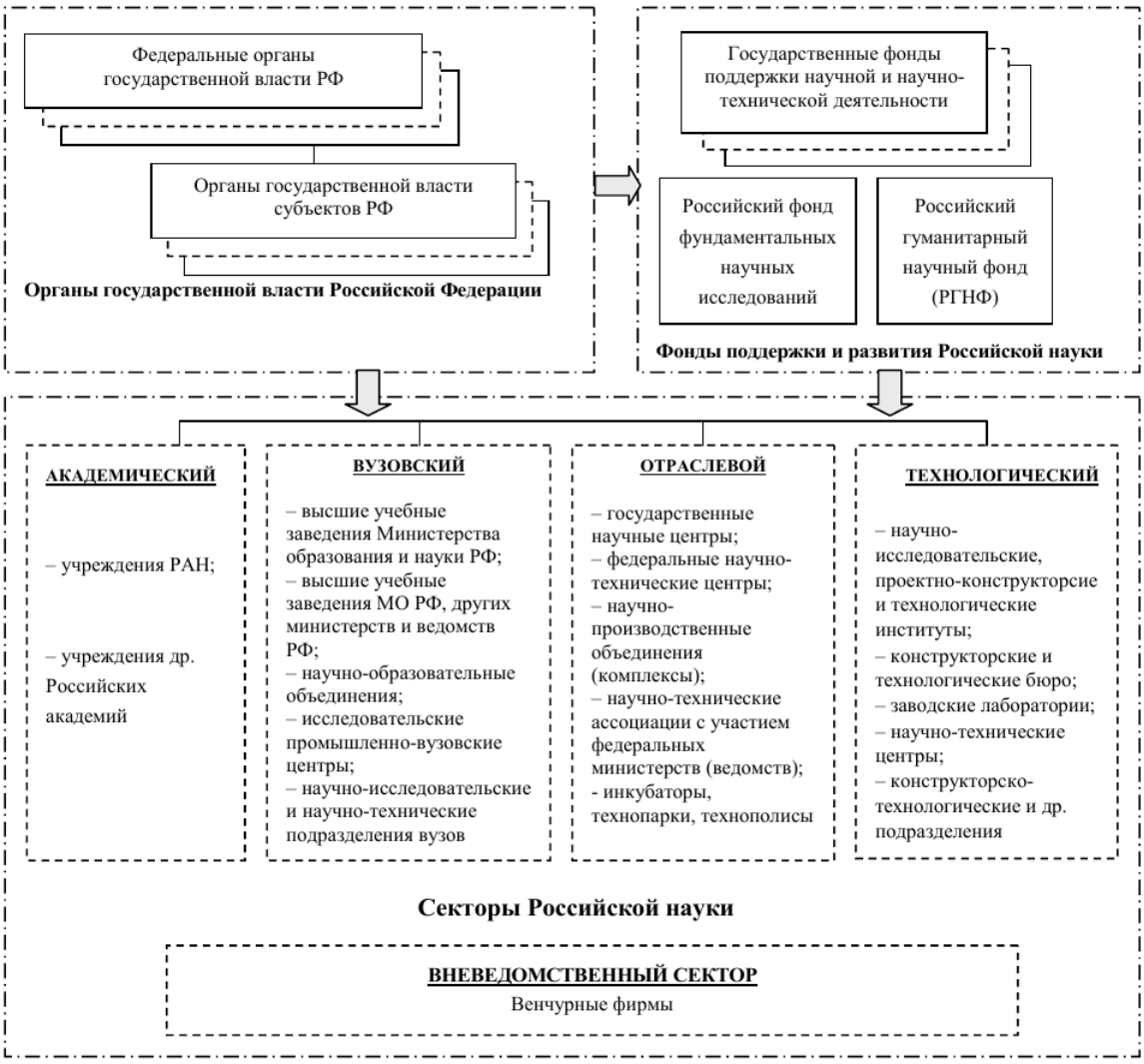

Рисунок 2 — Структура государственного управления научной деятельностью

осуществляющее в качестве основной научную или научно-техническую деятельность, подготовку научных работников и действующее в соответствии с учредительными документами научной организации.

В работе [78] описаны задачи, решаемые ВУЗом в рамках реализации стратегии исследовательского университета, подчёркивается важность научно-исследовательской деятельности ВУЗа и, в частности, проектной. Так же отмечается значимость в современной научной среде информационных платформ коммуникации и комплексного взаимодействия между студентами, сотрудниками и внешним миром.

Отметим, что проектный подход распространён в работах, посвящённых проблемам управления научной деятельностью в научных организациях. Данное направление опирается на теорию управления проектами (В.Н. Бурков, Д.А. Новиков, В.И. Воропаев, Д.И. Голенко-Гинзбург, И.И. Мазур, В.Д. Шапи-

Таблица 1 — Классификация российских научных организаций

Сектор	**Организация**
Академическая наука	Учреждения РАН и других российских академий
Вузовская наука	Научно-образовательные объединения
	Исследовательские промышленно-вузовские центры
	Научно-исследовательские и научно-технические подразделения вузов
Отраслевая наука	Государственные научные центры
	Федеральные научно-технические центры
	Научно-производственные объединения (комплексы)
	Научно-технические ассоциации с участием федеральных министерств (ведомств)
	Инкубаторы
	Технопарки
	Технополисы
Заводская наука	Научно-исследовательские, проектно-конструкторские и технологические институты
	Конструкторские и технологические бюро
	Заводские лаборатории
	Научно-технические центры
	Конструкторско-технологические и другие подразделения
Вневедомственная наука	Венчурные фирмы

ро) и методы математического моделирования, применяемые в решении задач планирования (К.А. Багриновский, В.Л. Макаров, Г.С. Поспелов, В.А. Ириков, Г.Г. Балаян). Согласно [62], проект – это ограниченное по времени целенаправленное изменение отдельной системы с установленными требованиями к качеству результатов, возможными рамками расхода средств и ресурсов и специфической организацией.

В работе [83] подробно исследованы проблемы управления научными проектами в ВУЗах и применяемые математические модели и методы управления. При этом авторы акцентируют своё внимание на проблеме управления исследованиями и разработками (ИР). Особое внимание уделяется информационно-логическим моделям научных исследований (ИЛМ) [54]. Для

реализации ВУЗом научных проектов, авторы предлагают сформировать систему управления, схема которой приведена на рисунке 3.

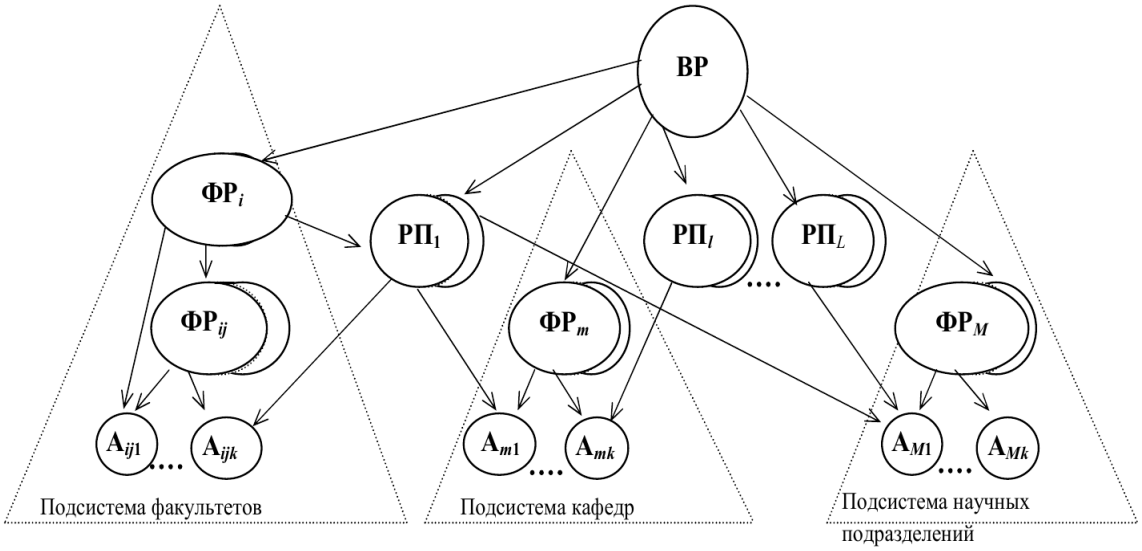

Рисунок 3 — Схема системы управления научными проектами ВУЗа

На основе анализа схемы 3, авторы [83] выделяют следующие основные задачи управления научными проектами в ВУЗе:

1. оценка результатов научных проектов;
2. планирование портфеля научных проектов;
3. распределение ресурсов в научных проектах;
4. стимулирование исполнителей научных проектов;
5. оперативное управление научными проектами.

Кроме проектного подхода к управлению научной деятельностью, широкое распространение получил процессный подход. Концепция процессного подхода подразумевает, что процесс управления может быть выражен в виде непрерывных взаимосвязанных действий – функций управления [80]. К функциям управления относятся: планирование, организация, программно-целевой подход, мотивация, обратная связь и координация. Функции управления объединяются процессами коммуникации и выработки мер воздействия. Схема процессного подхода приведена на рисунке 4 [80].

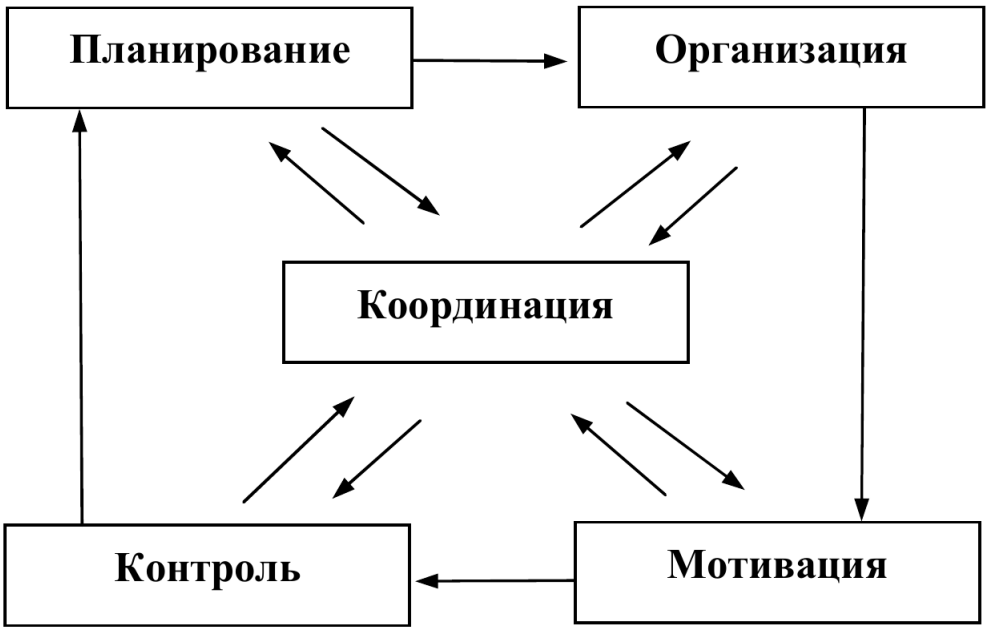

Рисунок 4 — Взаимосвязь основных функций управления научной деятельностью

Ещё одним важным объектом исследований в области управления научной деятельностью является коллектив учёных. Высокая значимость вклада научных коллективов в науку подчёркивается в [50]. При этом, анализируемые работы изучают различные аспекты деятельности научных коллективов и, как следствие, авторы используют разные обозначения научного коллектива, например: научный коллектив, коллективный агент научного производства [72], виртуальный научный коллектив [61] и т.д.

Управление научной деятельностью коллективов учёных рассматривается с множества различных точек зрения. Например, в работе [64] описываются социально-психологические аспекты управление научной деятельностью научных коллективов. В частности, выделены главные факторы, влияющие на продуктивность исследовательской деятельности научного коллектива. Кроме того, в работе [64] вводится понятие малой научной группы – это небольшое по численному составу объединение людей, образующееся на основе совместного решения научных проблем, имеющее структуру межличностных отношений, т.е. непосредственных контактов членов группы между собой и с руководством.

В работе [61] рассматриваются процессы формирования виртуальных научных коллективов, мотивация, введены шесть принципов, обеспечиваю-

щих устойчивость и стабильность. Так же рассмотрены правовые аспекты и организационные вопросы управления виртуальными научными коллективами. Методика оценки эффективности деятельности виртуальных научных коллективов приведена в работе [60].

1.2 Анализ современных подходов к оценке результатов научной деятельности

Ещё одной важной задачей управления научной деятельностью является оценка её результатов. На данный момент разработано множество подходов к оценке результатов научной деятельности, их можно разделить на две большие группы: качественные и количественные. Количественные подходы сводят обобщённую оценку к единому числовому представлению, что позволяет сопоставлять результаты естественным образом, в то время как качественные результаты некоторым образом описывают структуру результатов научной деятельности.

Большинство существующих количественных методов оценки результатов научной деятельности агентов научного производства основывается на формальных наукометрических показателях, относящихся к публикационной активности (ОПА) агентов. Среди наиболее распространённых наукометрических показателей можно выделить разнообразные индексы цитирования (ИЦ) и импакт-фактор (ИФ). Вычисление этих показателей, как правило, осуществляется на основе библиографических или реферативных баз данных, наиболее крупными и авторитетными среди которых являются Thomson Reuters Web of Science (основанная на ISI Web of Knowledge) [47] и Elsevier Scopus [43]. Сбором информации о публикациях, её обобщением и ведением баз данных занимаются крупные международные институты: Institute for Scientific Information (ISI), International Mathematical Union (IMU), International Council for Industrial and Applied Mathematics (ICAM) [1; 2].

Отдельно следует отметить наукометрическую базу данных ключевых научных показателей (Essential Science Indicators) компании Thomson Reuters, на основе которой ежегодно вычисляется значения импакт-фактора для входящих в неё журналов. Импакт-фактор, вычисляемый Thomson Reuters, оче-

видно, является наиболее авторитетной характеристикой научного журнала на сегодняшний день.

На территории Российской Федерации наиболее распространённым средством оценки результатов научной деятельности можно считать показатель национального российского индекса научного цитирования (РИНЦ), вычисляемого проектом «Научная электронная библиотека» [81]. Существенным преимуществом данной системы является широкий охват русскоязычных изданий, в то время как в вышеназванных системах такие издания практически не представлены.

Заметим, что все вышеназванные проекты не предоставляют открытого доступа к своим базам данных, что не позволяет учёным производить над ними исследования и вычислять произвольные оценки. Однако, существуют и открытые библиографические и реферативные базы данных, например PubMed [39], Elsevier Scirus [18], CiteSeerX [9], arXiv [3] и проект от компании Google – сервис Google Scholar [22], однако охватываемые ими тематики зачастую ограничены и практически не содержат материалов русскоязычных изданий. В таблице 2 представлены сравнительные характеристики наиболее известных библиографических баз данных. Заметим, что русскоязычные издания охвачены в полной мере лишь в двух последних проектах – РИНЦ и «МАРС» [35].

Основу расчетов разнообразных индексов (цитируемости, самоцитируемости, Хирша, импакт-фактора и мн. др.) составляют сугубо формальные количественные показатели числа публикаций и цитирований за определенные периоды времени. Многочисленные индексаторы / агрегаторы научного контента поставляют данные для исследований в области наукометрии, которые особенно популярны за рубежом. Исследователи оценивают возможность применения количества цитирований, вводят индексы для оценки эффективности научной деятельности, тестируют существующие индексаторы научного контента (Web of Science, Scopus, Google Scholar и др.), исследуют особенности оценки эффективности в гуманитарных, социальных, естественно-научных и технических направлениях [15; 16; 23; 31], изучают возрастные и гендерные аспекты эффективности и производительности в науке [11; 12], осуществляют анализ междисциплинарных взаимодействий [36; 49] и возможности исполь-

Таблица 2 — Сравнительные характеристики библиографических баз данных

Название	Владелец	Дисциплины	Доступ
Web of Science	Thomson Reuters	Произвольные	Подписка
Scopus	Elsevier	Произвольные	Подписка
arXiv	Cornell University	Физика, математика, компьютерные науки, статистика	Свободный
IEEE Xplore	IEEE	Компьютерные науки, инженерное дело, электроника	Подписка
Microsoft Academic Search	Microsoft	Компьютерные науки	Свободный
PubMed	U.S. National Institutes of Health	Биологические, медицинские	Свободный
Google Scholar	Google	Произвольные	Свободный
РИНЦ	Научная электронная библиотека	Произвольные	Закрытый
«МАРС»	АРБИКОН	Произвольные	Подписка

зования индексов для сопоставления национальных наук и научных специализаций [68; 76; 88].

Эти же формальные показатели используются как основа планирования научной деятельности, в том числе, в аналитических инструментах InCites (приложение, работающее на базе наукометрических показателей Web of Science) [25] и SciVal Spotlight (приложение, работающее на базе наукометрических показателей Scopus) [42], созданных для оценки текущей деятельности агентов научного производства и применяемых как инструмент принятия решения в области финансирования проектов и научных коллективов [90]. Заметим, что в данном случае под планированием научной деятельности, в первую очередь, подразумевается административное планирование поддержки научных коллективов. Однако, планирование научной деятельности со стороны собственно научного коллектива данные инструменты не предусматривают.

Среди наиболее распространённых наукометрических показателей в первую очередь следует отметить импакт-фактор. Импакт-фактор предназна-

чен для оценки и ранжирования журналов по среднему числу цитирований опубликованных в них статей. Однако, импакт-фактор зачастую используется и для оценки учёных, например в виде суммы импакт-факторов всех журналов, в которых были опубликованы работы учёного. Понятие импакт-фактора ввёл Ю. Гарфилд для оценки научных журналов как среднего числа цитирований статей за определённый промежуток времени [21].

Наряду с импакт-фактором, важнейшим наукометрическим показателем на данный момент является индекс цитирования Хирша (или h-индекс), названный в честь автора, Дж. Е. Хирша. Индекс Хирша, как и его производные, используется для оценки учёных. H-индекс данного учёного является наибольшим таким числом n, что у ученого есть n статей, на каждую из которых существует по крайней мере n ссылок [23]. Целью индекса Хирша является создание единой величины для оценки количества цитирований и их распределения [1]. Индекс Хирша был одним из первых в своём роде, что, вероятно, в большой мере обусловило его широкое распространение. Однако, он обладает и рядом известных недостатков. Например, h-индекс не может превышать общего количества статей учёного, поэтому, если учёный опубликовал мало работ, его h-индекс останется минимальным, несмотря на значимость самих работ [14]. В качестве примера можно вспомнить Э. Галуа, опубликовавшего всего четыре работы.

Помимо h-индекса, Дж. Е. Хирш предложил так же m-индекс. M-индекс учёного определяется как отношение его h-индекса к числу лет, прошедших после первой публикации этого учёного. Целью этого индекса является сокращение разницы величины показателя у молодых и опытных учёных [1; 23].

По мере выявления недостатков h-индекса учёные стали предлагать свои варианты индекса, лишённые тех или иных недостатков оригинала. Так, уже в 2006 году были предложены новые варианты индекса цитирования: g-индекс, $H^{(2)}$-индекс, A-индекс и т.д.

G-индекс учёного определяется как наибольшее n, для которого n наиболее цитируемых работ автора (ключевых публикаций) в общей сложности цитируются по крайней мере n^2 раз. Данный индекс, в отличии от h-индекса, учитывает высокую цитируемость отдельных работ учёного [17]. Так же g-индекс позволяет отслеживать изменения в составе ключевых публикаций це-

ной повышения требований к детализации всего списка публикаций учёного (проблема детализации).

$H^{(2)}$-индекс есть k где k – число первых публикаций, процитированных как минимум k^2 раз [29]. К преимуществам данного подхода относят пониженные требования к детализации списка публикаций учёного, а к недостаткам малую чувствительность к его изменениям [46].

A-индекс основан на h-индексе, однако его основным преимуществом является учёт непосредственного числа цитирований ключевых публикаций, составляющих ядро h-индекса. В отличии от g-индекса, A-индекс просто вычисляет среднее арифметическое количеств цитирований каждой из ключевых публикаций [27]:

$$A = \frac{1}{h}\sum_{j=1}^{h} cit_j, \tag{1}$$

где h – h-индекс, $cit_j, j = \overline{1,h}$ – число цитирований j-ой работы.

Отметим, что ядро индекса Хирша однозначно определяет A-индекс. При этом A-индекс использует тот же список публикаций, что и h-индекс, следовательно проблема детализации не возрастает по сравнению с h-индексом, в отличии от g-индекса. Так же заметим, что всегда выполняется неравенство $h \leqslant A$. Однако, в некоторых случаях A-индекс может давать некорректную оценку. Например, индексы h_a, h_b, A_a, A_b вычисленные для учёных a и b могут получить значения, удовлетворяющие следующим неравенствам: $h_a < h_b$ и $A_a > A_b$. Проблема заключается в знаменателе (1) в виде h-индекса: величина h-индекса негативно влияет на значение A-индекса. Для решения данной проблемы в [46] предложен R-индекс:

$$R = \sqrt{\sum_{j=1}^{h} cit_j},$$

или

$$R = \sqrt{Ah}.$$

Очевидно, для R-индекса, подобно A-индексу, всегда выполняется неравенство $h \leqslant R$. Кроме того, в [46] показано, что h, g, A и R-индексы на практике часто коррелируют между собой, при этом коэффициент корреляции между

R и g-индексами всегда выше, чем между R и h или g и h-индексами. Заметим, что корреляция между индексами цитирования отмечалась и в других работах. Например, в работе [24] показано, что h, g и A индексы коррелируют между собой в задаче оценки 99 университетов Тайвани.

Ещё одним индексом, предложенным в [46] является AR-индекс. В отличии от вышеперечисленных индексов, AR-индекс учитывает время (в годах), прошедшее с моментами опубликования работы. Это означает, что AR-индекс со временем уменьшается, стимулируя учёного работать дальше, в то время как остальные остаются неизменными. AR-индекс вычисляется следующим образом:

$$AR = \sqrt{\sum_{j=1}^{h} \frac{cit_j}{a_j}},$$

где $a_j, j = \overline{1,h}$ – год опубликования j-ой работы.

В работе [105] предложен индекс цитирования, учитывающий скрытую диффузию научных знаний. По мнению авторов, данный индекс позволяет легко идентифицировать скрытых инициаторов научного мэйнстрима. Диффузия знаний определена в [20] как подход на основе сетевого анализа индивидуальных цитирований. Предложенный показатель имеет две составляющих:

1. описывает видимую диффузию научных знаний;
2. описывает скрытую диффузию научных знаний и выражается через количество неявных цитирований.

Особенностью индекса является введение составляющей, учитывающей диффузию знаний. Величина индекса цитирования вычисляется следующим образом:

$$\begin{gathered} C_j = D_j + \alpha \cdot I_j, \\ D_j = |T_j|, \\ j = \overline{1,M}, \end{gathered}$$

где M – общее число публикаций, $\alpha \in [0;1]$ – весовой коэффициент важности неявного цитирования, D_j – обычный индекс цитирования, I_j – индекс

неявного цитирования.

$$I_j = \sum_{\forall i \in T_j, j \notin F_i} \frac{|T_i|}{N_i},$$

где $A_j = \langle j, F_j, T_j \rangle$ – публикация, $F_j = \left\{F_j^1, F_j^2, F_j^3, \ldots\right\}$ – множество номеров публикаций, на которые ссылается работа A_j, $T_j = \left\{T_j^1, T_j^2, T_j^3, \ldots\right\}$ – множество номеров публикаций, в которых цитируется робота A_j, $N_i = |F_i|$ – длина списка литературы в i-ой публикации.

В то же время все сильнее звучит критика подобного сугубо формального статистического подхода для оценки эффективности научной деятельности учёного, научного коллектива, организации и т.д. Критика относится не только к новым системам индексации (например, РИНЦ, обсуждение которого в российской научной среде очень заметно), но и к таким системам, как Web of Science или Scopus. Утверждается, что показатели цитируемости исследователей, импакт-факторов журналов часто не соответствуют истинному положению дел ни в отношении самой статистики, ни в отношении качественности исследований, порождают негативные процессы в науке, такие как citation-fishing, citation-bartering, множество форм манипулирования импакт-фактором (редакционное давление на авторов, «взращивание» собственных авторов, обзорные статьи с многочисленными ссылками и др.), дискриминацию национальных наук. Всё это подробно изложено в специальном Докладе Международного Союза математиков, анализирующем основанные на статистике цитируемости показатели и предостерегающий от их широкого использования [2; 30].

Несмотря на всю критику в адрес формальных методов оценки эффективности научной деятельности, данные методы позволяют делать мгновенные «срезы» состояния научных исследований и планировать административную поддержку отдельным учёным или научным коллективам. В то же время индексы, вычисляемые на основе публикационной активности и цитируемости за определенные промежутки времени, дают показатели эффективности результатов научной деятельности, но не могут быть положены в основу планирования собственной исследовательской деятельности. Тот факт, что исследователя А сегодня процитировали N раз, не дает основание утверждать, что через год его процитируют столько же раз, т.е. фактор цитируемости, взятый

сам по себе, не может использоваться для планирования своей дальнейшей деятельности.

Однако, нельзя отрицать важность наукометрических показателей как стимулов для агентов научного производства. Так, упомянутая компания Thomson Reuters регулярно проводит награждение успешных учёных, при этом успешность оценивается по индексу цитируемости в системе Web of Science. Более того, существуют рейтинги на основе наукометрических показателей для агентов научного производство разного уровня, вплоть до уровня государств. Например, проект SCImago Journal & Country Rank (SJR) ведёт рейтинг стран (и научных журналов) по данным системы Scopus [41]. В таблице 3 приведена часть данного рейтинга.

Таблица 3 — Рейтинг стран проекта SJR

№	Страна	ОПА	Цитирований	Ср. знач.	h-индекс
1	США	7.281.575	152.984.430	22,02	1.518
2	Китай	3.095.159	14.752.062	6,81	436
3	Великобритания	1.932.907	37.450.384	19,82	934
4	Германия	1.876.342	30.644.118	17,39	815
5	Япония	1.874.277	23.633.462	13,01	694
6	Франция	1.348.769	21.193.343	16,85	742
7	Канада	1.040.413	18.826.873	20,05	725
8	Италия	1.015.410	15.317.599	16,45	654
9	Индия	825.025	5.666.045	8,83	341
10	Испания	800.214	10.584.940	15,08	531
11	Австралия	723.460	11.447.009	18,24	583
12	Южная Корея	642.983	5.770.844	11,49	375
13	Российская Федерация	629.671	3.664.726	6	355

Вместе с критикой, в качестве альтернативы формальному наукометрическому подходу часто называют другое направления оценки эффективности научной деятельности – качественного экспертного анализа (метода экспертных оценок). Однако методы данного направления так же не лишены недостатков. В частности, они сопряжены с необходимостью привлечения экспертов, поэтому сложно представить их использование для объективной оценки результатов научной деятельности в больших масштабах. Тем не менее, такой опыт есть. Например, проект Австралийского исследовательского совета (Australian Research Council), который включает качественную оценку 20000

научных журналов различных дисциплин на основе рецензирования привлечёнными научными академиями, сообществами, исследователями и экспертами. На основе данного проекта был составлен рейтинг качества журналов (Excellence in Research Australia, ERA) 2010 года. Анализ данного рейтинга показал, что экспертная оценка многих журналов не соответствует их импакт-фактору [2].

Следует отметить, что многие статистические подходы так же не охватывают всю науку, например импакт-фактор учитывает цитирования из ограниченной базы журналов. При этом попадание журнала в базу данных зависит от его импакт-фактора.

Ф. Кемпбелл (главный редактор журнала Nature) в своей работе [8] утверждает, что существующая система оценки результатов должна претерпеть существенные изменения. В частности, ключевым направлением является увеличение детализации оценки, например в виде возможности цитирования подразделов статьи. Так же, Кемпбелл видит перспективу в отказе от «журнальных брендов» благодаря сервисам, подобным проекту PLoS ONE [38]. PLoS ONE – это сетевой журнал, издающийся Публичной научной библиотекой (the Public Library Science). Работы в этом журнале публикуются в открытом виде, при этом добавляется возможность комментирования работ. При таком подходе рецензентом может стать любой желающий. В то же время, редакции не приходится выбирать работы для публикации в очередном номере, т.к. природа сетевого журнала не накладывает каких-либо ограничений на количество публикуемых работ.

Можно так же заметить, что нарастает тенденция публикации т.н. «препринтов» – подготовленных для публикации работ, независимо от журналов. Такой подход позволяет обеспечить свободный доступ к самой работе и вынести её на обсуждение и оценку сообществом. Это направление делает результаты научной деятельности более доступными широкой общественности и позволяет избавиться от некоторых недостатков научных журналов, например коммерческой составляющей, ограниченному тиражу и сложности опубликования. В то же время, у данного направления есть и свои недостатки. К ним относится большое число псевдонаучных публикаций и отсутствие какой-либо объективной оценки публикуемых работ. Существующие сервисы компенси-

рует эти недостатки различными способами. Так, arXiv [3] – крупнейший бесплатный архив электронных публикаций научных статей естественнонаучных направлений, с 2004 года использует систему предварительного подтверждения статуса публикуемых работ.

1.3 Математические методы в управлении научной деятельностью и оценке её результатов

Можно выделить большую группу исследований в области управления научной деятельностью и оценки её результатов, использующих методы математического моделирования. Не смотря на то, что многие исследователи отдают предпочтение эмпирическим методам [64; 78], в данном направлении, как и во многих других, в последнее время проводится всё больше исследований, активно использующие методы математического моделирования. Данный процесс обусловлен как развитием математического аппарата в целом, так и прикладного аппарата в частности. Важную роль так же сыграло развитие вычислительной техники, позволяющей эффективно обрабатывать большие объёмы данных и решать сложные вычислительные задачи. Очевидно, математическое моделирование является необходимым инструментом в применении вычислительной техники для решения разнообразных задач, в т.ч. задач управления научной деятельностью.

В работе [67] проведено комплексное исследование проблем взаимодействия государства и науки, в том числе формирования научной политики. В частности, рассмотрены существующие модели экономического роста. В общем виде производственная функция (укрупнённая модель экономического роста), учитывающая сектор производства научных и технических знаний может быть записана следующим образом:

$$Y = F(K, L, k, h),$$

где K – совокупный производственный капитал, L – совокупные производственные затраты труда, k – совокупный интеллектуальный капитал, h – совокупные затраты интеллектуального труда.

Соответствующее равенство Эйлера:

$$F(K, L, k, h) = \frac{\partial F}{\partial K}K + \frac{\partial F}{\partial L}L + \frac{\partial F}{\partial k}k + \frac{\partial F}{\partial h}h,$$

где $\frac{\partial F}{\partial K}$ – равновесная процентная ставка на производственный капитал, $\frac{\partial F}{\partial L}$ – равновесная ставка реальной заработной платы производственного персонала, $\frac{\partial F}{\partial k}$ – равновесная процентная ставка на интеллектуальный капитал, $\frac{\partial F}{\partial h}$ – равновесная ставка реальной заработной платы работников интеллектуального труда.

В качестве примеров производственной функции Y в [67] рассмотрены следующие модели:

1. Модель Узавы (1965 г.):

   $$Y(t) = F(K(t), A(t), L_1(t)),$$

   где $K(t)$ – капитал в год t, $A(t)$ – эффективность труда, $L_1(t)$ – количество труда в материальном секторе в год t. При этом образовательный сектор воздействует через величину $A(t)$:

   $$\frac{dA}{dt} = A(t)f\left(\frac{L_2(t)}{L(t)}\right),$$

   где $L_2(t)$ – количество труда, занятого в образовательном секторе, $L(t)$ – общее количество труда в год t.

   В модели Узавы управление выпуском продукции осуществляется распределение трудовых ресурсов между между производством и образованием, а продукция распределяется между инвестициями и потреблением.

2. Статистическая Модель Барро (1991 г.), построенная на основе анализа показателей большого количества стран, определяет зависимость изменения валового внутреннего продукта от отношения числа учащихся к общему числу жителей страны:

   $$Y(t) = 0.0302 - 0.0075Y(0) + 0.025h_0(t) + 0.0305h_1(t) - 0.119g(t),$$

где $Y(0)$ – начальный уровень душевого потребления, $h_0(t)$ – степень охвата населения начальным образованием, $h_1(t)$ – степень охвата населения средним образованием, $g(t)$ – доля правительственного потребления.

3. В модели Ромера (1986 г.) рассматривается совокупность предприятий, характеризующихся производственной функцией вида $F(h, H, x)$, где h – знания предприятия, $H = \sum h$ – совокупные знания предприятий, x – ресурсы предприятий (капитал и труд). Если совокупность x постоянна, то влияние на выпуск интеллектуальной продукции определяется производственной функцией следующего вида:

$$f = k^a K^b,$$

где $a, b > 0$, $a + b > 1$.

4. В модели Лукаса (1986 г.) вводится следующая производственная функция:

$$Y = rK^a(ukL)^{1-b}k_0^c,$$

где r, a, b, c – статистические параметры, L – численность рабочей силы, k – уровень знаний работника моделируемого предприятия, k_0 – уровень знаний среднего работника в стране, $u(t)$ – доля труда в материальном производстве, $K(t)$ – физический капитал в год t.

Анализ рассмотренных моделей позволил их авторам разработать рекомендации для построения оптимальной траектории производственной функции Y.

В работе [83] решается задача планирования портфеля научных проектов [79] с применением теоретико-игровых моделей и теории управления.

Для этого рассматривается четырёхуровневая модель системы управления научными проектами, включающая следующие уровни:

1. высшее руководство (ВР);
2. функциональные руководители (ФР);
3. руководители научных проектов (РП);

4. исполнители.

В рассматриваемой модели использованы следующие обозначения:

- $N = \{1, 2, \ldots, n\}$ – множество агентов (исполнителей);
- $K = \{1, 2, \ldots, k\}$ – множество руководителей проектов;
- $M = \{1, 2, \ldots, m\}$множество функциональных руководителей;
- $y_i \in A_i \subseteq \mathfrak{R}^{n_i}, 0 \in A_i$ – действие i-го исполнителя, $i \in N$;
- $y = (y_1, y_2, \ldots, y_n) \in A' \subseteq \mathfrak{R}^{\sum_{i \in N} n_i}$ – вектор действий исполнителей;
- $c_i(y)$ – функция затрат i-го исполнителя, $i \in N$;
- $h_j(y)$ – функция дохода j-го руководителя проекта, $j \in K$;
- $H_l(y)$ – функция дохода l-го функционального руководителя, $l \in M$;
- $H_0(y)$ – функция дохода высшего руководства;
- $\sigma_{ij}(y)$ – функция стимулирования i-го агента со стороны j-го руководителя проекта, $i \in N, j \in K$;
- $v_{il}(y)$ – функция стимулирования i-го агента со стороны l-го функционального руководителя, $i \in N, l \in M$;
- $u_{jl}(y)$ – функция стимулирования j-го руководителя проекта со стороны l-го функционального руководителя, $j \in K, l \in M$;
- $s_j(y)$ – функция стимулирования j-го руководителя проекта со стороны высшего руководства, $j \in K$;
- $q_l(y)$ – функция стимулирования l-го функционального руководителя со стороны высшего руководства, $l \in M$.

Целевые функции (ЦФ) участников системы выглядят следующим образом:

- $f_i(\sigma_i, v_i, y) = \sum\limits_{j \in K} \sigma_{ij}(y) + \sum\limits_{l \in M} v_{il}(y) - c_i(y), i \in N$ – ЦФ агентов (исполнителей);

- $F_j(\sigma_j, u_j, s_j, y) = h_j(y) + \sum\limits_{l \in M} u_{jl}(y) + s_j(y) - \sum\limits_{i \in N} \sigma_{ij}(y), j \in K$ – ЦФ руководителей проектов;

- $\Phi_l(q_l, u_l, v_l, y) = H_l(y) + q_l(y) - \sum\limits_{j \in K} u_{jl}(y) - \sum\limits_{i \in N} v_{il}(y), l \in M$ – ЦФ функционального руководителя;

- $\Phi_0(y, s, q) = H_0(y) - \sum\limits_{l \in M} q_l(y) - \sum\limits_{j \in K} s_j(y)$ – ЦФ высшего руководства;

Рассмотренная задача неразрешима в общем случае, однако в работе [83] предложено решение рассмотренной задачи с применением принципа компенсации затрат [82], позволяющее получить согласованный план, максимизирующий сумму ЦФ всех участников системы.

В третьем типе задач управления научной деятельностью распространены работы, изучающие социологические и психологические аспекты научной деятельности с применением методов математического моделирования. Так, в работе [65] предложена когнитивная модель структуры личности как участника работы над научным проектом. В данной модели личность представлена в виде триады концептов:

- Y_1 – темп интеллектуальной деятельности;

- Y_2 – темп организационно-трудовой деятельности;

- Y_3 – темп изменения психологического состояния личности.

Модель представлена на рисунке 5.

где:

- k_{21} – отражает влияние интеллекта на трудовую деятельность;

- k_{12} – отражает активизацию трудовой деятельностью интеллекта;

- k_{13} – отражает влияние психического состояния на интеллектуальную деятельность;

- k_{31} – отражает влияние интеллекта на изменение состояния психики личности;

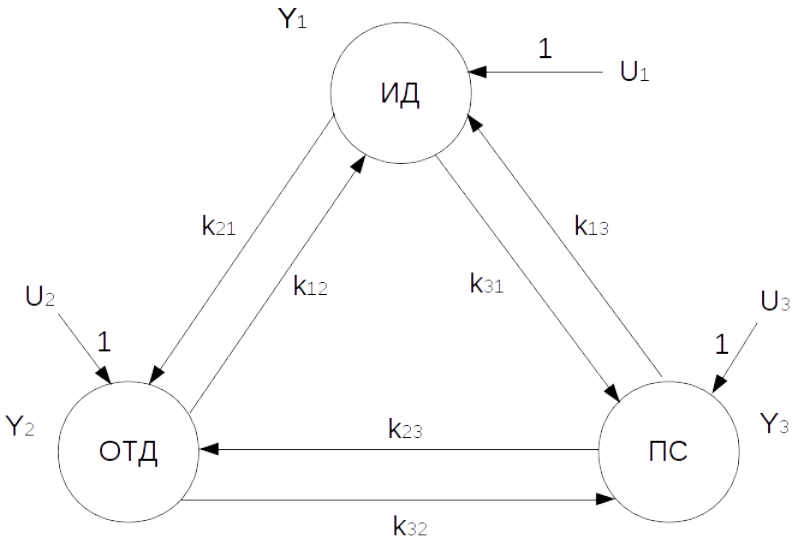

Рисунок 5 — Когнитивная модель структуры личности

- k_{23} – отражает влияние изменения состояния психики на трудовую деятельность;
- k_{32} – отражает влияние трудовой деятельности на изменение состояния психики.

Рассматриваемая модель описывается следующей системой уравнений:

$$T_1\dot{Y_1} + Y_1 = k_{12}Y_2 + k_{13}Y_3 + U_1,$$
$$T_2\dot{Y_2} + Y_2 = k_{21}Y_1 + K_{23}Y_3 + U_2,$$
$$T_3\dot{Y_3} + Y_3 = k_{31}Y_1 + k_{32}Y_2 + U_3,$$

где T_i – постоянные времени, учитывающие инерционность изменения концептов, $U = (U_1, U_2, U_3)$ – вектор внешних воздействий.

Решение этой системы уравнений в статике имеет вид:

$$\Delta Y_i = \Delta_i U,$$

где Δ – главный определитель Крамера.

Так же в работе [65] проведено моделирование, на основе результатов которого установлены свойства модели структуры личности и сделан вывод о возможности её исследования для анализа поведения малой научной группы.

В работе [72] рассмотрены социологические аспекты проблемы управ-

ления коллективным агентом научного производства с применением математического моделирования. В частности, решается задача управления социального управления структурой научного коллектива. Социальным управлением научно-исследовательским коллективом называется некоторым образом ориентированное воздействие на коллектив с целью приведения его в требуемое состояние.

Предварительно вводится модель научно-исследовательского коллектива в узком смысле:

$$\mathcal{M} = \langle \mathfrak{B}, \mathcal{R} \rangle ,$$

где где n-местное отношение $\mathcal{R}^n$ есть подмножество прямого произведения $\mathcal{R}^n = \mathcal{R}_1 \times \mathcal{R}_2 \times \ldots \times \mathcal{R}_n$, т.е. отношение ранга n представляет собой множество n упорядоченных наборов $\{r_1, r_2, \ldots, r_n\}$.

В общем случае модель научно-исследовательского коллектива представляется парой:

$$\mathcal{M} = \langle \mathfrak{G}, \mathsf{B} \rangle ,$$

где $\mathfrak{G}$ – структура коллектива, трактуемая как функция плотности вероятности социальных различий, а B – множество параметров, описывающих коллектив.

Следует отметить, что в рассмотренной работе подчёркивается теоретический характер полученных результатов.

Помимо рассмотренных выше индексов цитирования особого внимания заслуживает стохастическая модель процесса деятельности учёного, предложенная в работе [6]. Данная модель представляет процесс публикации новых работ учёным и последующее цитирование этих публикаций как процессы Пуассона [101]. Такой подход позволяет исследовать различные ситуации благодаря возможности изменения параметров интенсивности публикования новых работ, их цитирования и продолжительности карьеры учёного.

Модель абстрагирована от вопросов точного определения времени, например какой момент считать началом карьеры учёного? Возможны следующие варианты:

- вступление в должность;
- отправка первой публикации в редакцию;

- принятие первой публикации к печати;
- выход журнала с первой публикацией.

Вместо этого предлагается считать, что на данный момент учёный находится в T единице времени своей карьеры. За единицу времени могут приниматься различные величины, но обычно это год. Процессы, описываемые моделью, схематично представлены на рисунке 6:

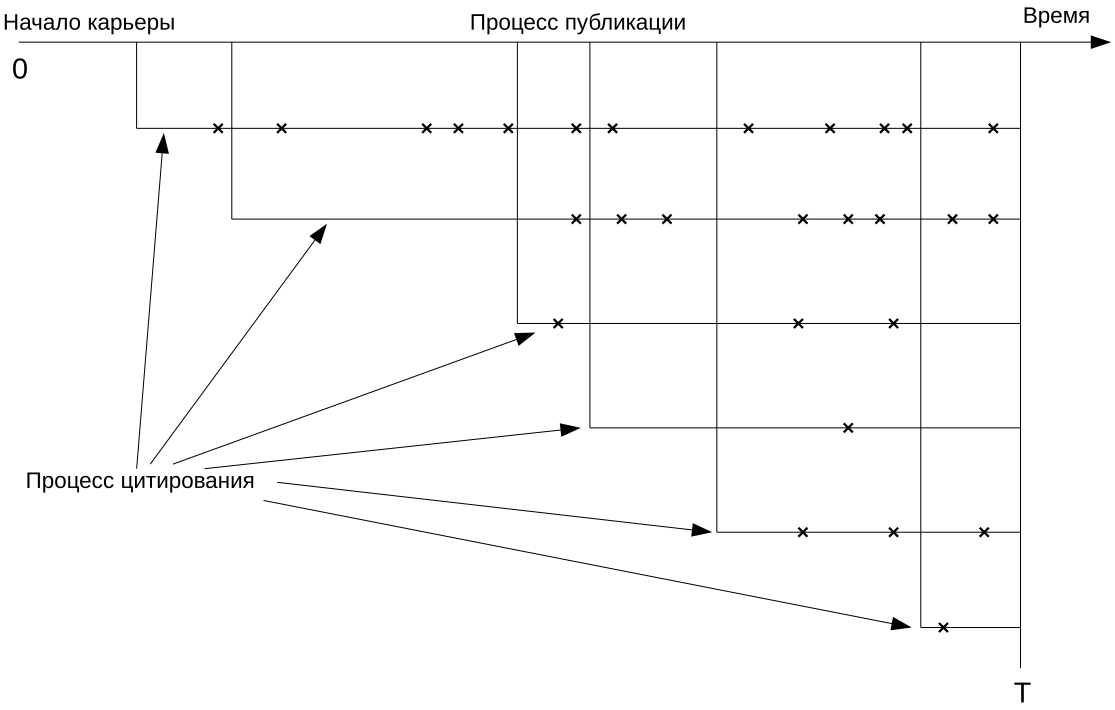

Рисунок 6 — Представление процессов публикации и цитирования

В стохастической модели используются следующие предположения:

1. С момента начала карьеры учёного (в нулевой момент времени) новые работы публикуется согласно процессу Пуассона с интенсивностью θ, что является средним количеством публикаций в единицу времени, называемое интенсивностью публикации (publication rate). Тогда количество публикаций Y_T учёного в момент времени T имеет следующее распределение:

$$P(Y_T = r) = e^{-\theta T}\frac{(\theta T)^r}{r!}, \tag{2}$$

где $r = 0, 1, 2, \ldots$ и $E[Y_T] = \theta T$.

2. Каждая публикация цитируется согласно процессу Пуассона с интенсивностью Λ, где Λ изменяется от публикации к публикации. Таким образом, Λ обозначает среднее количество цитирований публикации в единицу времени, называемое интенсивностью цитирования.

3. Интенсивность цитирования Λ для данного автора изменяется между его публикациями согласно гамма-распределению [59] с параметрами $\nu \geqslant 1$ (форма) и $\alpha > 0$ (масштаб). Тогда функция плотности вероятности для Λ определяется следующим образом:

$$f_\Lambda(\lambda) = \frac{\alpha^\nu}{\Gamma(\nu)} \lambda^{\nu-1} e^{-\alpha\lambda}, \tag{3}$$
$$0 < \lambda < \infty.$$

   Заметим, что $E[\Lambda] = \frac{\nu}{\alpha}$ есть общая средняя интенсивность цитирований или среднее количество цитирований одной случайной выбранной публикации данного учёного за единицу времени.

Простейшим результатом, полученным на основе модели, является распределение величины X_T – количество цитирований случайно взятой (выбранной) публикации данного учёного за время T. Учитывая вышеописанные предположения, данное распределение вводится следующим образом:

$$P(X_T = r) = \frac{\alpha}{(\nu-1)T} B\left(\frac{T}{\alpha+T}; r+1, \nu-1\right),$$
$$r = 0, 1, 2, \ldots$$

где

$$B(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x y^{a-1}(1-y)^{b-1} dy$$

есть функция распределения для бета-распределения первого типа с параметрами a и b.

Кроме того, возможно оценить ожидаемое число публикаций, получив-

ших как минимум n цитирований за время T:

$$E\left[N(0;T)\right] = \theta T,$$
$$E\left[N(n;T)\right] = \theta T\left(1 - \frac{\alpha}{(\nu-1)T}\sum_{r=0}^{n-1} B\left(\frac{T}{\alpha+T}; r+1, \nu-1\right)\right),$$
$$n = 0, 1, 2, \ldots$$

Доказательства приведены в [6].

Стохастическая модель была изначально предложена в 1992 году в работе [5] а в 2007 году применена для исследования поведения h-индекса [6]. Позже на её основе в работе [7] было проведено исследование изменения величин ядра h-индекса и A-индекса с течением времени (относительно увеличения h-индекса). В частности, авторами были сделаны выводы, что h-индекс, при прочих равных:

1. приблизительно пропорционален к текущей продолжительности карьеры учёного T;
2. приблизительно равен линейной функции логарифма интенсивности публикации учёного;
3. приблизительно равен линейной функции логарифма средней интенсивности цитирований публикаций учёного.

Можно выделить широкий класс методов на основе графических моделей [71; 75; 84; 103]. К этому классу относятся методы управления научной деятельностью и оценки её результатов, так или иначе использующих графы и сетевой анализ [10; 37]. В частности, наиболее распространены т.н. научные карты.

Понятие научной карты применяют к достаточно обширному списку методов представления данных о состоянии различных научных систем. Обычно подобные карты отображают взаимосвязи между различными элементами моделируемой системы, например между научными направлениями.

Наиболее крупным и успешным проектом, реализующим подобную концепцию, является Map Of Science от SciTech Strategies [32]. Целью проекта является создание глобальной научной карты на основе данных о научных пуб-

ликациях из разных областей науки. В качестве источника данных о публикациях проект использует библиографическую базу данных Institute for Scientific Information, с 2004 года поек получил доступ к базе данных Scopus, а с 2012 года - к базе USPTO (база патентов США). Большой объём данных, доступных проекту, обуславливает высокую точность научной карты. Полученная карта используется различными организациями в целях анализа и прогнозирования состояния как науки в целом, так и её отдельных объектов. Для построения научных карт проект Map Of Science использует два основных подхода: анализ связей публикаций на основе прямого цитирования («Academic Lineage») и кластеризацию на основе совместного цитирования («Problem Genealogies»).

Авторами Map of Science проделана большая работа по совершенствованию методики построения научных карт и её приложений. Так, в работе [33] предложена новая универсальная модель, позволяющая визуализировать глобальную научную карту, включая как гуманитарные, так и естественные науки. В качестве преимуществ авторы отмечают возможность визуальной оценки ключевых предметных областей науки, их размера, подобия и междисциплинарных связей. Авторы подчёркивают важность достигнутой структурной точности полученной модели. Кроме того, в работе предложены несколько метрик подобия научных документов. Пример полученной карты с одной из таких метрик приведён на рисунке 7.

В другой работе [44] описана методика выявления формирующихся направлений в науке и технике на основе научных карт. Позиционируется авторами как представляющая интерес для лиц принимающих решения (ЛПР) на государственном уровне а также в индустрии. В работе рассматриваются 2 модели на основе прямых цитирований и совместных цитирований научных документов, на которых производится кластерный анализ. Показано, что в результате применения предложенной методики получается список формирующихся направлений, пригодный для изучения ЛПР.

Ещё одна работе [28] посвящена изучения проблем переходов между локальными и глобальными научными картами, строящимися на основе научных документов. Предложен метод, позволяющий увеличить точность как локальных, так и глобальных карт.

Существуют и другие подходы на основе графических моделей. В част-

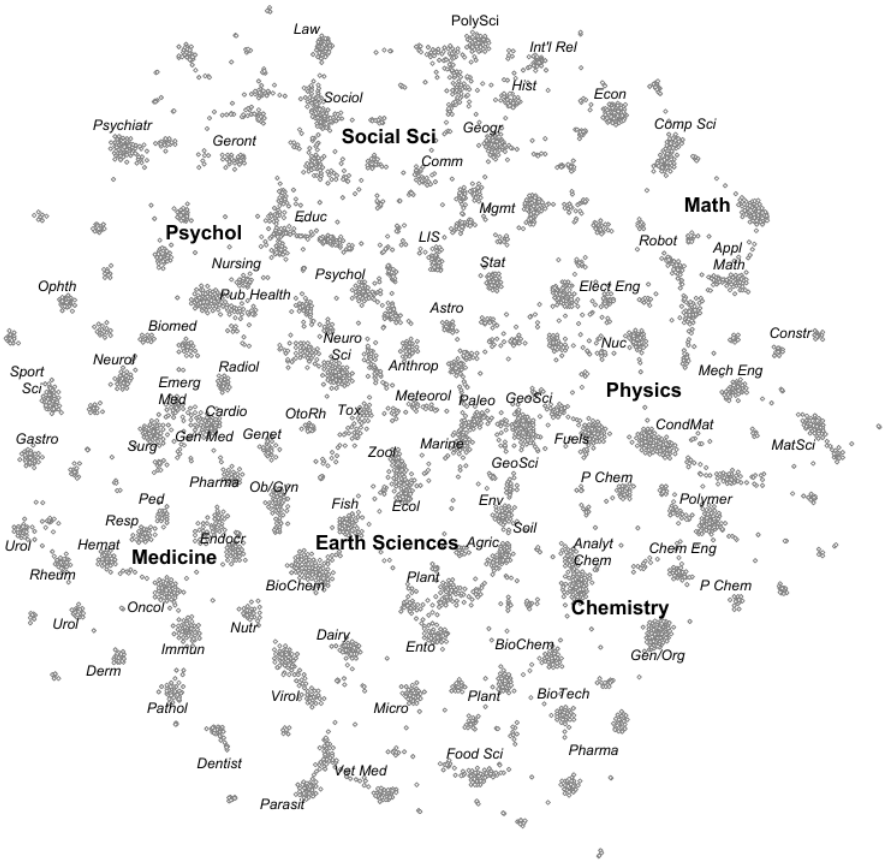

Рисунок 7 — Пример научной карты

ности, интересен подход, включающий круг исследований, так или иначе использующих содержимое опубликованных работ (текст, аннотации или ключевые слова). Обычно, подобные модели используются для изучения конкретных предметных областей. Например, в работе [34] предложена графическая модель, позволяющая оценить значимость отдельных терминов и их взаимосвязей для когнитивной нейробиологии. Авторы считают полученные результаты полезными для выявления перспективных научных направлений а так же систематических отклонений в проведении и представлении исследований.

1.4 Выводы и постановка цели исследования

Проведённый анализ существующих подходов, методов и математических моделей, используемых при решении задач управления научной деятельностью, показал востребованность решений, позволяющих повысить эффективность деятельности агентов научного производства. Существующие модели рассматривают различные аспекты управления научной деятельностью, например такие как распределение ресурсов, социальные и психологические

факторы в научном коллективе. Однако, большинство существующих методов и моделей не учитываю особенности предметных областей, разрабатываемых агентами научного производства а также процесс их изменения с течением времени, т.е. исследовательскую траекторию. Оптимизируя исследовательскую траекторию по некоторому критерию, позволяющему оценить результативность агента научного производства, можно добиться существенного повышения его эффективности.

В то же время, многие существующие методы повышения эффективности деятельности агентов научного производства подразумевают использование некоторого фиксированного способа оценки результатов научной деятельности, что сужает потенциальную сферу их применения. Очевидно, выбор метода оценки результатов научной деятельности является важным этапом в решении задачи повышения эффективности деятельности агентов научного производства. В ходе проведённого анализа существующих методов оценки научной деятельности были выделены 3 класса по способу интерпретации результатов:

1. Методы на основе наукометрических показателей. Часто критикуются за недостоверность и прочие недостатки, тем не менее, активно применяются во многих организациях.

2. Методы на основе экспертных оценок. Трудоёмки, однако позволяют получить более качественные оценки, однако могут обладать недостатком субъективности.

3. Интерактивные методы, основанные на применении интернет-сервисов для публикации научных работ. Относительно новое направление. Фактически, обладает преимуществами экспертных методов, но в меньшей мере подвержено субъективности, т.к. в оценка может производиться большим количеством экспертов.

На основе анализа существующих подходов к оценке результатов научной деятельности можно сделать вывод, что не смотря на значительную критику, подходы на основе наукометрических показателей остаются самыми широко распространёнными. Так же показано, что каждый подход имеет свои пре-

имущества и недостатки. Очевидно, для разных задач подходят разные методы оценки результатов научной деятельности, что необходимо учитывать при разработке новых методов и моделей управления научной деятельностью, т.е. при разработке новой методики должна обеспечиваться достаточная универсальность.

В результате была сформулирована цель данного исследования: разработка математических моделей, алгоритмического и программного обеспечения для моделирования, прогнозирования и оптимизации исследовательских траекторий агентов научного производства.

Для достижения поставленной цели решаются следующие задачи:

1. Провести анализ существующих подходов, методов и математических моделей, используемых при решении задач управления научной деятельностью. Провести классификацию задач управления научной деятельностью и методов оценки результатов научной деятельности.

2. Разработать математическую модель предметной области агента научного производства.

3. Разработать математическую модель исследовательской траектории агента научного производства.

4. Разработать методику построения оптимальных исследовательских траекторий агентов научного производства.

5. Разработать информационную систему для моделирования, оценки и оптимизации исследовательских траекторий агентов научного производства.

6. Провести апробацию полученных результатов на задачах построения оптимальных исследовательских траекторий для двух различных агентов научного производства: научного журнала и научного коллектива.

Задача 1 решена в 1 главе, задачи 2-4 решаются во второй главе. Решению 5 задачи посвящены третья глава. Задача 6 рассматривается в четвёртой главе.

2 Разработка методики построения оптимальной исследовательской траектории

2.1 Разработка графосемантической модели предметной области

Графосемантическое моделирование представляет собой метод моделирования предметных областей посредством выделения структурных связей между семантическими компонентами, составляющими семантические поля моделируемых областей. Основное условие, позволяющее использовать описываемый метод – наличие таких связей между компонентами. Метод графосемантического моделирования позволяет представить набор данных в виде системы, в которой каждый из компонентов имеет иерархическую и топологическую определенность по отношению к другим компонентам и всей системе в целом. Эта структурная контекстуальность, в свою очередь, позволяет интерпретировать каждый компонент системы. Метод графосемантического моделирования использовался во многих научных и прикладных областях:

- для реконструкции имиджевых портретов и аудитории известных брендов и компаний на основе мнений Интернет–пользователей;
- для создания медиа–планов рекламных кампаний;
- для мониторинга СМИ и Интернета по актуальным проблемам образования и здравоохранения для Министерства образования и науки РФ и Министерства здравоохранения и социального развития РФ;
- для оценки эквивалентности и адекватности переводных текстов оригинальным;
- для когнитивного моделирования процессов аналитической деятельности;
- для реконструкции образа мира и медиа–образов российской политической элиты;

- для построения информационно–когнитивных моделей научных картин мира и др. (обзор применений метода графосемантического моделирования [56]).

Основными понятиями метода графосемантического моделирования являются:

- Контекст – основная структурная сущность модели предметной области, определяет связи между смысловыми компонентами предметной области;
- Метатип – тип данных, определяющий конкретную описательную характеристику контекстов (источник, территориальная принадлежность, возраст и т.д.);
- Метаполе – значение определённого метатипа для данного контекста;
- Семантический компонент – сущность, характеризующая существование определённой семантики (значения, объекта и т.п.) в контексте;
- Семантическое поле – сущность, включающая множество компонентов, через которые определяется её присутствие в контекстах (в обобщённом методе может включать другие поля, определяя иерархию);
- Семантическая карта – двумерная матрица, содержащая численные характеристики сил взаимных связей семантических полей (в обобщённой модели может быть многомерной и определяться не только для полей);
- Семантический граф – способ представления семантической карты, позволяющий наглядно оценить связность семантических полей.

Метод графосемантического моделирования состоит из следующих шагов:

1. Сбор исходных данных, описывающих предметную область (контекстов);
2. Ввод описательной информации (заполнение метаполей) в собранных контекстах;

3. Выделение семантических компонентов в контекстах;
4. Объединение семантических компонентов в семантические поля;
5. Формирование выборок контекстов на основе метаполей;
6. Вычисление семантических карт;
7. Визуализация (построение семантического графа);
8. Анализ полученных данных.

Для наглядного представления описываемого метода, рассмотрим пример графосемантической модели, схема которой представлена на рисунке 8.

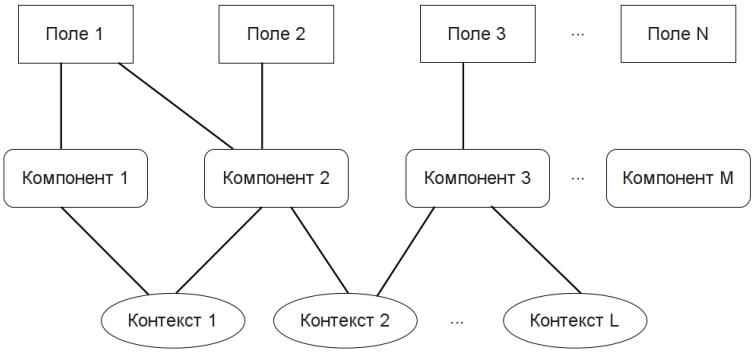

Рисунок 8 — Схематичное представление примера графосемантической модели

Как видно из рисунка 1, в данной модели присутствуют L контекстов, M компонентов и N полей. Следует учитывать, что представленная на рисунке 1 схема является упрощённой. В частности, на схеме не приведены мета–поля.

Введём формальное математическое описание графосемантической модели:

- S – множество символьных строк неограниченной длины;
- $\Sigma \subset S$ – множество контекстов;
- $C \subset S$ – множество компонентов;
- $F \subset S$ – множество семантических полей;
- T – множество типов мета–полей;

- $\Psi(\Sigma, T) = \{\psi : \psi(\sigma, t), \forall t \in T, \forall \sigma \in \Sigma\}$ – функция, определяющая множество мета–полей для данного множества контекстов;

- $\Phi(C, \Sigma) = \{\phi : \phi(c, \sigma) \in 0, 1, \forall c \in C, \forall \sigma \in \Sigma\}$ – функция, определяющая множество связей компонентов с контекстами;

- $\Lambda(C, F) = \{\lambda : \lambda(c, f) \in \{0, 1\}, \forall c \in C, \forall f \in F\}$ – функция, определяющая множество связей компонентов с полями;

- $\Gamma(F, \Sigma) = \{\gamma : \gamma(f_1, f_2, \Sigma) \in \mathbb{N}, \forall f_1, f_2 \in F, f_1 \neq f_2\}$ – функция, определяющая множество числовых значений (сил, весов) связей между полями;

- $\Omega(\Sigma, C, F, T, \Psi(\Sigma), \Phi(C, \Sigma), \Lambda(C, F), \Gamma(F, \Sigma))$ – графосемантическая модель.

Заметим, что в задаче моделирования предметных областей, множество семантических полей F является общим для всех предметных областей, принадлежащих одной научной отрасли.

Так же можно обозначить множество Z, содержащее факты существования связей полей с контекстами. Такое множество вычисляется на основе множеств F и Σ, и является вспомогательным для нахождения $\Gamma(F)$

$$Z(F, \Sigma) = \{\zeta : \zeta(f, \sigma) \in \mathbb{N}, \forall f \in F, \forall \sigma \in \Sigma\}.$$

Определим основные функции, используемые в обозначениях:

- $\psi(\sigma, t), t \in T, \sigma \in \Sigma$ – функция, определяющая значение метаполя типа t для контекста σ;

- $\phi(c, \sigma) = \begin{cases} 0, & c \notin \sigma \\ 1, & c \in \sigma \end{cases}, c \in C, \sigma \in \Sigma$– функция, определяющая факт существования связи между компонентом c и контекстом σ, где $c \in \sigma, \forall c \in C, \forall \sigma \in \Sigma$ означает присутствие компонента c в контексте σ;

- $\lambda(c, f) = \begin{cases} 0, & c \notin f \\ 1, & c \in f \end{cases}, c \in C, f \in F$ – функция, определяющая факт

существования связи между компонентом c и полем f, где $c \in f, \forall c \in C, \forall f \in F$ означает присутствие (связь) компонента c в поле f;

- $\zeta(f, \sigma) = \sum\limits_{c \in C} \phi(c, \sigma) \cdot \lambda(c, f), f \in F, \sigma \in \Sigma$ – функция, определяющая силу связи между полем f и контекстом σ через компоненты $c \in C$.

Таким образом, функция γ может быть определена двумя способами:

$$\gamma(f_1, f_2, \Sigma) = \left| \{\sigma : \sigma \in \Sigma, \zeta(f_1, \sigma) > 0, \zeta(f_2, \sigma) > 0\} \right|, \tag{4}$$

$$\gamma(f_1, f_2, \Sigma) = \sum_{\sigma \in \Sigma} \min\{\zeta(f_1, \sigma), 1\} \cdot \min\{\zeta(f_2, \sigma), 1\}.$$

Для рассмотренного примера описанные множества будут содержать следующие элементы:

- Σ= { «Контекст 1», «Контекст 2», . . . , «Контекст L» };
- C = { «Компонент 1», «Компонент 2», «Компонент 3», . . . , «Компонент M» };
- F = { «Поле 1», «Поле 2», «Поле 3», . . . , «Поле N» };
- T = { «Название» }.

Функции $\Psi(\Sigma, T)$, $\Phi(C, \Sigma)$, $\Lambda(C, F)$, $Z(F, \Sigma)$ и $\Gamma(F,\Sigma)$ могут быть представлены в табличном виде. В таблице 4 представлена функция $\Psi(\Sigma, T)$.

Таблица 4 — Функция $\Psi(\Sigma, T)$

$\Psi(\Sigma, T)$	Название
Контекст 1	«Контекст 1»
Контекст 2	«Контекст 2»
. . .	. . .
Контекст L	«Контекст L»

Замечание: множество T может включать произвольное количество строковых, целочисленных, временных и других типов данных.

В таблице 5 представлена функция $\Phi(C, \Sigma)$.

Таблица 5 — Функция $\Phi(C, \Sigma)$

$\Phi(C, \Sigma)$	Компонент 1	Компонент 2	Компонент 3	...	Компонент M
Контекст 1	1	1	0	...	0
Контекст 2	0	1	1	...	0
...	...	...	...	...	...
Контекст L	0	0	1	...	0

В таблице 6 представлена функция $\Lambda(C, F)$.

Таблица 6 — Функция $\Lambda(C, F)$

$\Lambda(C, F)$	Поле 1	Поле 2	Поле 3	...	Поле N
Компонент 1	1	0	0	...	0
Компонент 2	1	1	0	...	0
Компонент 3	0	0	1	...	0
...	...	...	...	...	...
Компонент M	0	0	0	...	0

В таблице 7 представлена функция $Z(F, \Sigma)$.

Таблица 7 — Функция $Z(F, \Sigma)$

$Z(F, \Sigma)$	Поле 1	Поле 2	Поле 3	...	Поле N
Контекст 1	1	1	0	...	0
Контекст 2	1	1	1	...	0
...	...	...	...	...	...
Контекст L	0	0	1	...	0

Табличное представление функции $\Gamma(F,\Sigma)$ называют семантической картой. В столбцах и строках такой таблицы записываются названия полей, а в ячейках – числовое значение силы связи между полями, т.е. число контекстов, в которых поля встретились вместе. Семантическая карта, соответствующая приведённому на рисунке 8 примеру, может быть представлена в виде таблицы 8.

Таблица 8 — Функция $\Gamma(F,\Sigma)$

$\Gamma(F,\Sigma)$	Поле 1	Поле 2	Поле 3	...	Поле N
Поле 1	–	2	1	...	0
Поле 2	2	–	1	...	0
Поле 3	1	1	–		0
...	...	...	...	...	...
Поле N	0	0	0	...	–

Очевидно, семантическая карта в большой степени зависит от остальных элементов графосемантической модели. В частности, семантическая карта строится на основе подмножества (выборки) контекстов. В терминах объектно–ориентированного программирования, $\Gamma(F,\Sigma)$ является методом модели Ω, а Σ, C, F, T, $\Psi(\Sigma,T)$, $\Phi(C,\Sigma)$, $\Lambda(C,F)$ – её свойствами. Очевидно, $Z(F,\Sigma)$ так же является методом Ω. Выборка контекстов производится на основе выборки мета–полей. Например, выборка мета–полей, для которых значением метатипа «age» («возраст») является 25, может быть записана следующим образом:

$$\Psi^* = \left|\{\psi : \psi = 25, \psi \in \Psi(\Sigma, \{"age"\})\}\right|.$$

Следовательно, выборку множества контекстов, содержащих выбранные метаполя, можно записать как

$$\Sigma^* = \{\sigma : \psi(\sigma,t) \in \Psi^*, t \in \{"age"\}, \sigma \in \Sigma\},$$

тогда $\Gamma(F,\Sigma^*)$ – семантическая карта на основе данной выборки.

Важным инструментом графосемантического моделирования является семантический граф. Семантический граф – граф, вершинами которого являются поля, а рёбра описывают связи между полями. Семантический граф можно рассматривать как мультиграф, в котором одна связь между полями соответствует одному ребру, однако обычно число связей используют в качестве веса ребра.

Семантический граф можно определить с помощью множества полей F и матрицы смежности A, которая строится на основе семантической карты

$\Gamma(F, \Sigma)$:

$$A = \{a_{ij}\}, i = \overline{1,n}, j = \overline{1,n}, \tag{5}$$

$$a_{ij} = \begin{cases} \gamma(f_i, f_j, \Sigma), & i \neq j, \gamma(f_i, f_j, \Sigma) > 0, \\ 0, & i \neq j, \gamma(f_i, f_j, \Sigma) = 0, \\ 0, & i = j. \end{cases} \tag{6}$$

$$n = \left|F\right|, \tag{7}$$

где n – число полей.

Определим основные свойства матрицы A:

$$a_{ij} = a_{ji}, \forall i = \overline{1,n}, j = \overline{1,n}, \tag{8}$$

$$a_{ij} = 0, i = j, \forall i = \overline{1,n}, j = \overline{1,n}, \tag{9}$$

$$a_{ij} \in \mathbb{Z}, \forall i = \overline{1,n}, j = \overline{1,n}, \tag{10}$$

$$a_{ij} \geqslant 0, \forall i = \overline{1,n}, j = \overline{1,n}. \tag{11}$$

Свойство (8) означает симметричность, (9) – нулевая главная диагональ, (10) – целые элементы, (11) – неотрицательные элементы.

Из (8–11) следует, что граф, который определяет матрица A, является неориентированным взвешенным графом без петель [84; 92; 103].

Матрица смежности A содержит веса ребёр графа, однако вершинам графа так же могут быть присвоены веса. В графосемантической модели весами графа являются частотности полей. Частотность поля – число контекстов, с которыми связано поле через произвольное число компонентов:

$$\nu(f,\Sigma) = \left|\{\sigma : \sigma \in \Sigma, \zeta(f, \sigma) > 0\}\right|, \forall f \in F, \tag{12}$$

или

$$\nu(f, \Sigma) = \sum_{\sigma \in \Sigma} \min\{\zeta(f, \sigma), 1\}. \tag{13}$$

Заметим, что

$$\nu(f, \Sigma) \in \mathbb{Z}, \forall f \in F. \tag{14}$$

Таким образом, на основе (5–13), семантический граф для модели Ω определяется тройкой:

$$G_\Omega = (V, A, \nu), V = \{f : \nu(f) > 0, f \in F\}, \tag{15}$$

где V – множество вершин графа.

Из вышесказанного следует, что семантический граф G_Ω является полносвязным неориентированным графом без петель, следовательно, количество уникальных элементов в матрице A, равное количеству рёбер графа G_Ω, равно $\frac{n(n-1)}{2}$.

Поскольку в прикладных исследованиях графосемантическая модель может включать большое количество полей, построение графа и его визуализация может быть очень трудоёмким процессом. Наглядность – важное требование к результирующему графу, поскольку он является одним из ключевых объектов анализа. Как следствие, граф строится в полуавтоматическом режиме с использованием специальных программных средств, например Gephi. Семантический граф, соответствующий рассмотренному примеру, представлен на рисунке 9.

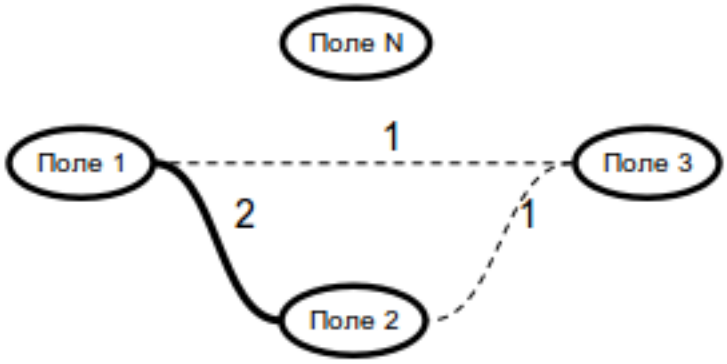

Рисунок 9 — Пример семантического графа

В рассмотренной графосемантической модели частотность полей $\nu(f, \Sigma)$ и сила связи между ними $\gamma(f_1, f_2, \Sigma)$ являются абсолютными значениями. Заменив их относительными, можно перейти к вероятностной модели [4; 55].

Заменим частотность поля f вероятностью его появления в отдельном контексте (публикации):

$$\widetilde{\nu}(f, \Sigma) = P(f). \tag{16}$$

Вероятность $P(f)$ вычисляется как отношение числа контекстов, в которых поле встретилось, к общему числу контекстов:

$$P(f) = \frac{|\{\sigma : \sigma \in \Sigma, \zeta(f, \sigma) > 0\}|}{|\Sigma|},$$

учитывая (12) и (16):

$$\widetilde{\nu}(f, \Sigma) = \frac{\nu(f, \Sigma)}{l}. \tag{17}$$

где $l = |\Sigma|$ – число контекстов (публикаций).

Элементы семантической карты $\gamma(f_1, f_2, \Sigma)$ (силу связи полей) заменим вероятностью совметной встречи этих полей в контексте (публикации):

$$\widetilde{\gamma}(f_1, f_2, \Sigma) = P(f_1, f_2).$$

Вероятность $P(f_1, f_2)$ может быть найдены через условную вероятность $P(f_1 \mid f_2)$ или $P(f_2 \mid f_1)$:

$$P(f_1 \mid f_2) = \frac{P(f_1, f_2)}{P(f_2)},$$

$$P(f_2 \mid f_1) = \frac{P(f_1, f_2)}{P(f_1)},$$

$$P(f_1, f_2) = P(f_1 \mid f_2)P(f_2) = P(f_2 \mid f_1)P(f_1). \tag{18}$$

Величина $P(f_2 \mid f_1)$ означает вероятность появления поля f_2 в контексте, в котором уже присутствует поле f_1. Она может быть вычислена как отношение мощности пересечения множеств контекстов, в которых присутствуют поля f_1 и f_2 к мощности множества контекстов, в которых присутствует поле f_1:

$$P(f_2 \mid f_1) = \frac{|\{\sigma : \sigma \in \Sigma, \zeta(f_1, \sigma) > 0\} \cap \{\sigma : \sigma \in \Sigma, \zeta(f_2, \sigma) > 0\}|}{|\{\sigma : \sigma \in \Sigma, \zeta(f_1, \sigma) > 0\}|}. \tag{19}$$

Используя равенства (4), (12), (18) и (19), определим значения элементов $\widetilde{\Gamma}(F,\Sigma)$:

$$\widetilde{\gamma}(f_1,f_2,\Sigma)=\frac{\gamma(f_1,f_2,\Sigma)}{\nu(f_1,\Sigma)}\cdot P(f_1),$$

подставив (16) и (17), получим:

$$\widetilde{\gamma}(f_1,f_2,\Sigma)=\frac{\gamma(f_1,f_2,\Sigma)}{l}. \tag{20}$$

По аналогии определим новую матрицу смежности:

$$\widetilde{A}=\{\widetilde{\mathrm{a}}_{ij}\}, i=\overline{1,n}, j=\overline{1,n}, \tag{21}$$

$$\widetilde{\mathrm{a}}_{ij}=\begin{cases}\widetilde{\gamma}(f_i,f_j,\Sigma), & i\neq j, \widetilde{\gamma}(f_i,f_j,\Sigma)>0,\\ 0, & i\neq j, \widetilde{\gamma}(f_i,f_j,\Sigma)=0,\\ 0, & i=j.\end{cases} \tag{22}$$

Очевидно, свойства матрицы $\widetilde{A}$ совпадают со свойствами матрицы A. На основе (16) и (21–22) определим вероятностный семантический граф:

$$\widetilde{G}_\Omega=(V,\widetilde{A},\widetilde{\nu}), V=\{f:\nu(f)>0, f\in F\}. \tag{23}$$

В вероятностном семантическом графе $\widetilde{G}_\Omega$ вес вершины является вероятностью существования связи поля, соответствующего вершине, с произвольным контекстом, а вес ребра является вероятностью одновременного существования связей полей, соединяемых данным ребром, с произвольным контекстом. Вероятностный семантический граф является частным случаем графической вероятностной модели [13; 26].

Рассмотрим задачу оценки вероятности связывания определённых полей с новым контекстах, добавляемым в графосемантическую модель Ω. Для этого введём понятие набора полей. Набор полей – совокупность элементов множества F, связанных с одним контекстом σ (будем говорить, что набор полей F_σ присутствует в контексте σ):

$$\mathrm{F}_\sigma=\{f:f\in F,\zeta(f,\sigma)>0\}, F_\sigma\subset F, \sigma\in\Sigma. \tag{24}$$

Поскольку функция $\min\{\zeta(f,\sigma),1\} \in \{0,1\}$ определяет факт связи поля f с контекстом σ, набору F_σ можно поставить в соответствие бинарный вектор vf длины n, в котором 1 означает существование связи f с σ, а 0 – её отсутствие:

$$vf_i = \min\{\zeta(f_i,\sigma),1\}, \sigma \in \Sigma, f_i \in F, i = \overline{1,n}.$$

Следовательно, количество возможных наборов F_σ можно вычислить следующим образом:

$$\overline{A}_2^n = 2^n. \tag{25}$$

Очевидно, векторы vf являются строками таблицы 7.

Оценим вероятность присутствия заданного набора полей в контексте:

$$\mathrm{P}(F_\sigma) = P(f_1)P(f_2 \mid f_1)P(f_3 \mid f_1,f_2)...P(f_k \mid f_1,f_2,...,f_{k-1}), \tag{26}$$

где $P(F_\sigma)$ – вероятность присутствия набора полей F_σ в контексте σ, $k = |F_\sigma|$ – количество полей в наборе.

Выражение (26) позволяет оценить вероятность присутствия конкретных полей в контексте σ, однако не учитывает присутствие или отсутствие других полей из F. Оценку вероятности присутствия в контексте σ набор полей F_σ при отсутствии других полей можно представить следующим образом:

$$P(F_\sigma, \overline{F \setminus F_\sigma}) = P(F_\sigma)\prod_{i=1}^{n-k} P(\overline{f_{k+i}} \mid f_1,f_2,...,f_k,\overline{f_{k+1}},...,\overline{f_{k+i-1}}), \tag{27}$$

где $\overline{F \setminus F_\sigma}$ означает отсутствие в контексте σ полей, не входящих в набор F_σ, а $\overline{f_i}, i = \overline{k+1,n}$ – отсутствие поля f_i в контексте σ.

Определим вероятность отсутствия поля в контексте и соответствующие совместные вероятности:

$$P(\overline{f}) = 1 - \widetilde{\nu}(f) = 1 - \frac{\nu(f)}{l}, f \in F,$$
$$P(\overline{f_i}, \overline{f_j}) = \frac{|\{\sigma : \sigma \in \Sigma, \zeta(f_i, \sigma) = 0\} \cap \{\sigma : \sigma \in \Sigma, \zeta(f_j, \sigma) = 0\}|}{l},$$
$$P(f_i, \overline{f_j}) = \frac{|\{\sigma : \sigma \in \Sigma, \zeta(f_i, \sigma) = 1\} \cap \{\sigma : \sigma \in \Sigma, \zeta(f_j, \sigma) = 0\}|}{l},$$
$$i = \overline{1, n}, j = \overline{1, n}, i \neq j.$$

Очевидно, для вычисления компонентов выражения (26) вида $P(f_i \mid f_1, f_2, ..., f_{i-1}), i \geqslant 3$ семантическая карта и, соответственно, матрица смежности должны иметь размерность i. Для больших значений i вычисления становятся слишком трудоёмкими, т.к. сложность алгоритма экспоненциальна относительно числа полей и линейна относительно числа контекстов. Для решения этой проблемы воспользуемся следующим методом – применим к $P(f_i \mid f_1, f_2, ..., f_{i-1}), i \geqslant 3$ теорему Байеса [58; 59]:

$$P(f_i \mid f_1, f_2, ..., f_{i-1}) = \frac{P(f_1, f_2, ..., f_{i-1} \mid f_i) P(f_i)}{P(f_1, f_2, ..., f_{i-1})}. \tag{28}$$

Как следует из (24), знаменатель в (28) так же является вероятностью присутствия набора полей, и его значение может быть вычислено с помощью (26). Таким образом, (26) – рекуррентная формула. Для вычисления числителя необходимо произвести факторизацию распределения $P(f_1, f_2, ..., f_{i-1} \mid f_i)$. Введём допущение об условной независимости полей $f_1, f_2, ..., f_{i-1}$ при условии f_i[99]:

$$P(f_1, f_2, ..., f_{i-1} \mid f_i) = \prod_{j=1}^{i-1} P(f_j \mid f_i). \tag{29}$$

Тогда можно записать (28) следующим образом:

$$P(f_i \mid f_1, f_2, ..., f_{i-1}) = \frac{P(f_i) \prod_{j=1}^{i-1} P(f_j \mid f_i)}{P(f_1, f_2, ..., f_{i-1})},$$

и подставить в (26):

$$\mathrm{P}(F_\sigma) = P(f_1)P(f_2 \mid f_1)\prod_{i=3}^{k}\frac{P(f_i)\prod_{j=1}^{i-1}P(f_j \mid f_i)}{P(f_1, f_2, ..., f_{i-1})}. \quad (30)$$

Произведя аналогичные преобразования с (27) и подставив (30), получим:

$$P(F_\sigma, \overline{F \setminus F_\sigma}) = \mathrm{P}(F_\sigma)\prod_{i=1}^{n-k}\frac{P(\overline{f_{k+i}})\prod_{j=1}^{k}P(f_j \mid \overline{f_{k+i}})\prod_{j=1}^{i-1}P(\overline{f_{k+j}} \mid \overline{f_{k+i}})}{P(f_1, f_2, ..., f_k, \overline{f_{k+1}}, ..., \overline{f_{k+i-1}})}. \quad (31)$$

Определим операцию суперпозиции графосемантических моделей Ω и Ω_1:

$$\begin{gathered}\widehat{\Omega} = \Omega + \Omega_1, \quad (32)\\ F_{\widehat{\Omega}} = F_\Omega \cup F_{\Omega_1},\\ \Sigma_{\widehat{\Omega}} = \Sigma_\Omega \cup \Sigma_{\Omega_1},\\ C_{\widehat{\Omega}} = C_\Omega \cup C_{\Omega_1},\\ T_{\widehat{\Omega}} = T_\Omega \cup T_{\Omega_1}\\ \Psi(\Sigma_{\widehat{\Omega}}, T_{\widehat{\Omega}}) = \Psi(\Sigma_\Omega, T_\Omega) \cup \Psi(\Sigma_{\Omega_1}, T_{\Omega_1}),\\ \Phi(C_{\widehat{\Omega}}, \Sigma_{\widehat{\Omega}}) = \Phi(C_\Omega, \Sigma_{\widehat{\Omega}}) \cup \Phi(C_{\Omega_1}, \Sigma_{\widehat{\Omega}}),\\ \Lambda(C_{\widehat{\Omega}}, F_{\widehat{\Omega}}) = \Lambda(C_\Omega, F_{\widehat{\Omega}}) \cup \Lambda(C_{\Omega_1}, F_{\widehat{\Omega}}).\end{gathered}$$

Добавление контекста σ_1 к существующей графосемантической модели Ω можно рассматривать как суперпозицию моделей Ω и модели Ω_1, содержащую один контекст σ_1. Рассмотрим графосемантическую модель Ω_1. Положим, что для σ_1 определён набор полей F_{σ_1}. Для такой графосемантической модели могут быть построены семантические графы G_{Ω_1} и $\widetilde{G}_{\Omega_1}$, элементы ко-

торых будут определены следующим образом:

$$a_{ij} = \begin{cases} 1, & f_i \in F_{\sigma_1}, f_j \in F_{\sigma_1}, \\ 0, & f_i \notin F_{\sigma_1}, \\ 0, & f_j \notin F_{\sigma_1}, \end{cases}$$

$$\nu(f) = \begin{cases} 1, & f \in F_{\sigma_1}, \forall f \in F, \\ 0, & f \notin F_{\sigma_1}, \forall f \in F, \end{cases}$$

$$i = \overline{1,n}, j = \overline{1,n},$$

кроме того:

$$a_{ij} = \widetilde{a}_{ij}, i = \overline{1,n}, j = \overline{1,n}, \tag{33}$$

$$\nu_i(f) = \widetilde{\nu}_i(f), i = \overline{1,n}, \forall f \in F. \tag{34}$$

Из (33–34) следует, что графы G_{Ω_1} и $\widetilde{G}_{\Omega_1}$ равны, т.е. $G_{\Omega_1} = \widetilde{G}_{\Omega_1}$, поэтому будем рассматривать только G_{Ω_1}. Очевидно, граф G_{Ω_1} зависит лишь от набора полей F_{σ_1}. Следовательно, любой набор F_σ определяет семантический граф G_{Ω_1}.

Очевидно, графосемантическую модель Ω можно рассматривать как динамическую систему, переход между состояниями которыми происходит при изменении любого из параметров $\Sigma, F, \Phi(C,\Sigma), \Lambda(C,F)$. Кроме того, любую графосемантическую модель Ω можно представить в виде последовательной суперпозиции моделей с одним контекстом, которую можно записать в виде рекуррентного выражения:

$$\Omega^{(i+1)} = \Omega^{(i)} + \Omega_1^{(i)}, i = \overline{1,l-1}.$$

При таком представлении операция суперпозиции (32) изменяет состояние графосемантической модели, а $\Omega^{(i)}, i = \overline{1,l-1}$ – есть последовательность состояний модели.

Рассмотрим графосемантическую модель с фиксированным множеством полей F и ограниченным числом контекстов l как случайный процесс. Очевидно, множество состояний для такой системы конечно. Кроме того, возмож-

но определение вероятности перехода между состояниями с помощью (31). Таким образом, процесс изменения графосемантической модели Ω можно описать как марковский процесс [77]. Предположим, что в общем случае состояние $\Omega^{(i+1)}$ зависит только от $\Omega^{(i)}$, тогда процесс представим в виде дискретного марковского процесса первого порядка:

$$P(X_{n+1} = \Omega^{(n+1)} \mid X_0 = \Omega^{(0)}, ..., X_{\mathrm{n}} = \Omega^{(n)}) = P(X_{n+1} = \Omega^{(n+1)} \mid X_n = \Omega^{(n)}). \tag{35}$$

где $\{X_n\}$ – пространство состояний системы Ω.

Пусть $N = |\{X_n\}|$ – число возможных состояний рассматриваемого процесса, α – распределение вероятностей начального состояния, $P = \{p_{ij}\}$ – матрица перехода. Очевидно, из каждого состояния, кроме последнего, система может с ненулевой вероятностью перейти максимум в одно из 2^n состояний. Определим элементы P:

$$p_{ij} = \begin{cases} P(F_{\sigma_1}, \overline{F \setminus F_{\sigma_1}}), & \exists F_{\sigma_1} : \Omega^{(j)} = \Omega^{(i)} + \Omega_1, \\ 0, & !\exists F_{\sigma_1} : \Omega^{(j)} = \Omega^{(i)} + \Omega_1, \end{cases} \tag{36}$$
$$i = \overline{1, N}, j = \overline{1, N}, F_{\sigma_1} \in F.$$

Заметим, что в результате допущения об условной независимости присутствия полей в контексте при заданном условии (29), матрица P с элементами (36) в общем случае не является стохастической [77]:

$$\sum_{j=1}^{N} p_{ij} \leqslant 1, i = \overline{1, N}.$$

Кроме того, при отсутствии в текущем состоянии контекстов с некоторым набором полей F_σ, переходам через данный набор полей будет соответствовать нулевая вероятность. Для решения данной проблемы можно распределить «потерянную» вероятность $1 - \sum_{j=1}^{N} p_{ij}$ между наборами F_σ, не представленными в текущем состоянии системы.

Рассмотренное представление графосемантической модели в виде дискретного марковского процесса можно использовать для решения ряда задач, например для оценки вероятности перехода модели из заданного исходного

Таблица 9 — Функция $Z(\mathrm{F}_{\Omega_{ABC}}, \Sigma_{\Omega_{ABC}})$

$Z(\mathrm{F}_{\Omega_{ABC}}, \Sigma_{\Omega_{ABC}})$	Контекст 1	Контекст 2	Контекст 3
A	1	1	0
B	1	1	0
C	1	0	1

состояние в конечное за k шагов:

$$P(X_r = \Omega^{(i)} \mid X_{r+k} = \Omega^{(j)}) = (P^n)_{ij}. \tag{37}$$

Поставим задачу нахождения наиболее вероятного набора полей F_σ^* в новом контексте:

$$F_\sigma^* = \arg\max_{F_\sigma \subset F} P(F_\sigma).$$

Очевидно, возможность применения аппарата теории дискретных марковских процессов для анализа графосемантических моделей как динамических систем существенно расширяет перспективы применения этого метода. Большинство объектов, к которым применим данный метод, являются динамическими системами, например: предметные области, электронные СМИ, социальные медиа и т.д. Исследование моделей этих объектов как случайных процессов может выявить новые свойства, а также определить ключевые особенности их динамики. Так же следует отметить возможность анализа графосемантических моделей изменяющихся предметных областей.

Рассмотрим описанный подход на примере. В качестве исходной модели используем Ω_{ABC}, семантический граф которой изображён на рисунке 10. Множество контекстов данной модели состоит из трёх контекстов: $\Sigma_{\Omega_{ABC}} = \{\sigma_1, \sigma_2, \sigma_3\}$. Множество полей так же состоит из трёх полей: $F_{\Omega_{ABC}} = \{A, B, C\}$. В данном примере не будем рассматривать множества контекстов и метаполей. Функция $Z(\mathrm{F}_{\Omega_{ABC}}, \Sigma_{\Omega_{ABC}})$ (множество связей контекстов с полями) представлена в таблице 9.

Как следует из (25), для данного множества полей $F_{\Omega_{ABC}}$ возможно $2^3 = 8$ различных наборов полей, т.е. модель Ω_{ABC} может перейти в восемь различных состояний за один шаг. Оценим вероятности присутствия каждого из восьми наборов полей в новом контексте используя (31). Данные вероят-

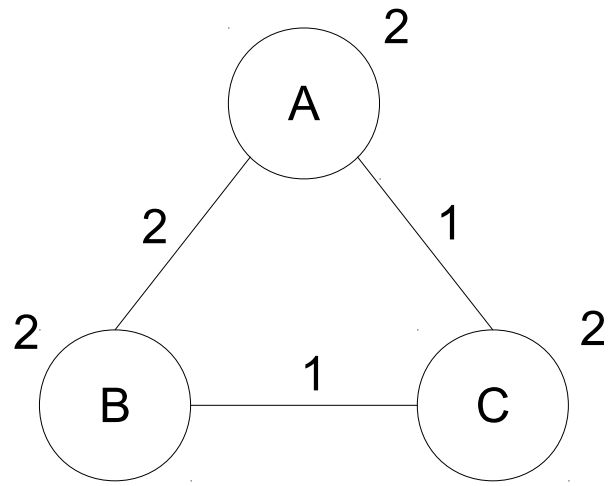

Рисунок 10 — Семантический граф $G_{\Omega_{ABC}}$

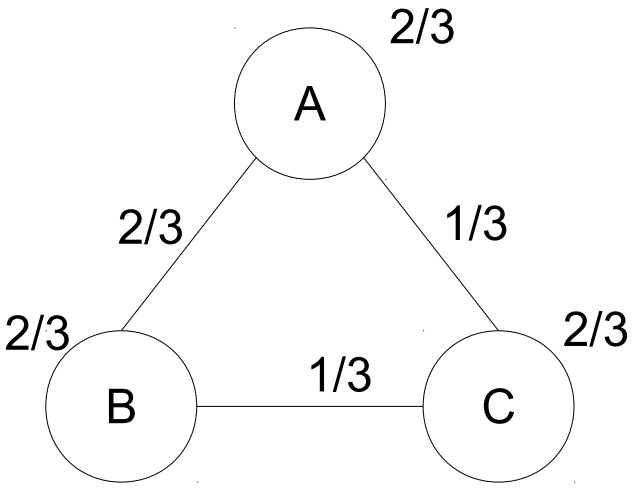

Рисунок 11 — Вероятностный семантический граф $\widetilde{G}_{\Omega_{ABC}}$

ности оцениваются на основе вероятностного семантического графа $\widetilde{G}_{\Omega_{ABC}}$, изображённого на рисунке 11.

В качестве примера выведем выражение для вычисления вероятности присутствия в новом контексте набора полей $\{AB\}$:

$$
\begin{gathered}
P(\{AB\}, \overline{\{C\}}) = P(A)P(B \mid A)P(\overline{C}|A, B), \\
P(\overline{C}|A, B) = \frac{P(A, B \mid \overline{C})P(\overline{C})}{P(A, B)}, \\
P(A, B \mid \overline{C}) = P(A \mid \overline{C})P(B \mid \overline{C}), \\
P(\{AB\}, \overline{\{C\}}) = P(A)P(B \mid A)\frac{P(A \mid \overline{C})P(B \mid \overline{C})P(\overline{C})}{P(A, B)}.
\end{gathered} \tag{38}
$$

Таблица 10 — Оценки вероятности присутствия различных наборов полей

F_σ	$P(F_\sigma, \overline{F \setminus F_\sigma})$	F_σ	$P(F_\sigma, \overline{F \setminus F_\sigma})$
$\emptyset$	0	$\{AB\}$	$\frac{1}{3}$
$\{A\}$	0	$\{AC\}$	0
$\{B\}$	0	$\{BC\}$	0
$\{C\}$	$\frac{1}{3}$	$\{ABC\}$	$\frac{1}{6}$

Таблица 11 — Перераспределённые оценки вероятности присутствия различных наборов полей

F_σ	$P(F_\sigma, \overline{F \setminus F_\sigma})$	F_σ	$P(F_\sigma, \overline{F \setminus F_\sigma})$
$\emptyset$	$\frac{1}{30}$	$\{AB\}$	$\frac{1}{3}$
$\{A\}$	$\frac{1}{30}$	$\{AC\}$	$\frac{1}{30}$
$\{B\}$	$\frac{1}{30}$	$\{BC\}$	$\frac{1}{30}$
$\{C\}$	$\frac{1}{3}$	$\{ABC\}$	$\frac{1}{6}$

В данном примере значение вероятности присутствия в новом контексте набора полей $\{AB\}$, вычисленное согласно (38), равно $\frac{1}{3}$, что не противоречит логическому заключению, поскольку набор полей $\{AB\}$ встречается в модели Ω_{ABC} в одном контексте из трёх возможных. В таблице 10 приведены значения вероятностей для каждого из восьми возможных наборов полей.

Как видно из таблицы 10, $\sum P(F_\sigma) = \frac{5}{6} < 1$, т.е. $\frac{1}{6}$ «потеряна». Очевидно, проблема заключается в ложности допущения (29) для набора полей $\{ABC\}$. Воспользуемся вышеописанным методом решения данной проблемы: распределим недостающую вероятность между наборами полей с нулевой вероятностью. В данном примере не будем использовать алгоритм дисконтирования, результат приведён в таблице 11.

Рассмотрим метод графосемантического моделирования применительно к задаче анализа научной деятельности. В качестве моделируемого объекта будем рассматривать агентов научного производства, т.е. любых участников научного производства, характеризуемых публикационной активностью. К агентам научного производства можно отнести отдельных учёных, научные коллективы, организации, журналы, целые научные направления и отрасли, всю науку.

Методику графосемантического моделирования предметных областей агентов научного производства можно разделить на несколько этапов. На пер-

вом этапе производится сбор исходных данных, описывающих результаты научной деятельности. Исходными данными в данной задаче являются объекты публикационной активности (точнее, их библиографические данные): статьи, тезисы, монографии и т.д. Такой выбор исходных данных обусловлен их высокой доступностью, применимостью к различным агентам научного производства и высокой доступностью, по сравнению с данными анкетирования или экспертными оценками.

В общем случае, описание объекта публикационной активности может содержать разное количество полей с произвольным количеством значений в каждом полем (например: список авторов, ключевые слова, список литературы), т.е. представляет собой квазиструктурированный документ. Ниже перечислены наиболее часто встречаемые поля в описаниях объектов публикационной активности:

1. Название – название работы.
2. Аннотация – ключевой объект предметной области, содержит авторское описание публикации, включающее все прочие моделируемые объекты.
3. Автор – информация об авторе публикации, может включать электронную почту автора, место работы, должность, учёную степень. Обязательно включает фамилию и инициалы. Так же, в случае использования гипертекстовой разметки, может содержать ссылку на документ, уникальным образом идентифицирующий автора, например на профиль в системе индекса цитирования. Поле «Автор» может включать как информацию об одном авторе, так и о нескольких.
4. Тип – тип публикации: «статья в журнале», «статья в сборнике», «монография» и т.д.
5. Язык – язык публикации: русский, английский и т.д.
6. Номер – номер периодического издания, в котором опубликована работа.
7. Год – год опубликования работы.
8. Страницы – номера страниц издания, содержащих публикацию.

9. Журнал – включает название издания, в котором опубликована работа, название издательства и прочие данные.

10. Рубрика – рубрика, в которой опубликована работа.

11. Ключевые слова – наиболее важное поле документа, содержащее список ключевых слов, посредством которых автор публикации описывает предметную область своей работы.

12. Список литературы – библиографический список литературных источников, использованных автором в своей работе.

Схема предметной области представлена ниже:

Для данной предметной области была построена графосемантическая модель Ω_S:

На основе вышеописанных исходных данных определим множество контекстов графосемантической модели Σ_Ω, множество мета-типов T и множество мета-полей $\Psi(\Sigma_\Omega, T)$. Пусть контексту $\sigma \in \Sigma$ соответствует объект публикационной активности (аннотация), а множество мета-типов представлено в таблице 12.

Таблица 12 — Множество мета-типов

Название типа	Тип	Входит в	Множ.
Название	Строка	–	Нет
Авторы	Структура	–	Да
ФИО	Строка	Авторы	Нет
Учёная степень	Строка	Авторы	Нет
Организация	Строка	Авторы	Нет
Город	Авторы	Авторы	Нет
Журнал	Структура	–	Нет
Название	Строка	Журнал	Нет
Издательство	Строка	Журнал	Нет
Номер	Число	Журнал	Нет
Год	Число	–	Нет
Рубрика	Строка	–	Нет
Ключевые слова	Строка	–	Да

На втором этапе строится графосемантическая модель предметной области, описываемой исходными данными. Для этого каждому объекту публи-

кационной активности ставится в соответствие контекст и заполняются его метаданные (название публикации, год опубликования, автор, организация, журнал и т.д.). Метаданные имеют большое значение, т.к. они позволяют детализировать графосемантическую модель и строить срезы. Например, с помощью метаданных можно выделить предметную область отдельного автора или журнала. Далее, в каждом объекте публикационной активности выделяются ключевые слова и добавляются в соответствующие контексты в качестве семантических компонентов. Результатом данного этапа является частично сформированная графосемантическая модель, содержащая отдельные контексты, не связанные между собой. Такая модель может быть использована для решения некоторых задач, например для оценки близости публикаций на основе представленных в них ключевых слов. Однако, из-за большого разнообразия ключевых слов, результативность данного подхода незначительна. Кроме того, при наличии перекрёстных ссылок в метаданных (извлечённых из списков цитирований), данная модель может использоваться для построения карт совместных цитирований публикаций.

Для выполнения третьего этапа требуется привлечение экспертов моделируемой научной отрасли. В задачи экспертов входит определении списка семантических полей, описывающих научную отрасль (например: «микроэкономика»), и установление связей этих полей с доступными семантическими компонентами (извлечёнными на втором этапе ключевыми словами). Отметим, что связи между полями и компонентами относятся к типу «многие ко многим». Количество семантических полей зависит от выбранного уровня детализации модели. Модель может включать несколько уровней детализации (иерархическая графосемантическая модель), при этом поля нижестоящих уровней включаются в вышестоящие поля, образуя иерархию. Такой подход позволяет точнее описывать моделируемые предметные области и производить анализ модели на любом из доступных уровней детализации.

Полученная на третьем этапе графосемантическая модель является завершенной и может быть использована для построения семантических карт, графов и полевого анализа. Кроме того, модель может дополняться новыми контекстами и допускает изменение семантических полей и их связей с компонентами.

Очевидно, в каждом контексте σ присутствует набор семантических полей F_σ, определённых экспертом и связанных с этим контекстом посредством ключевых слов (семантических компонентов). Фактически, набор полей F_σ определяет единичную предметную область $F_i = F_\sigma, \forall i = \overline{1, 2^n}$ (или, в задаче моделирования предметных областей агентов научного производства, единичную научную предметную область). Следует заметить, что в детализированных моделях с большим числом семантических полей и контекстов число различных единичных предметных областей (2^n) может быть очень велико. Тем не менее, несколько контекстов в одном проекте могут содержать одинаковые наборы полей F_σ, особенно если объекты моделируемой предметной области, на основе которых сформировано множество контекстов Σ, связаны каким-либо образом. Примером такой связи может быть принадлежность публикаций одному автору или журналу, т.е. тематическая связь.

На практике неудобно работать с единичными предметными областями как с множествами F_i или бинарными векторами vf, однако все возможные единичные предметные области составляют конечное множество и, следовательно, могут быть перенумерованы. Заметим, что бинарный вектор vf представим как запись целого числа в двоичной системе счисления и присвоим соответствующую запись этого числа в десятичной системе счисления в качестве номера единичной предметной области. Очевидно, обратное преобразование может быть выполнено как перевод номера единичной предметной области в двоичную систему счисления и представление в виде вектора vf. Ниже представлен пример нумерации единичной предметной области:

$$F = \{A, B, C\},$$
$$F_i = \{A, C\},$$
$$vf = (1; 0; 1),$$
$$N_{F_i} = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = 5.$$

Отдельные единичные предметные области могут быть проинтерпретированы экспертом как по семантическим полям, входящим в наборы полей F_σ, так и по контекстам σ, содержащим эти наборы. Однако, выделение значимых единичных предметных областей из общего множества может быть затруд-

нительно, кроме того, не учитываются близкие по содержанию предметные области. Поэтому необходима группировка единичных предметных областей. Одним из очевидных решений данной задачи является кластерный анализ единичную предметных областей. Кластерный анализ производился над множеством контекстов, параметрами были бинарные вектора vf, содержащие наборы полей соответствующих контекстов. Поскольку каждому контексту соответствует единичная предметная область, можно использовать результат в качестве оценки подобия публикаций и соответствующих им единичных предметных областей.

Существует множество алгоритмов кластерного анализа: . Наиболее простым и часто используемым является алгоритм K-means.

Альтернативой описанному алгоритму является алгоритм C-means. Его можно рассматривать как модифицированную версию K-means. Важным отличием K-means от C-means является то, что последний возвращает нечёткие значения, т.е. степень принадлежности элементов к каждому кластеру. Алгоритм нечёткой кластеризации C-means основывается на минимизации целевой функции $J(x, c, u, m)$:

$$J(x, c, u, m) = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2, m \in R, m > 1, \tag{39}$$

где m – нечёткий параметр (выбирается экспертом), u – степень принадлежности кластеру, $\|*\|$ – норма , характеризующая близость элементов анализируемого пространства. Процесс оптимизации целевой функции заключается в итеративном пересчёте степеней $u_{ij}, i = \overline{1, N}$ и центров кластеров $c_j, j = \overline{1, C}$:

$$u_{ij} = \frac{1}{\sum\limits_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}, \tag{40}$$

$$c_j = \frac{\sum\limits_{i=1}^{N} u_{ij}^{m} x_i}{\sum\limits_{i=1}^{N} u_{ij}^{m}}. \tag{41}$$

Условие окончания итерации: $\max\limits_{ij}\left(u_{ij}^{(k+1)}-u_{ij}^{(k)}\right)<\varepsilon$, где ε – заданный критерий, $\varepsilon>0$.

2.2 Разработка математической модели исследовательской траектории

По мере публикации новых работ, агент научного производства осуществляет переход из одной предметной области в другую, тем самым формируя исследовательскую траекторию.

Математическая модель исследовательской траектории представляет собой множество состояний предметной области агента научного производства, описываемых графосемантической моделью $\Omega_t, t=\overline{1,T}$, где T – количество состояний в исследовательской траектории. Таким образом, модель исследовательской траектории может быть представлена следующим образом:

$$Y=\left\{y(t):y(t)=\Omega_t,t=\overline{1,T}\right\}.$$

Кроме того, в упрощённом виде исследовательскую траекторию можно представить в виде последовательности семантических карт, соответствующих графосемантическим моделям предметных областей каждого состояния:

$$Y=\left\{y(t):y(t)=A(t),i=\overline{1,n},j=\overline{1,n},t=\overline{1,T}\right\},$$

где $A(t)$ – матрица смежности семантического графа модели Ω_t.

Поскольку матрица A симметрична и имеет нулевую главную диагональ (свойства 1 и 2), достаточно использовать верхнетреугольную часть матрицы A:

$$Y=\left\{y(t):y(t)\in R^{p_A},t=\overline{1,T}\right\}, \tag{42}$$

$$y_k(t)=a_{ij}(t),i=\overline{1,n},j=\overline{1,n},j>i, \tag{43}$$

$$k=n+(i-1)\cdot n+j-i=i\cdot(n-1)+j, \tag{44}$$

где p_A число параметров (43) (количество уникальных значений матрицы A):

$$p_A = \sum_{k=1}^{n-1} k = \frac{n(n-1)}{2}. \tag{45}$$

Далее в качестве модели исследовательской траектории будет использоваться (42-44).

Публикуя новые работы, агент научного производства переходит к новому состоянию своей исследовательской траектории. При этом изменение модели предметной области определяется единичной предметной областью объекта публикационной активности. Под очередным научным исследованием понимается не завершённый научный проект, а лишь его часть, завершающаяся публикацией, выступлением на конференции и т.п.

Деятельность агентов научного производства является длительным, трудоёмким и невоспроизводимым процессом, что делает практически невозможным проведение натурных экспериментов с реальными агентами научного производства. Поэтому, для исследования процесса изменения состояния предметной области было применено имитационное моделирование. В качестве имитационной модели используется описанный выше дискретный марковский процесс первого порядка, т.е. марковская цепь. Ниже приведено описание процесса имитации.

Пусть задана графосемантическая модель научной предметной области агента научного производства Ω и модель изменения её состояния в виде марковского процесса первого порядка (35) с матрицей перехода $p_{N\times N}$ (36). Тогда, имитационное моделирование первых T переходов может быть осуществлено следующим образом:

1. В качестве начального состояние X_0 выбирается заданная модель Ω;
2. Для каждого шага $t = \overline{1,T}$:
   (a) выбирается отрезок $x : \left[0; \sum_{i=1}^{N} p_{ti}\right]$;
   (b) генерируется («разыгрывается») случайное число r, равномерно распределённое на отрезке x;

(c) в качестве состояния X_t выбирается такое состояние $\Omega^{(j)}$, для которого выполняется условие:

$$\sum_{i=1}^{j-1} p_{ti} < r \leqslant \sum_{i=1}^{j} p_{ti}.$$

(d) переход к шагу $t+1$.

3. Полученная последовательность состояний $X_t, t = \overline{0,T}$ принимается в качестве результата имитационного моделирования.

Принцип, используемый на шагах (b),(c) так же называется «принципом рулетки» [85].

Полученная последовательность $X_t, t = \overline{0,T}$ определяет прогнозируемую исследовательскую траекторию агента научного производства.

2.3 Постановка и решение задачи оптимизации исследовательской траектории

Задача поиска оптимальной исследовательской траектории агента научного производства была сформулирована в форме задачи оптимального управления. Для этого необходимо представить научную предметную область как динамическую систему. Определим основные элементы динамической системы:

- t – временной параметр;
- t_0 – начальное значение времени (время исходного состояния научной предметной области);
- T – конечное значение времени (время, к которому должно быть достигнуто целевое состояние научной предметной области);
- p – число анализируемых параметров научной предметной области;
- $y_i(t), i = \overline{1,p}, \forall t \in [t_0, T]$ – параметры научной предметной области (фазовые переменные);

- $y(t) = (y_1(t), \ldots, y_{\text{p}}(t)), \forall t \in [t_0, T]$ – исследовательская траектория предметной области агента научного производства;
- $Y \subset \mathbb{R}^p$ – фазовое пространство, $y(t) \in Y, \forall t \in [t_0, T]$;
- $y(t_0) \in Y$ – начальное состояние предметной области агента научного производства;
- $y(T) \in Y$ – конечное состояние предметной области агента научного производства;
- r – число возможных управляющих воздействий;
- $u_i(t), i = \overline{1, r}, \forall t \in [t_0, T]$ – управляющие воздействия;
- $u(t) = (u_1(t), \ldots, u_r(t)), \forall t \in [t_0, T]$ – управление;
- $U \subset \mathbb{R}^r$ – область управления, $u(t) \in U, \forall t \in [t_0, T]$.

Допустим, что $u_i(t), i = \overline{1, r}$ – кусочно–непрерывные функции $\forall t \in [t_0, T]$, тогда научная предметная область как динамическая система может быть представлен в виде системы дифференциальных уравнений:

$$\begin{cases} \frac{dy_1}{dt} = g_1(t, y(t), u(t)), \\ \frac{dy_2}{dt} = g_2(t, y(t), u(t)), \\ \ldots \\ \frac{dy_p}{dt} = g_p(t, y(t), u(t)). \end{cases} \quad \forall t \in [t_0, T]. \tag{46}$$

Система (46) может быть представлена в векторной форме:

$$\frac{dy}{dt} = g(t, y(t), u(t)), \forall t \in [t_0, T], \tag{47}$$

где

$$\frac{dy}{dt} = \begin{pmatrix} \frac{dy_1}{dt} \\ \frac{dy_2}{dt} \\ \vdots \\ \frac{dy_p}{dt} \end{pmatrix}, \; g(t, y(t), u(t)) = \begin{pmatrix} g_1(t, y(t), u(t)) \\ g_2(t, y(t), u(t)) \\ \vdots \\ g_p(t, y(t), u(t)) \end{pmatrix}.$$

Так же для постановки задачи оптимального управления необходимо задать критерий качества управления, который в интегральном виде записывается следующим образом:

$$J(u) = \int_{t_0}^{T} F(t, y(t), u(t))dt, \tag{48}$$

где $F(t, y(t), u(t))$ – функция оценки качества процесса в момент времени t.

На основе вышеописанного, поставим задачу оптимального управления научной предметной областью (47) с критерием качества (48):

$$\frac{dy}{dt} = g(t, y(t), u(t)), \tag{49}$$

$$J(u) = \int_{t_0}^{T} F(t, y(t), u(t))dt \to \max, \tag{50}$$

$$y(t) \in Y, u(t) \in U, t \in [t_0, T]. \tag{51}$$

Заметим, что критерий качества (50) так же может быть представлен в следующем виде:

$$J(u) = \int_{t_0}^{T} F(t, y(t), u(t))dt \to \max.$$

Задача оптимального управления научной предметной областью (49–51) записана в общем виде, для практического применения необходимо определить значения входящих в неё параметров. В данном случае, значения параметров должны определяться на основе графосемантической модели.

Поскольку исходными данными для моделирования являются статьи (работы), состояние модели предметной области агента научного производства может изменяться только с выходом в печать новых работ. Следовательно, временной параметр t может принимать лишь значения, соответствующие моментам выхода в печать отдельных работ, т.е. принадлежит дискретному мно-

жеству. Допустим, агенту научного производства необходимо достичь некоторой цели, опубликовав не более T работ. Тогда публикации можно перенумеровать числами от 0 до $T-1$, и параметр t будет принимать эти номера в качестве значений.

Таким образом, рассматриваемая модель научной предметной области является дискретным объектом управления. Для дискретного объекта управления ставится дискретная задача оптимального управления. Для этого необходимо определить множество значений t

$$t = \overline{0, T-1}, \tag{52}$$

дискретную траекторию системы

$$y(t+1) = g(t, y(t), u(t)), \tag{53}$$

и критерий качества

$$J(u) = \sum_{t=t_0}^{T} F(t, y(t), u(t)). \tag{54}$$

Учитывая (52–54), дискретная задача оптимального управления научной предметной областью ставится следующим образом:

$$y(t+1) = g(t, y(t), u(t)), \tag{55}$$

$$J(u) = \sum_{t=0}^{T-1} F(t, y(t), u(t)) \to \max, \tag{56}$$

$$y(t) \in Y, u(t) \in U, t = \overline{0, T-1}. \tag{57}$$

За основу фазового пространства возьмём элементы вероятностного семантического графа (15) $\widetilde{A}$ и $\widetilde{\nu}(f, \Sigma), \forall f \in F$. Так как предметная область агента научного производства будет изменяться с каждым состоянием систе-

мы, граф $\widetilde{G}$ и его элементы будут зависеть от параметра t

$$\widetilde{G} = \widetilde{G}(t),$$
$$\widetilde{a}_{ij} = \widetilde{a}_{ij}(t),$$
$$\widetilde{A} = \widetilde{A}(t),$$
$$\widetilde{\nu} = \widetilde{\nu}(t).$$

Определим параметры для $\widetilde{A}$ и $\widetilde{\nu}$ отдельно. Пусть первые n параметров – вероятности появления полей предметной области в работе, тогда:

$$y_k(t) = \widetilde{\nu}(f_k, \Sigma, t), k = \overline{1, n}, f_k \in F. \tag{58}$$

Поскольку матрица $\widetilde{A}$ симметрична и имеет нулевую главную диагональ (свойства 1 и 2), достаточно использовать верхнетреугольную часть матрицы $\widetilde{A}$:

$$y_k(t) = a_{ij}(t), i = \overline{1, n}, j = \overline{1, n}, j > i, \tag{59}$$
$$k = n + (i-1) \cdot n + j - i = i \cdot (n-1) + j.$$

Число параметров (59) можно вычислить следующим образом:

$$p_{\widetilde{A}} = \sum_{k=1}^{n-1} k = \frac{n(n-1)}{2}. \tag{60}$$

Определим фазовое пространство Y, учитывая (10), (14) и (60):

$$Y \in \mathbb{R}^{|p|},$$

где

$$p = n + p_{\widetilde{A}}.$$

Поскольку разрабатываемая система оперирует предметной областью агента научного производства (объектом управления) на уровне модели предметной области, вид доступных управляющих воздействий определяется данной моделью. В рассмотренной графосемантической модели такими воздействиями, изменяющими её состояние, являются публикации (контексты). Фак-

тически, публикуя новые работы, агент научного производства изменяет свою предметную область. Таким образом, очередное управляющее воздействие состоит в задании тематики очередного научного исследования.

Определим тематику научного исследования как множество полей, участвующих в этом исследовании. Тогда управление можно записать следующим образом:

$$u_i(t) = \begin{cases} 1, & f_i \in D_t \\ 0, & f_i \notin D_t \end{cases}, i = \overline{1,n}, f_i \in F \tag{61}$$

где D_t – направление очередного научного исследования, f_i – присутствие поля f_i в результирующей публикации.

Таким образом, $u(t)$ является бинарным вектором, определяющим направление исследования. Заметим так же, что $r = n$.

Определим закон изменения состояния предметной области агента научного производства на основе (55) и (61). Для нахождения закона изменения первых n параметров, необходимо выяснить, как изменяется вероятность встречи соответствующих полей, входящих в очередное направление исследования $u(t)$:

$$\widetilde{\nu}^*(f_{\text{k}}, \Sigma) = \frac{\nu^*(f_k, \Sigma)}{l}, k = \overline{1,n},$$

где $\widetilde{\nu}^*(f_k, \Sigma)$ – значение вероятности встречи поля после публикации новой работы, $\nu^*(f_k, \Sigma)$ – абсолютное значение частотности поля после публикации новой работы.

Очевидно, в силу (61), абсолютное значение $\nu(f_k, \Sigma)$ увеличится на $u_k(t)$, таким образом:

$$\nu^*(f_k, \Sigma) = \nu(f_k, \Sigma) + u_k(t), k = \overline{1,n},$$

$$\widetilde{\nu}^*(f_k, \Sigma) = \frac{\nu(f_k, \Sigma) + u_k(t)}{l}, k = \overline{1,n}. \tag{62}$$

Из (58) и (62) следует закон изменения первых n фазовых параметров предметной области агента научного производства y_k:

$$g_k(t, y(t), u(t)) = y_k(t) + \frac{u_k(t)}{l}, k = \overline{1, n}. \tag{63}$$

Чтобы определить закон изменения остальных фазовых переменных предметной области агента научного производства, рассмотрим как изменяется $\widetilde{\Gamma}(F, \Sigma)$ при появлении новой работы по направлению исследования $u(t)$:

$$\widetilde{\gamma}^*(f_i, f_j, \Sigma) = \frac{\gamma^*(f_i, f_j, \Sigma)}{l}, i = \overline{1, n}, j = \overline{1, n},$$

где $\widetilde{\gamma}^*(f_i, f_j, \Sigma)$ – значение вероятности совместной встречи полей f_i и f_j после публикации новой работы, $\gamma^*(f_i, f_j, \Sigma)$ – число совместных появлений полей f_i и f_j после публикации новой работы.

Поскольку значение $\gamma(f_i, f_j, \Sigma)$ увеличится на единицу только в случае присутствия в новой работе как поля f_i, так и поля f_j, а также учитывая (61), можно записать:

$$\begin{gathered}
\gamma^*(f_i, f_j, \Sigma) = \gamma(f_i, f_j, \Sigma) + u_i(t) \cdot u_j(t), \\
\widetilde{\gamma}^*(f_i, f_j, \Sigma) = \frac{\gamma(f_i, f_j, \Sigma) + u_i(t) \cdot u_j(t)}{l}, \\
i = \overline{1, n}, j = \overline{1, n}, i \neq j.
\end{gathered} \tag{64}$$

На основе (59) и (64) определим закон изменения фазовых переменных предметной области агента научного производства y_k при $k = \overline{n+1, p}$:

$$g_k(t, y(t), u(t)) = y_k(t) + \frac{u_i(t) \cdot u_j(t)}{l}, i = \overline{1, n}, j = \overline{1, n}, j > i, \tag{65}$$

$$k = i \cdot (n-1) + j.$$

Запишем закон изменения фазовых переменных предметной области агента научного производства (63), (65) в матричном виде:

$$g(t, y(t), u(t)) = Ay(t) + u^T(t)Bu(t). \tag{66}$$

Определим матрицы A и B:

$$A = \{a_{ij}\}, i = \overline{1,n}, j = \overline{1,n},$$

$$a_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} i = \overline{1,n}, j = \overline{1,n},$$

$$B = \{b_{ij}\}, i = \overline{1,n}, j = \overline{1,n},$$

$$b_{ij} = \begin{cases} 1, & k = i \cdot (n-1) + j, \\ 0, & k \neq i \cdot (n-1) + j, \end{cases} i = \overline{1,n}, j = \overline{1,n}.$$

Теперь (55) можно записать с использованием (66) следующим образом

$$y(t+1) = Ay(t) + u^T(t)Bu(t). \tag{67}$$

Из (67) следует, что рассматриваемая система является стационарной и автономной.

Очевидно, ограничения на переменные управления могут быть обусловлены постановкой конкретной задачи, которая, в свою очередь, может зависеть от типа агента научного производства. Так, например, для научного журнала накладывается существенное ограничение на управляющие воздействия в виде доступных для публикации работ у редакции. Однако, могут быть сформулированы некоторые общие ограничения, характерные для всего класса решаемых задач.

В первую очередь, к таким ограничениям относится ограничение на переход в другую частнонаучную предметную область. Такой переход подразумевает значительную смену направления деятельности агента научного производства и, следовательно, не входит в класс решаемых задач. Однако, сам переход может быть определён нечётко, за счёт чего достигается некоторая гибкость в возможностях применения предлагаемой методики. Для формального определения первого ограничения введём функцию $r_j(u)$ – степень принадлежности единичной предметной области, определяемой управляющим воз-

действием u, j-му ведущему научному направлению:

$$r_j(u) = \frac{1}{\sum\limits_{k=1}^{C} \left(\frac{\|u-c_j\|}{\|u-c_k\|} \right)^{\frac{2}{m-1}}}, \tag{68}$$

где m – параметр нечёткой кластеризации из (39-41), c_j – центр j-го кластера.

С учётом (68) первое ограничение может быть записано следующим образом:

$$r_k(u) \geqslant r_j(u),$$
$$\forall j = \overline{1, C},$$
$$j \neq k,$$

где C – количество кластеров (ведущих научных направлений) из (39-41), k – текущее направление.

В данной задаче ограничения необходимо представить в виде аддитивных штрафов к качеству управления:

$$R_1(t, y(t), u(t)) = 1 - r_k(u(t)), \tag{69}$$

под текущим кластером понимается частнонаучная предметная область, соответствующая последнему состоянию (на данном шаге) исследовательской траектории.

Вторым ограничением является гладкость траектории – отсутствие резких смен направления научной деятельности. При этом может учитываться вся предыдущая траектория агента научного производства, её часть (например, за последние N лет), либо только последний шаг. Поскольку управляющие воздействия представлены бинарными векторами, в качестве математической формализации данного ограничения предлагается использовать расстояние Хэмминга. Расстояние Хэмминга между двумя бинарными векторами x и y определяется следующим образом:

$$d_{xy} = \sum_{k=1}^{p} |x_k - y_k| ,$$

где p – размерность векторов.

Запись второго ограничения в виде аддитивного вектора для всей предшествующей исследовательской траектории агента научного производства принимает следующий вид:

$$R_2(t, y(t), u(t)) = \frac{1}{n} \min_{z \in Z} \sum_{k=1}^{p} |x_k - z_k| , \tag{70}$$

где Z – предшествующая исследовательская траектория агента научного производства, p – размерность векторов (равна числу семантических полей), $\frac{1}{n}$ – коэффициент нормализации.

Третьим является ограничение на количество семантических полей в выбираемой предметной области. Формально, количество единиц в бинарном векторе управляющего воздействия:

$$R_3(t, y(t), u(t)) = \begin{cases} 1, & \sum_{k=1}^{p} u(t)_k > P \\ 0, & \sum_{k=1}^{p} u(t)_k \leqslant P \end{cases} \tag{71}$$

где P – заданное ограничение на количество полей в выбираемой предметной области.

Четвёртое ограничение – на отклонение от прогнозируемой исследовательской траектории. Данное ограничение позволяет задать оптимизируемой траектории степень близости к прогнозируемой траектории агента научного производства. Математически его можно представить как норму разности между семантическими картами соответствующих предметных областей на каждом шаге:

$$R_4(t, y(t), u(t)) = \|\text{g(t,y(t),u(t))} - A(t)\| , \tag{72}$$

где $A(t)$ – семантическая карта ПрО прогнозируемой ИТ на шаге t, $\|$g(t,y(t),u(t))-A(t)$\|$ – удалённость ПрО оптимизируемой ИТ от ПрО прогнозируемой ИТ на шаге t, норма разности семантических карт соответствующих ПрО.

Чтобы определить критерий качества управления $u(t)$, необходимо предварительно ввести оценку качества научной предметной области $I(x)$. Для определения $I(x)$ может использоваться любая из вышеописанных оценок со

значениями, сводимыми к числовой величине, например средний ИФ статей в ПрО x. Однако, графосемантическая модель ПрО предполагает большую область определения оценки $I(x)$, вычислить которую во многих случаях не представляется возможным. Например, в случае использования экспертных оценок, экспертам необходимо оценить 2^n ПрО. Для решения данной проблемы предлагается использовать аппроксимацию функции $I(x)$. При таком подходе значение оценки вычисляется для нескольких предметных областей (точках области определения функции $I(x)$), затем по ним строится аппроксимирующая функция $\widetilde{I}(x)$, определяющая качество каждой допустимой ПрО x:

$$X = (x^1, x^2, \ldots, x^p), \tag{73}$$

$$\widetilde{I}(x^i) = I(x^i), i = \overline{1, p}, \tag{74}$$

где X – множество точек, для которых определена оценка $I(x)$, $|X| = p$.

Учитывая условия задачи, в качестве инструмента для построения аппроксимирующей функции $\widetilde{I}(x)$ была выбрана искусственная нейронная сеть. Искусственная нейронная сеть – математическая модель биологической нейронной сети – сети нервных клеток (нейронов) живого организма. ИНС получили широкое распространение в решении широкого круга задач: прогнозировании, распознавании образов, аппроксимации и т.д.

Ключевой особенностью ИНС является возможность автоматического перепрограммирования за счёт применения алгоритма обучения. Исходными данными для алгоритма обучения ИНС является обучающая выборка, вид которой определяется типом алгоритма обучения: с учителем или без. Для алгоритма обучения с учителем обучающая выборка включает набор допустимых входов совместно с соответствующими им ожидаемыми значениями выходов ИНС. Для алгоритма обучения без учителя обучающая выборка содержит только допустимые входные данные.

ИНС используются для решения большого количества задач: классификации, прогнозирования, аппроксимации и т.д. Тем не менее, у большинства классических нейросетевых моделей есть существенный недостаток: они представляют собой чёрный ящик. Фактически, это означает, что эксперт не

может интерпретировать состояние нейронной сети и произвольно влиять на результат её работы. Однако, существуют сети, работающие по принципу белого ящика, т.е. с интерпретируемым состоянием. Примером такой сети является ИНС Такаги-Сугено-Канга, основанная на аппарате нечёткой логики.

Ключевым элементом данной сети является нечёткой логический вывод Сугено. Его особенностью является чёткий результат, это означает, что выход сети не нуждается в дефаззификации.

Состояние ИНС Такаги-Сугено-Канга характеризуется набором линейных и нелинейных параметров. Линейные параметры являются частью следствий нечётких правил вывода Сугено и могут быть представлены как гиперплоскости в пространстве с числом измерений, равным числу входов сети. Нечёткие параметры находятся в посылках нечётких правил вывода Сугено. Эти параметры используются в функциях принадлежности, они могут быть визуализированы с помощью самих функций. Такая структура сети позволяет гибко настраивать вывод сети после обучения, что делает сеть удобной для решения многих задач.

Структура ИНС Такаги-Сугено-Канга схематично изображена на рисунке 12.

Слои ИНС Такаги-Сугено-Канга выполняют следующие функции:

1. Фаззификация входных переменных: $\mu_A(x_j)$, $j = \overline{1, N}$. На этом слое находятся нелинейные параметры $c_j^{(k)}, \sigma_j^{(k)}, b_j^{(k)}$, корректируемые в процессе обучения.

2. Агрегация – определение уровня активации каждого правила, вычисляется следующим образом:

$$\mu_A^{(k)} = \prod_{j=1}^{N} \mu_A(x_j), k = \overline{1, M}.$$

Данный слой не изменяется в процессе обучения.

3. Вычисление правых частей правил

$$y_k = \left(p_{k0} + \sum_{j=1}^{N} p_{kj} x_j \right) \mu_A^{(k)},$$

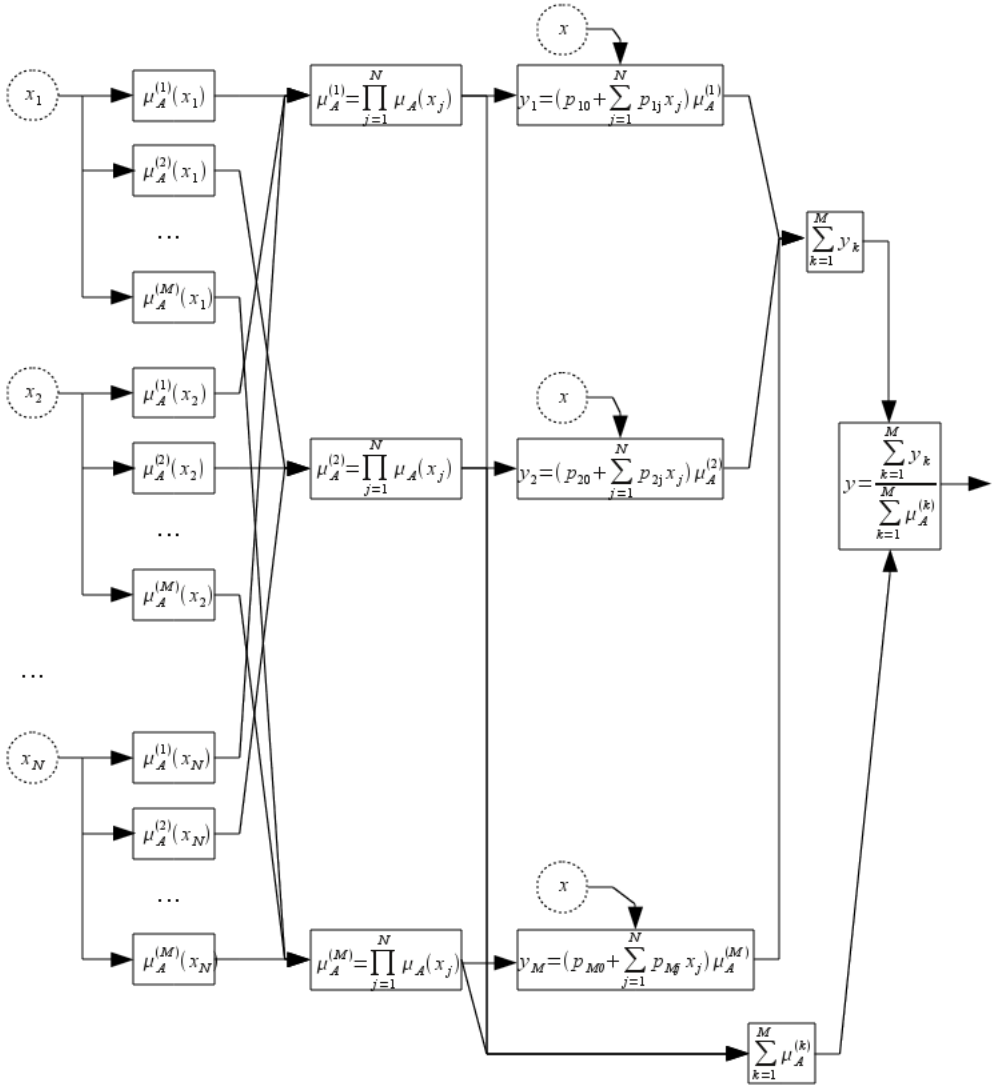

Рисунок 12 — Структура ИНС Такаги-Сугено-Канга

на этом слое обучаются линейные веса p_{kj} для $k = \overline{1, M}$ и $j = \overline{1, N}$.

4. На этом слое всего два суммирующих нейрона:

$$\sum_{k=1}^{M} y_k,$$

и

$$\sum_{k=1}^{M} \mu_A^{(k)}.$$

5. Нормализация:

$$y = \frac{\sum\limits_{k=1}^{M} y_k}{\sum\limits_{k=1}^{M} \mu_A^{(k)}}.$$

Таким образом, общая формула нечёткого вывода в ИНС Такаги-Сугено-Канга принимает следующий вид:

$$y(x) = \frac{1}{\sum\limits_{k=1}^{M} \prod\limits_{j=1}^{N} \mu_A^{(k)}(x_j)} \sum_{k=1}^{M} \prod_{j=1}^{N} \mu_A^{(k)}(x_j) \left(p_{k0} + \sum_{j=1}^{N} p_{kj} x_j \right).$$

Помимо описанных преимуществ, ИНС Такаги-Сугено-Канга обладает и рядом недостатков, важнейшим из которых является высокая трудоёмкость алгоритма обучения, связанного с корректированием линейных и нелинейных параметров. Для устранения данного недостатка был произведён анализ возможности повышения скорости обучения за счёт параллелизации с применением технологии OpenCL. Результаты анализа подробно описаны в работах
.

Однако, не смотря на полученный выигрыш в скорости, при большом количестве входных данных (семантических полей), обучение сети зачастую требует неприемлемо большие промежутки времени.

Таким образом, критерий качества управления научной предметной областью можно записать следующим образом:

$$J(u) = \sum_{t=0}^{T} \widetilde{I}(u(t)), \tag{75}$$

где $u(t) = (u_1(t), \ldots, u_r(t)), \forall t \in [t_0, T]$ – управление исследовательской траекторией предметной области агента научного производства.

В полученный критерий качества (75) введём вышеописанные ограничения в виде аддитивных штрафов:

$$J(u) = \sum_{t=0}^{T} \left(\alpha_0 \widetilde{I}(u(t)) - \sum_{i=1}^{4} \alpha_i R_i(t, y(t), u(t)) \right), \tag{76}$$

где α_0 – коэффициент значимости оценки, $\alpha_i, i = \overline{1,4}$ – коэффициент значимости i-го штрафа,

$$\alpha_i \geqslant 0, i = \overline{0,4},$$
$$\sum_{i=0}^{4} \alpha_i = 1.$$

Наиболее распространённым методом решения дискретных задач оптимального управления является метод динамического программирования Беллмана. Данный метод позволяет найти решение целого класса задач оптимального управления.

Существенной проблемой метода динамического программирования является высокая вычислительная сложность в задачах с фазовыми пространствами большой размерности. Кроме того, нахождение функции Беллмана в аналитическом виде является нетривиальной и, во многих случаях, очень сложной задачей.

Наиболее простой альтернативой является метод полного перебора всех возможных траекторий, вычисление критерия качества для каждой из них и выбор оптимальной. В большинстве случаев данный метод неприменим из-за огромной вычислительной сложности (перебор всех возможных траекторий), однако в некоторых задачах его удаётся существенно оптимизировать, например за счёт отсечения заведомо неэффективных траекторий.

В данной работе был выбран компромиссный метод решения задачи оптимального управления на основе глобальной оптимизации критерия качества управления $J(u)$ с помощью генетического алгоритма. Генетический алгоритм так же подразумевает перебор траекторий и оценку их качества («приспособленности»), однако, в отличии от полного перебора, оцениваются не все траектории, а выбираемые на основе принципа, имитирующего поведение биологических организмов. Генетические алгоритмы, в отличие от стандартного метода Беллмана, хорошо приспособлены для работы с бинарными данными.

Основной идеей данного метода является глобальная оптимизация ИТ АНП по критерию качества управления $J(u)$:

$$J(\widehat{u}) = \max_{u \in U} J(u).$$

Однако, управление $u(t)$, как правило, не присутствует в явном виде в (76). Для решения этой проблемы, воспользуемся модифицированной функцией Беллмана . Оригинальная функция Беллмана на каждом шаге выбирает частичную оптимальную траекторию начиная с этого шага . Для реализации рассматриваемого метода в этом нет необходимости, т.к. траектория оптимизируется глобально, поэтому модифицируем функцию Беллмана следующим образом:

$$\psi^*(y_0, t_0, \widehat{u}) = \max_{u \in U} \psi^*(y_0, t_0 + 1, u), \tag{77}$$

$$\psi^*(y, t, u) = \alpha F(y, u(t), t) + \psi^*(f(y, u(t), t), t + 1, u), t = \overline{t_0 + 1, T - 1}, \tag{78}$$

$$\psi^*(y, T, u) = (1 - \alpha)\Phi(y). \tag{79}$$

Очевидно, функция (77-79) осуществляет глобальную оптимизацию траектории по управлению $u(t)$, максимизируя критерий качества $J(u)$ (76) за счёт рекуррентных вызовов.

Генетический алгоритм относится к итерационным алгоритмам глобальной оптимизации. Основным объектом, с которым работает генетический алгоритм, является особь, соответствующая одному допустимому решению задачи. Каждая особь характеризуется вектором параметров, называемым хромосомой. Элементы хромосом называются генами. В данной задаче особью является траектория ПрО АНП y, а управление $u \in U$ – хромосомой. В качестве генов используется управление на заданном шаге $u(t), t = \overline{t_0, T - 1}$.

Перед началом итерации ГА выполняется генерация исходной популяции – множества особей со случайными хромосомами:

$$P^1 = \left\{p_1^1 \dots p_N^1\right\}, p_i^1 \in Y, i = \overline{1, N},$$

где N – размер популяции, зависит от задачи и определяется экспертом, P^1 – исходная популяция, p_i^1 – i-ая особь исходной популяции.

Каждая итерация состоит из следующих шагов:

1. Оценка приспособленности каждой особи. Приспособленность определяется эффективностью особи в качестве решения поставленной задачи. Обычно, в ГА максимизируется значение функции приспособленности, т.е. чем больше её значение, тем лучше приспособлена особь. В данной задаче функция приспособленности определена равная качеству управления:

$$Fitness(u) = J(u),$$

где $Fitness(u)$ – приспособленность особи с хромосомой $u \in U$.

2. Селекция. На данном этапе происходит выбор фиксированного числа особей для следующего этапа. Выбор может производиться разными способами, например выбираются пары наиболее приспособленных особей. Однако, рекомендуется выбирать особей из всей популяции, иначе алгоритм может зайти в локальный экстремум. В данной задаче выбор производится случайным образом, однако каждая особь выбирается с вероятностью, пропорциональной её приспособленности (принцип рулетки).

3. Скрещивание. Выбранные на этапе селекции (предки), используются в качестве операндов для оператора скрещивания, в результате которой генерируются новые особи (потомки), наследующие гены предков в определённом сочетании. Способ сочетания генов предков в потомках индивидуально определяется для каждой задачи. При этом возможно использование оператора скрещивания, генерирующего на основе v предков w потомков (в таком случае возможна тонкая настройка алгоритма с помощью параметров v и w).

2.4 Выводы по главе

1. Разработана графосемантическая модель предметной области агента научного производства. В основе лежит метод графосемантического мо-

делирования. Так же предложена методика выделения частнонаучных предметных областей на основе агоритма нечёткого кластерного анализа C-means.

2. Разработана математическая модель исследовательской траектории агента научного производства. На основе полученной модели разработана методика прогнозирования исследовательских траекторий агентов научного производства с применением имитационного моделирования.

3. Поставлена задача оптимального управления предметной областью агента научного производства для поиска оптимальной исследовательской траектории.

4. Предложен метод решения задачи оптимального управления предметной областью агента научного производства на основе метода динамического программирования Беллмана.

5. Предложен метод аппроксимации оценки качества предметной области агента научного производства на основе нечёткой нейронной сети Такаги-Сугено-Канга.

6. Разработан численный метод решения задачи оптимального управления предметной областью агента научного производства на основе генетического алгоритма.

3 Программное обеспечение для моделирования и оптимизации исследовательских траекторий

3.1 Информационная система «Семограф»

Применение вышеописаных моделей к практическим задачам с большими объёмами входных данных в ручном режиме является крайне трудоёмким процессом. Поэтому в рамках данной работы была реализована информационная система (ИС) «Семограф» [95], позволяющая автоматизировать большую часть трудоёмких операций [95; 96]. ИС «Семограф» создавалась как пилотный проект и начала функционировать в 2010 году в виде прототипа с минимальной функциональностью. Несмотря на отсутствие какой бы то ни было целевой рекламы, в ИС «Семограф» на сегодняшний момент зарегистрирован 340 пользователей, создано 589 проектов, внесено в базу данных 226050 контекстов, выделено 3380 полей и 321171 компонентов. Эти данные являются свидетельством востребованности системы, того инструментария, который предлагается исследователю. Если на начальном этапе ИС создавалась для программной реализации метода графосемантического моделирования, то в настоящее время функциональные возможности ИС значительно превосходят первоначальные требования и сама ИС может быть определена как информационно–экспертная система, предназначенная для извлечения знаний о предметных областях из информационных массивов, включающих текстовые выборки, метаданные, семантические компоненты и семантические поля. Некоторые инструменты, разработанные как часть ИС «Семограф», такие как система метаданных, стали неотъемлемой частью метода графосемантического моделирования. На данный момент ИС «Семограф» является источников большого объёма обработанных данных, описывающим различные предметные области. Разработанная система зарегистрирована в Федеральной службе по интеллектуальной собственности, патентам и товарным знакам, копия соответствующего свидетельства находится в приложении 1.

ИС «Семограф» представляет собой распределённую информационную систему с web-интерфейсом. Подобные системы так же называют веб-приложениями. Выбор такого типа программного обеспечения обусловлен ак-

тивизацией развития новых Интернет-технологий, объединённых понятием «Web 2.0». Кроме того, веб-приложения удобнее для конечного пользователя по сравнению с традиционными, поскольку не требуют установки дополнительного прикладного программного обеспечения, для работы с ними достаточно стандартного веб-клиента (браузера веб-страниц). К преимуществам веб-приложений так же можно отнести кросплатформенность (переносимость между программно-аппаратными платформами) – веб-приложение будет корректно работать в любой операционной системе, где доступен браузер (включая мобильные устройства).

Любое web-приложение представляет собой распределённую систему клиент-серверной архитектуры [97]. При этом архитектура уровней системы может варьироваться. С развитием Web 2.0 все большее число функций переносится на клиент. Это связано с возросшей вычислительной мощностью клиентских устройств и производительностью браузеров. Если раньше web-сервер генерировал статическую HTML-страницу (HyperText Markup Language, язык гипертекстовой разметки) а задачей клиента была лишь визуализация этой страницы, то теперь сервер, как правило, лишь передаёт браузеру набор данных, где производится конструирование DOM (Document Object Model, Объектная Модель Документа) и визуализация страницы. Помимо снижения нагрузки на сервер, это так же позволяет динамически изменять web-страницу и реализовать интерактивный пользовательский интерфейс. Ключевыми факторами в этом процессе стало внедрение языка программирования JavaScript в браузер, появление технологии AJAX (Asynchronous JavaScript and XML) и концепции RIA (Rich Internet Application – веб-приложения с «насыщенным» интерфейсом). На рисунке 13 приведена диаграмма процесса изменения архитектуры клиент-сервер web-приложений.

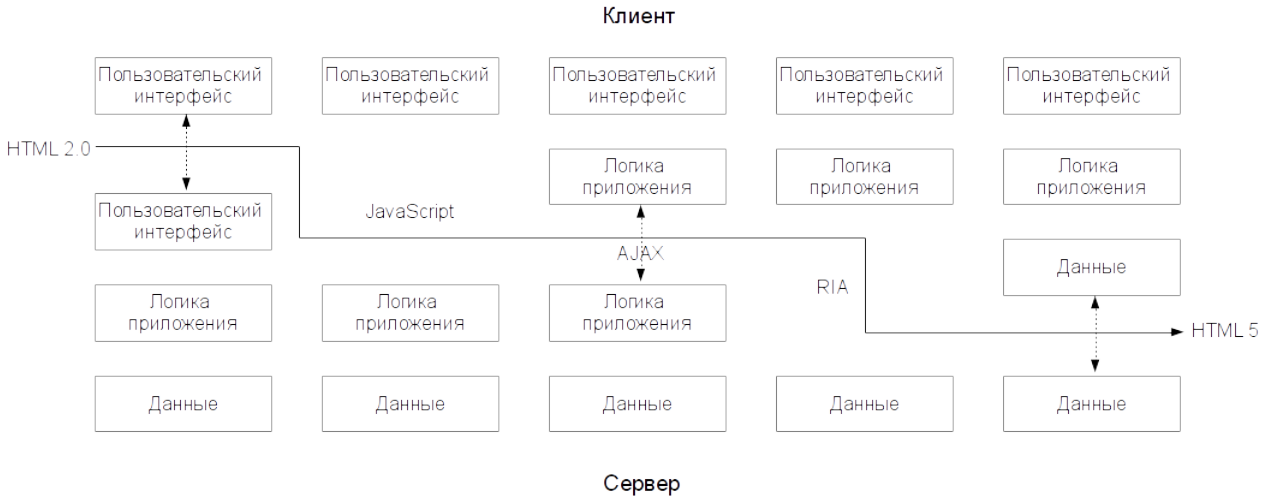

Рисунок 13 — Развитие архитектуры клиент-сервер

Разработка информационной системы выполнялась в соответствии с комплексом стандартов ISO/IES 12207-2002, ISO TC97/SC7 #383 (Criteria for Evaluation of Software) и ГОСТ 34. При разработке ИС «Семограф» за основу были взяты такие базовые свойства программного обеспечения, как масштабируемость, интероперабельность и переносимость. ИС разрабатывалась с использованием современных технологий и подходов, среди которых:

1. HTML5 – стандарт World Wide Web Consortium (W3C), пришедший на смену устаревшим HTML4.1 и Adobe Flash. Предоставляя куда более широкие возможности представления информации и взаимодействия с пользователем, HTML5 при этом не требует от пользователя установки дополнительных плагинов и расширений для браузера.

2. CSS3 – как и HTML5, является стандартом W3C (и обычно они используются совместно). CSS3 позволяет добиться качественно нового уровня пользовательского интерфейса с использованием совершенно новых возможностей каскадных таблиц стилей 3 уровня. Благодаря использованию этих стандартов как основополагающих, для работы с ИС пользователю необходим лишь современный веб-браузер с поддержкой этих стандартов.

3. Coding by convention – парадигма «программирование по соглашению», лежит в основе выбранной инфраструктуры Groovy/Grails, позволяет существенно сократить время разработки.

4. OAuth 2.0 – протокол аутентификации, обеспечивающий прозрачный доступ сторонних приложений к данным ИС.

5. REST – Representational State Transfer, вид прикладного интерфейса распределённых ИС, организует удобный доступ к любым информационным ресурсам.

6. WebSockets – часть стандарта HTML5, позволяют обеспечить синхронизацию общих данных между клиентами, при этом инициатором обмена данными выступает web-сервер, а не клиент.

7. OpenCL – технология выполнения параллельных вычислений на графическом процессоре, используется для повышения эффективности параллельных алгоритмов обработки данных.

При разработке ИС «Семограф» предпочтение отдавалось свободному программному обеспечению. Так, в серверной части задействованы следующие программные продукты:

1. MySQL – реляционная система управления базами данных, используется для хранения данных.

2. OpenJDK 7 – платформа, являющаяся основой для большинства компонентов системы, реализует стандарт Java 7.

3. Groovy/Grails – инфраструктура для создания веб-сайтов на основе языка Groovy. Язык Groovy представляет собой динамически типизируемый язык программирования высокого уровня, компилируемый в байт-код виртуальной машины Java (JVM). Такое решение позволяет совместить преимущества от использования обширной экосистемы Java с относительной простотой разработки сложных приложений, свойственной динамически типизируемым языкам.

4. Solr – сервер для индексации и полнотекстового поиска текстовой информации, основан на Apache Lucene.

В клиентской части (веб-страница, исполняется в браузере пользователя) применяются:

1. JavaScript – динамический язык, доступный в любом браузере, используется для реализации интерактивности веб-страниц. На данный момент является единственным широко распространённым языком программирования для клиентской части web-приложений, не требуеющим установки дополнительных приложений и модулей.

2. jQuery – широко распространённая JavaScript-библиотека, основной задачей которой является упрощения взаимодействия разработчика с объектной моделью HTML-документа.

3. UnderscoreJS – JavaScript-библиотека, расширяющая возможности языка для работы с массивами и объектами.

4. BackboneJS – JavaScript-библиотека, упрощающая реализацию паттерна проектирования MVC (Model View Controller). Используется для реализации динамически изменяющегося пользовательского интерфейса.

5. DataJS – JavaScript-библиотека для хранения и манипулирования структурированными данными. Используется для представления всех данных на клиенте.

Для разработки ИС так же применялись следующие средства:

1. Springsource Tool Suite/Eclipse – интегрированная среда разработки (IDE).

2. Git – система контроля версий, используется для хранения и работы с исходным кодом.

3. R – язык программирования высокого уровня, предназначенный для статистической обработки данных.

4. Gephi – программное средство для построения и анализа графов.

Важной особенностью разработанной ИС «Семограф» является применение концепции рабочих столов, агрегирующих рабочее окружение пользователя вокруг некоторой задачи (например набор окон, отображающих различные

части одного проекта либо логически связанных проектов). Рабочий стол веб-приложения аналогичен по своим возможностям рабочему столу операционной системы, но работает в одном окне веб-браузера и позволяет пользователю сосредоточиться над решаемой задачей без необходимости переключения окон и вкладок браузера. Рабочий стол в ИС «Семограф» позволяет одновременно отображать элементы одного или нескольких проектов в отдельных окнах, которые могут быть перемещены, свёрнуты или закрыты. Пример такого окна со списком контекстов приведён на рисунке 14.

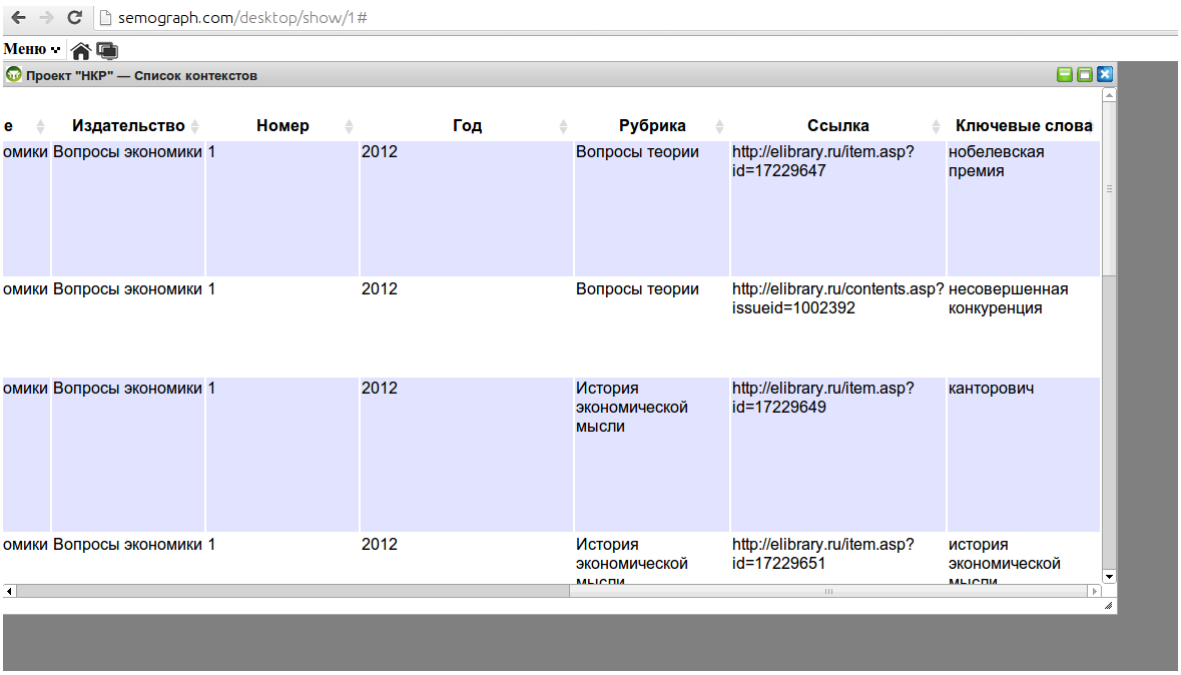

Рисунок 14 — Пример окна ИС «Семограф»

3.2 Структура данных информационной системы

Описанные выше преимущества актуальны и для информационных систем. Однако, текущие ограничения платформы Web 2.0 (в частности, отсутствие поддержки распределённых систем объектов) не позволяют эффективно оперировать сложноструктурированными данными, что затрудняет реализацию информационных систем в виде веб-приложений.

К данным со сложной структурой (сложноструктурированным) относятся массивы, списки, деревья, графы, сети и их комбинации [57; 66; 69; 94]. Основная проблема состоит в сложности поддержки актуальности DOM веб-страницы и синхронизации с сервером. Кроме того, в современных веб-

приложениях (особенно, в информационных системах) объёмы данных, с которыми работает пользователь, могут быть очень большими. Передача таких объёмов данных через Интернет может негативно сказаться на скорости реакции пользовательского интерфейса приложения. Следовательно, важной задачей является минимизация объёма передаваемых данных, необходимых для выполнения той или иной операции и устранение повторной загрузки данных.

В качестве решения данной проблемы предлагается ввести отдельный (вспомогательный) уровень данных на стороне клиента. Следует отметить, что стандарт HTML5 добавил в браузеры поддержку LocalStorage – реляционной БД, позволяющей веб-приложениям перманентно хранить в браузере произвольные данные. Однако, современные методы разработки информационных систем подразумевают использование средств более высокого уровня, чем прямое общение с БД через SQL, таких как ORM (Object-Relational Mapping, Объектно-Реляционное Отображение) [40]. Тем не менее, LocalStorage предоставляет возможность хранения данных на стороне клиента, что позволяет избежать необходимость повторной загрузки не изменявшихся данных при каждом открытии приложения. Наиболее логичным решением является реализация дополнительного уровня абстракции между LocalStorage и веб-приложением, предоставляющим удобный интерфейс программирования и использующий возможности LocalStorage.

В ходе реализации такого уровня абстракции, решающего описанные выше проблемы, была разработана концепция унифицированного ориентированного графа (UOG, Unified Oriented Graph). UOG представляет собой частный случай сетевой базы данных. Сетевые базы данных менее эффективны в плане скорости доступа и занимаемого объёма по отношению к реляционным БД, однако они позволяют естественным образом представлять сложносвязные структуры данных. Кроме того, схема сетевой базы данных, как правило, не накладывает жёстких ограничений на типы хранимых данных, что может быть полезным при хранении сложных структур данных.

Основной идеей унифицированного ориентированного графа является представление всех полученных с сервера данных в единой структуре (графе, сетевой БД) на клиенте, для обеспечения единого (унифицированного) интерфейса доступа к данным и ассоциативным связям между ними (ориен-

тированность). Для этого ассоциативные связи представляются в виде узлов графа, наряду с прочими данными.

Концепция UOG близка к модели OEM (Object Exchange Model), разработанной для представления слабоструктурированных данных [70; 86]. Однако, в отличии от модели OEM, UOG имеет более жёсткую схему (которая является частью самого графа), позволяющая производить валидацию данных и контролировать их структуру до синхронизации с сервером. Схема UOG задаётся априорно и не допускает появления в модели неструктурированных данных.

При использовании концепции унифицированного ориентированного графа можно выделить следующие основные задачи: обмен данными с сервером, выборка узлов, визуализация данных, слежение за обновлениями данных.

Поскольку web-приложения не поддерживают распределённые системы объектов (CORBA, DCOM, GLOBE, SOAP), для решения задачи обмена данными (протокола синхронизации уровней представления данных сервера и клиента) возможно использование одного из способов обмена сообщениями в распределённых системах, таких как XML-RPC, JSON, REST. XML-RPC (Remote Procedure Call, удалённый вызов процедур на основе XML) является одним из первых протоколов, обеспечивающих удалённый вызов функций [97]. Как правило, он не используется в web-приложениях, поскольку с Web 2.0 появились более удобные способы обмена данными. JSON (JavaScript Object Notation) – текстовый формат данных, предоставляющий прозрачную сериализацию основных структур JavaScript и легко читаемый людьми. REST (Representational State Transfer, Передача состояния представления) – набор простых принципов доступа и манипуляции данными через унифицированный интерфейс, реализующий набор операций CRUD (Create, Read, Update, Delete) поверх протокола HTTP (HyperText Transport Protocol) [48]. В качестве формата данных REST обычно использует JSON. Подход REST представляется наиболее естественным для синхронизации данных между разными уровнями представления данных.

Основными понятиями REST являются ресурс и коллекция ресурсов.

Ресурсом является любая сущность, которой оперирует ИС. Любой ресурс входит в некоторую коллекцию ресурсов, в пределах которой он должен иметь уникальный идентификатор. Например:

1. /rest/contexts – коллекция контекстов;
2. /rest/contexts/15 – контекст с идентификатором 15;
3. /rest/contexts/15/components – коллекция компонентов контекста с идентификатором 15.

Следует подчеркнуть, что REST не накладывает никаких ограничений на способ представления адресов ресурсов и коллекций. Так, пример 3 может быть представлен следующим образом: /rest/components/15?context_id=15 – коллекция компонентов с параметром запроса <идентификатор_контекста=15>.

В рассматриваемой задаче необходимо учесть, что нужно предоставить доступ к связям данных на стороне сервера как к ресурсам. Для доступа к связям данных с помощью REST можно использовать три подхода:

1. Присвоить каждой связи уникальный идентификатор относительно других связей этого же типа. То есть, необходимо завести отдельную коллекцию связей: /rest/context_has_comopnent/183 – у контекста есть компонент, связь с идентификатором 183. Такой подход удобен, поскольку не требует никаких изменений в логике UOG для работы с узлами, хранящими связи. Однако, хранение идентификатора связи влечёт накладные расходы на сервере БД. В реляционных БД связи типа многие-ко-многим, как правило, хранятся в виде отдельной таблицы, содержащей кортежи (<идентификатор первого связываемого объекта>; <идентификатор второго связываемого объекта>; [<дополнительные поля>]) [40], следовательно, включение идентификатора связи в кортеж может увеличить размер записи в 1,5 раза в случае отсутствия дополнительных полей. Учитывая потенциально большое количество связей (относительно обычных сущностей), это серьёзный недостаток данного подхода.
2. Очевидно, коллекция связей может быть выражена как подколлекция ресурсов, связываемых ею. Однако, такой способ представления

оставляет неоднозначность выбора уникального идентификатора связи: /rest/contexts/15/components/3 – компонент с идентификатором 3 контекста с идентификатором 15. /rest/components/3/contexts/15 – компонент с идентификатором 3 контекста с идентификатором 15.

3. Использование параметров строки запроса. Для этого подхода, как и для первого, требуется коллекция связей. Однако, внутри неё ресурсам (связям) не будут присваиваться уникальные идентификаторы. Вместо этого обращение к ресурсам будет осуществляться через параметры строки запроса: /rest/context_has_component?context_id=15&component_id=3 – компонент с идентификатором 3 контекста с идентификатором 15. Данный подход лишён недостатков предыдущих подходов, но следует заметить, что приведённый пример фактически является вызовом метода list, т.е. получения коллекции. Фактически, это означает, что результатом выполнения такого запроса будет массив объектов. Однако, уникальность связей должна обеспечиваться на уровне БД.

Очевидно, сложные структуры данных подразумевают сложную логику представления. Для реализации интерфейса таких информационных систем зачастую используют паттерн MVC (Model-View-Controller, Модель-Представление-Контроллер) [87; 102]. Основополагающей идеей данного паттерна является отделение данных от представления. Это означает, что одни и те же данные могут быть одновременно представлены разными способами, при этом обновление данных в модели автоматически вызывает обновление всех представлений этих данных посредством контроллера (обычно с помощью паттерна «Наблюдатель»).

Рассмотрим особенности использования UOG в качестве модели в составе паттерна MVC. При реализации этого паттерна часто бывает необходимо представлять целые коллекции объектов как единое целое. Удобство UOG заключается в возможности динамического формирования произвольных коллекций из узлов графа (срезов) посредством запросов с возможностью наблюдения за изменениями в этих коллекциях. Эта возможность основывается на механизме выборок узлов из графа, описанном ниже. Значительным достоинством данного подхода является возможность повторного использования пред-

ставлений для визуализации (в том числе одновременной) узлов одного типа, выбранных на основе связей с узлами различных типов, путём простой замены запроса. На рисунке 15 представлена схема клиент-серверного приложения с использованием UOG.

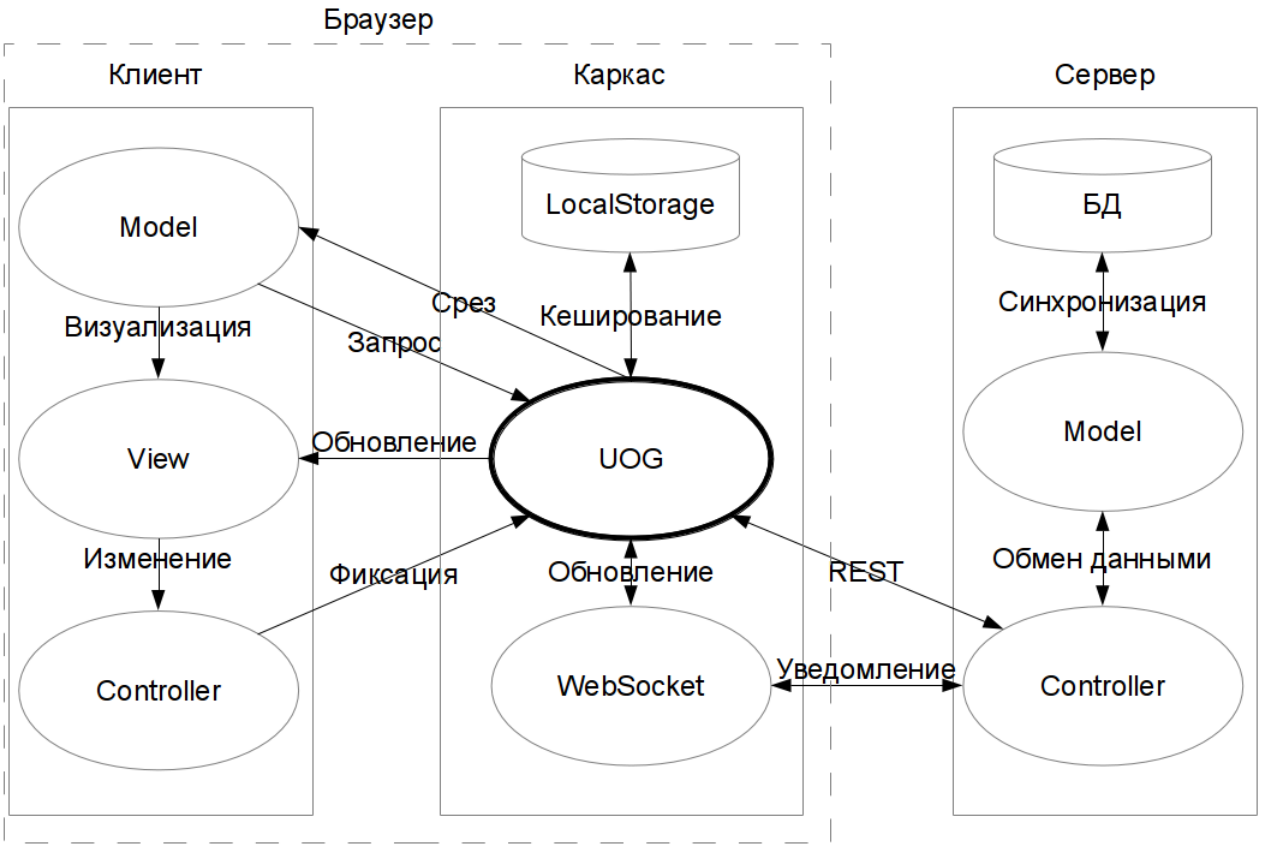

Рисунок 15 — UOG в клиент-серверном приложении

Помимо визуализации, использование UOG в качестве модели упрощает редактирование данных и синхронизацию с сервером, поскольку предоставляет единый интерфейс доступа к данным, их созданию и изменению.

Как уже отмечалось, во многих информационных системах объёмы данных, с которыми работает пользователь, слишком велики, чтобы загружать их единовременно. Поэтому необходим механизм отложенной загрузки данных, т.е. узлов графа в концепции UOG. Очевидно, в большинстве случаев объём необходимых пользователю данных значительно меньше объёма всех доступных ему данных, следовательно, механизм отложенной загрузки должен загружать только необходимые данные, с минимальной избыточностью,

благодаря чему снизится время загрузки (и, как следствие, время реакции интерфейса) и нагрузка на сервер. В рамках концепции UOG были разработаны два подхода к реализации механизма отложенной загрузки данных:

1. Узлы-пустышки (Dummy Nodes);
2. Ленивые узлы (Lazy Nodes).

Для выборки узлов из графа используется алгоритм, принимающий на вход объект JavaScript (ассоциативный массив) с параметрами запроса и последовательно анализирующий доступные узлы. Это же механизм используется для наблюдения за новыми и удаляемыми узлами. Для оптимизации поиска узлы могут быть упорядочены по типу и другим вспомогательным полям.

Описанный подход был опробован на второй версии системы графосемантического моделирования «Семограф». UOG разрабатывалась для хранения данных в ИС «Семограф» на стороне клиента. Первоначально для этих целей использовались классические коллекции (реализация из библиотеки BackboneJS), однако, особенности структур данных «Семографа» значительно усложняли этот подход (ядро структуры схематично представлено на рисунке 16). Такая структура стала следствием денормализации реляционной модели, проведённой с целью устранения дублирования больших объёмов текстовых данных.

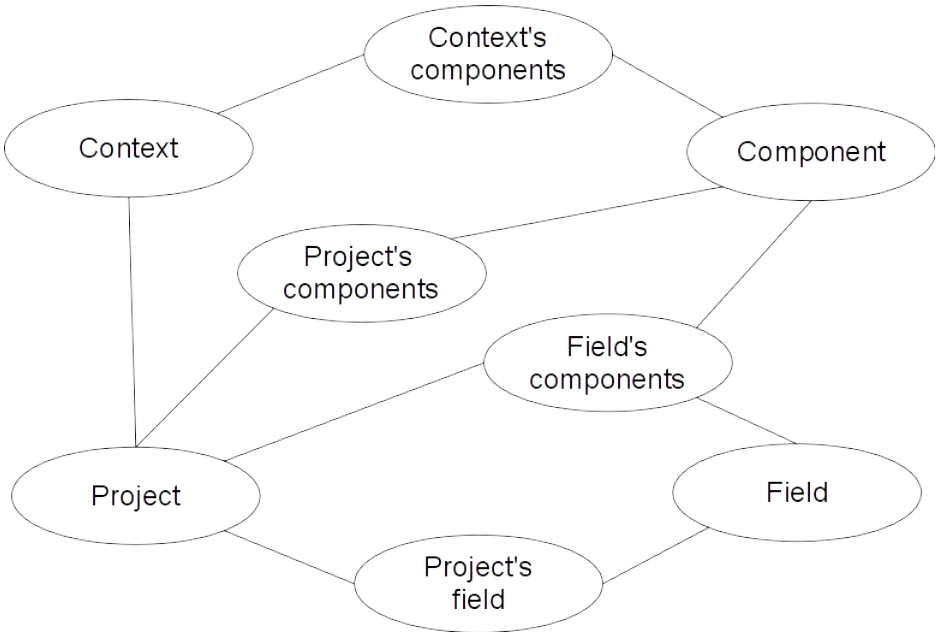

Рисунок 16 — Схема ядра структуры данных ИС «Семограф»

Проблема заключается в наличии сложных взаимосвязей между типами данных (например один компонент может одновременно принадлежать проекту, нескольким контекстам этого проекта и нескольким полям этого проекта). Для манипулирования подобными данными были испытаны два подхода:

1. хранение нескольких экземпляров одного объекта (в данном случае, компонента), используя для поддержания целостности некоторый дескриптор (например, первичный ключ из БД);

2. использование проекций коллекций (срезов), подразумевает наличие общей коллекции для хранения однотипных объектов и управляющего объекта (аггрегирующей коллекции), осуществляющего выборку из общей коллекции необходимых объектов.

Первый подход неудобен из-за возможной потери целостности данных, второй – из-за трудоемкости контроля и синхронизации управляющего объекта, коллекции и необходимости поддержки нескольких объектов управления. При всестороннем рассмотрении проблемы, в частности, представлении структуры данных в виде, изображённом на рисунке 16, стало очевидно, что самым логичным способом хранения этих данных является представление в виде графа. Эта идея стала основой концепции UOG.

Помимо ядра структуры данных (приведённой на рисунке 16), UOG включает большое количество переферийных и вспомогательных данных, не связанных непосредственно с реляционной моделью данных ИС. К ним относятся: HTML-шаблоны, элементы интерфейса (окна, их положение и размер), сведения о рабочих столах. Подобное единое представление данных позволило организовать унифицированный интерфейс доступа к любым данным клиента. В ИС «Семограф» этот интерфейс реализован посредством специального языка запросов, основанного на библиотеке DataJS. Особенности данных системы потребовали существенной доработки и оптимизации интерпретатора запросов.

3.3 Выводы по главе

1. Разработана информационная система «Семограф» для моделирования, оценки и оптимизации исследовательских траекторий агентов научного производства. Разработанная система зарегистрирована в Федеральной службе по интеллектуальной собственности, патентам и товарным знакам.

2. Разработаны модули на языке R, реализующие трудоёмкие численные методов.

3. Предложена концепция унифицированного объектного графа (UOG), обеспечивающего представление всех полученных с сервера данных в единой структуре (графе, сетевой БД) на клиенте для обеспечения единого (унифицированного) интерфейса доступа к данным и ассоциативным связям между ними. Так же применение модели на основе UOG упрощает синхронизацию данных с сервером.

4 Решение практических задач

4.1 Построение оптимальной исследовательской траектории научного журнала

4.1.1 Графосемантическая модель предметной области журнала «Вопросов экономики»

В качестве одного из агентов научного производства для апробации полученных теоретических результатов был выбран научный журнал «Вопросы экономики». «Вопросы экономики» является ведущим в России теоретическим и научно-практическим журналом общеэкономического содержания. Данный журнал находится в списке наиболее рейтинговых журналов РФ. С 2007 г. «Вопросы экономики» был включен в список российских научных журналов ВАК Минобрнауки России. Импакт-фактор издания – 3,354 (импакт-фактор РИНЦ 2013), журнал включен в международные базы цитирования. По мнению ряда ведущих экономистов РФ, журнал «Вопросы экономики» играет системообразующую роль в российской экономике. «Вопросы экономики» называют журналом академическим и специализированным, на протяжении многих лет поддерживающим высокие стандарты качества (Г. Фетисов, М. Ершов и др.). Аудиторией журнала являются экономисты-исследователи, преподаватели и студенты вузов, руководители федеральных и региональных органов власти, отвечающие за разработку экономической политики, аналитические подразделения крупных предприятий, корпораций и банков.

Заметим, что даже не смотря на статус выбранного журнала, и в предположении охвата им значительной части предметной области экономических наук, нельзя исключать возможную субъективность в выборе публикуемого материала. Для снижения влияния данного фактора необходимо построение макромодели предметной области экономических наук на основе источника более широкого профиля. Кроме того, макромодель может быть использована для выбора целевых предметных областей на этапе решения задачи оптимального управления предметной областью выбранного агента научного производства.

Очевидно, важным фактором успешной научной деятельности является финансовая поддержка со стороны различных государственных и негосударственных фондов. Кроме того, поддержка в виде грантов отдельных научных работ (проектов) характеризует актуальность их предметных областей, соответствие современным требованиям науки. Следовательно, в качестве макромодели может быть использована графосемантическая модель предметной области научных проектов, поддерживаемых научными фондами. Для «Вопросов экономики» была разработана функция оценки качества предметной области $I(x)$ на основе дополнительной макромодели предметной области Ω_{rfh}, основанной на данных о поддержанных Российским Гуманитарным Научным Фондом (РГНФ) проектах.

На материале журнала «Вопросы экономики» за 2010-2013 годы была построена графосемантическая модель Ω_{ve}. Модель Ω_{ve} включает 2055 компонентов и 414 контекстов. Поскольку функция оценки качества предметной области $I(x)$, используемая при построении оптимальной исследовательской траектории для «Вопросов экономики», основана на модели Ω_{rfh}, множества семантических полей моделей Ω_{ve} и Ω_{rfh} должны совпадать:

$$F_{\Omega_{rfh}} = F_{\Omega_{ve}}.$$

В таблице 13 приведено множество семантических полей, полученное при построении макромодели Ω_{rfh}, а также их частотности в соответствующих моделях Ω_{ve} и Ω_{rfh}. Семантический граф полученной модели Ω_{ve} приведён на рисунке 17.

Таблица 13 — Семантические поля графосемантических моделей Ω_{ve}, Ω_{rfh}

№	Поле	Частотность	
		ВЭ	РГНФ
1	Методология исследований	93	128
2	Научные направления	120	175
3	Виды деятельности	54	74
4	Территориальные образования и географическое расположение	80	101
5	Институциональные структуры	133	111
6	Денежно-кредитная политика	193	144
7	Безопасность	180	132
8	Фундаментальная наука	95	134
9	Система образования	129	130
10	Политическая деятельность	208	153
11	Макроэкономика	259	207
12	Микроэкономика	170	185
13	Демография и миграционные процессы	75	85
14	Качество жизни населения	154	129
15	Ресурсный потенциал	186	175
16	Особенности рыночного хозяйствования	137	94
17	Стратегические направления развития экономики	189	196

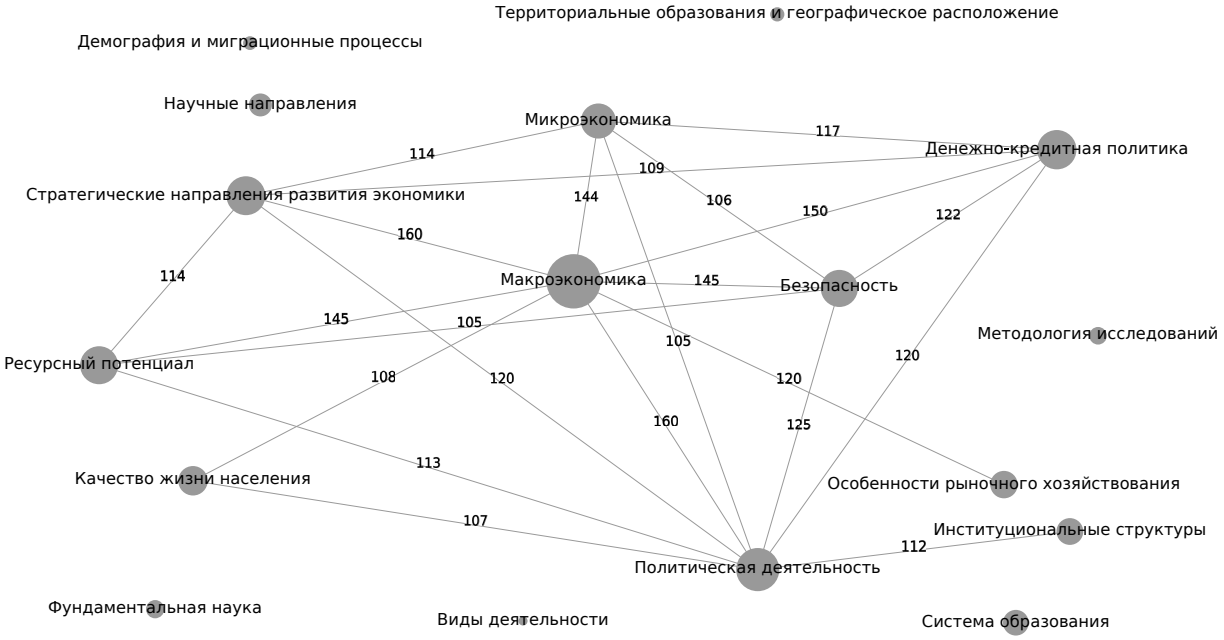

Рисунок 17 — Семантический граф исходной модели предметной области «Вопросов экономики»

4.1.2 Макромодель предметной области

В ходе построения макромодели предметной области экономических наук Ω_{rfh} были собраны исходные данные о 889 поддержанных проектах РГНФ в 2009-2013 годах в области знаний «02 - Экономические науки». РГНФ размещает основную информацию о поддержанных проектах в открытом доступе на своём сайте [89].

Собранные данные включают следующие типы описательной информации (мета–данных):

1. Название проекта;
2. Область знаний;
3. Тип проекта;
4. Название проекта;
5. Год начала проекта;
6. Ссылка на описание проекта на сайте РГНФ;
7. Ключевые слова, описывающие публикацию с точки зрения её авторов.

Среди описанных типов мета–данных особый интерес представляет «Тип проекта». Поля данного типа содержат одно из возможных 13 значений:

1. а - проект проведения научных исследований, выполняемый научным коллективом или отдельным учёным;
2. а1 - проект проведения научных исследований, выполняемый коллективом (до 9 человек) молодых учёных под руководством ведущего учёного без ограничения возраста;
3. а2 - проект проведения научных исследований, выполняемый коллективом (до 10 человек), состоящим полностью из молодых учёных, включая руководителя;
4. а(м) - совместный проект проведения научных исследований, выполняемый научным коллективом (международный конкурс);

5. а(р) - проект проведения научных исследований, выполняемый научным коллективом или отдельным учёным (региональный конкурс);

6. а(ф) - проект проведения научных исследований, выполняемый отдельным учёным;

7. а(ц) - проект проведения междисциплинарных исследований с изданием научных трудов по результатам исследований;

8. в - проект создания и приобретения программного обеспечения для информационных систем в области гуманитарных наук, способствующих распространению гуманитарных знаний в обществе;

9. г - проект организации в рамках реализации научных проектов мероприятий, в том числе конференций и семинаров;

10. г(р) - проект организации в рамках реализации научных проектов мероприятий, в том числе конференций и семинаров (региональный конкурс);

11. д - проект издания научных трудов по результатам научных исследований, проводимых в рамках научных проектов, профинансированных РГНФ;

12. е - проект экспедиций, полевых и социологических исследований, научно-реставрационных работ, необходимых для получения новых данных в области гуманитарных наук;

13. к - проект подготовки научно-популярного издания;

«Тип проекта» играет важное значение, поскольку естественным образом определяет классы моделей предметных областей научных проектов, разделяя их по целям, составу исполнителей, территориальным и другим признакам. На рисунке 18 приведено распределение количества проектов по приведённым выше значениям.

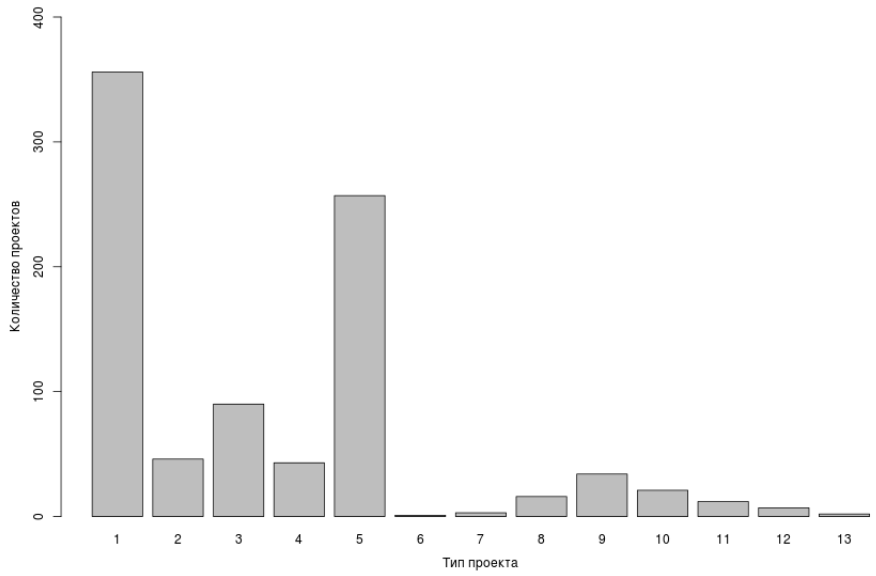

Рисунок 18 — Распределение количества проектов по типам проектов

Как видно из рисунка 18, наибольшее число поддержанных проектов приходится на 1 и 5 типы. В данном исследовании рассматривается только множество проектов первого типа (а - проект проведения научных исследований, выполняемый научным коллективом или отдельным учёным), как наиболее унифицированное и широко представленное. При построении графосемантической модели предметной области на основе данного множества проектов было сформировано множество контекстов. Полученное множество включает 356 контекстов.

В процессе подготовки исходных данных было выделено 4468 уникальных ключевых слов. Из наиболее частотных ключевых слов было сформировано множество семантических компонентов C (3830 элементов). Полученные компоненты были объединены экспертом в семантические поля. Для этого были определены множества семантических полей F (17 полей) и множество связей поле–компонент $\Lambda(C, F)$ (2833 связей). Полученное множество семантических полей приведено в таблице 13. Для выполнения данной работы были привлечены учёные-экономисты.

На рисунке 19 приведено распределение количества контекстов по количеству связанных полей (мощности предметной области).

Из рисунка 19 видно, что значительное количество контекстов не связано с семантическими полями. Появление таких контекстов возможно в резуль-

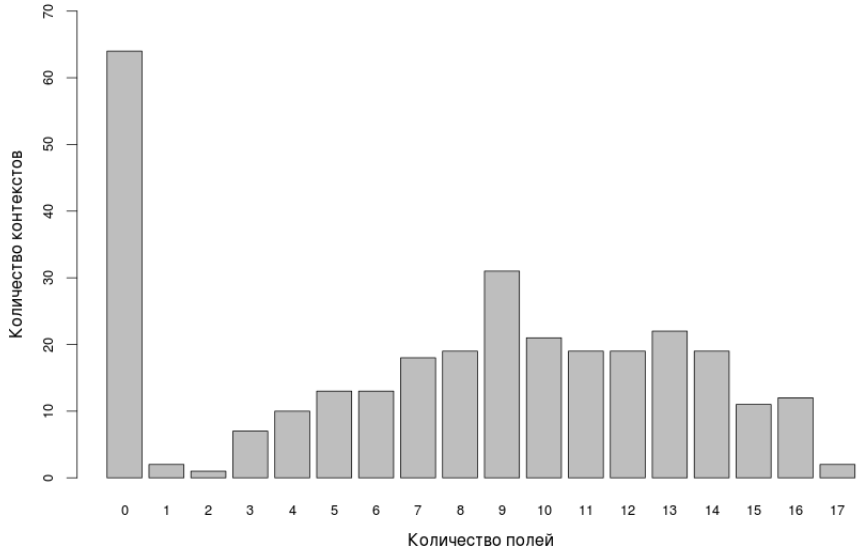

Рисунок 19 — Распределение количества контекстов по количеству связанных полей

тате отсутствия ключевых слов в описании исходного проекта, либо использования низкочастотных ключевых слов, не связанных с семантическими полями в процессе построения модели ($|C| < |\Lambda(C, F)|$). Подобные контексты, не содержащие предметной области в виде связей с семантическими полями, не представляют интерес в данном исследовании. Фактически, при анализе графосемантической модели предметной области рассматриваются только связи между семантическими полями, возникающие при связи контекста с более чем одним семантическим полем, поэтому контексты, связанные с одним полем так же не представляют интереса. Как следствие, из рассматриваемой выборки были исключены контексты, связанные с менее чем двумя семантическими полями, благодаря чему размер выборки сократился до 237 контекстов.

В результате была получена графосемантическая модель Ω_{rfh} и построен семантический граф, представленный на рисунке 20.

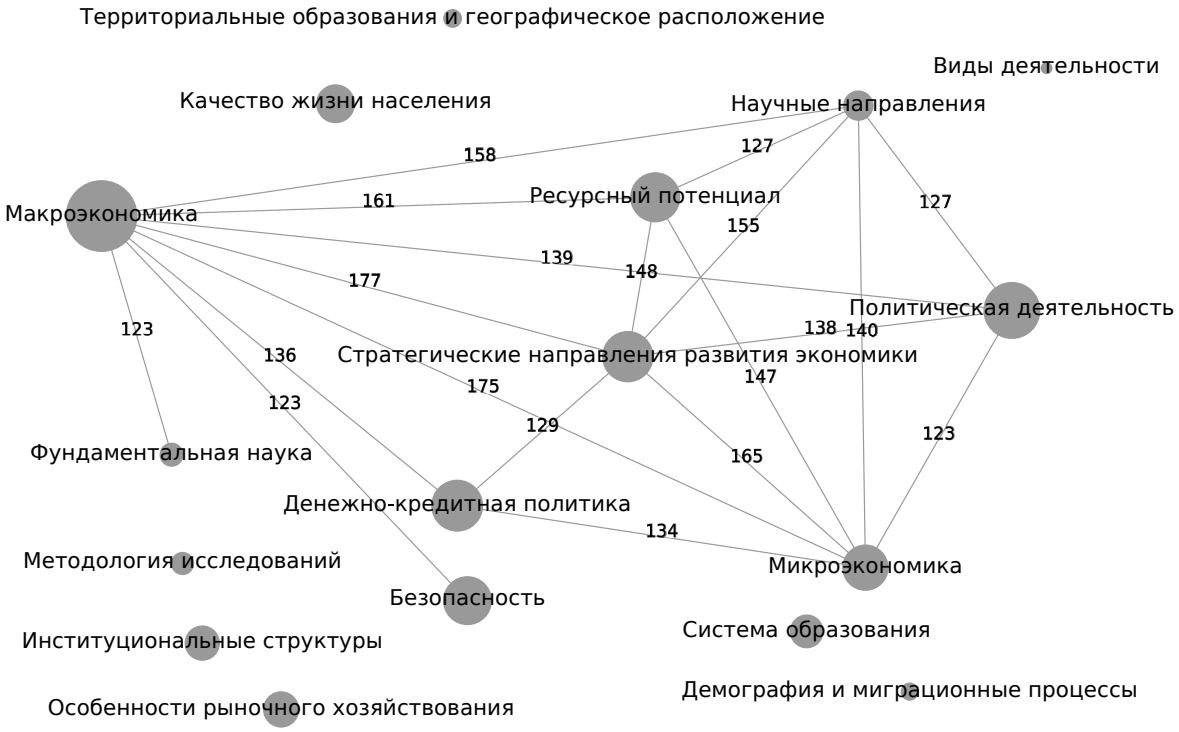

Рисунок 20 — Семантический граф исходной модели предметной области Ω_{rfh}

Для подробного изучения структуры модели предметной области в ней были выделены 5 кластеров (частнонаучных предметных областей), как было описано выше. Полученные нами кластеры определяют частнонаучные предметные области, их описание приведено в таблице 14. Для каждого кластера приведён наиболее репрезентативный проект – проект, единичная предметная область которого наиболее близка к центру кластера (для определения расстояния применяется Евклидова норма). Следует отметить, что в описании каждого кластера представлен один репрезентативный проект, однако их может быть несколько (находящихся на одном расстоянии от центра кластера), но их единичные предметные области совпадают.

Частнонаучные предметные области фактически описывают наиболее актуальные направления современной науки в рассматриваемой области знаний. При этом выбор количества кластеров позволяет регулировать детализацию модели. В рассмотренном случае представлен достаточно низкий уровень детализации, о чем свидетельствует высокая мощность наборов полей, составляющих центры полученных кластеров. Следует отметить кластер № 5, в центре которого представлены все доступные семантические поля. Данный кластер является устойчивым (сохраняется при изменении параметра m и ко-

Таблица 14 — Описание частнонаучных предметных областей

№	Наборы полей	Наиболее репрезентативные проекты
1	7, 9, 13, 15	Разработка принципов функционирования региональной инновационной системы
2	3, 5, 6, 7, 9, 10, 12, 15, 16	Разработка методологии моделирования процессов преодоления социодемографического кризиса в России
3	7, 8, 9, 10, 13, 14, 15, 17	Методология формирования экономики знаний
4	1, 4, 5, 7, 8, 9, 10, 11, 12, 13, 15, 17	Системный анализ стратегий устойчивого развития на примере Бурятской части Байкальского региона
5	1-17	Технологии мониторинга и прогнозирования приоритетных направлений развития региональной инновационной и кластерной политики в лесном хозяйстве

личества кластеров) и включает контексты, не содержащие выраженных признаков прочих кластеров.

Значительный интерес представляет изменение предметных областей во времени. Для изучения этого процесса частнонаучные предметные области были визуализированы в виде графика, представленного на рисунке 21. На этом графике отображены кумулятивные сумму количества статей в каждом из рассматриваемых направлений (частнонаучных предметных областей) за рассматриваемый период (2010-2013 гг.).

Следующим этапом исследования было прогнозирование состояния изучаемой предметной области (исследовательской траектории РГНФ) на 2014 год. Для построения прогноза использовалась вышеописанная методика на основе имитационного моделирования.

Для оценки ошибки прогноза был применён ретроспективный анализ. На основе контекстов за 2009-2012 гг. была построена исходная модель Ω^1_{rfh}, используемая в качестве исходной для прогноза состояния научной предметной области на 2013 г. Реальное количество контекстов в 2013 году составляет

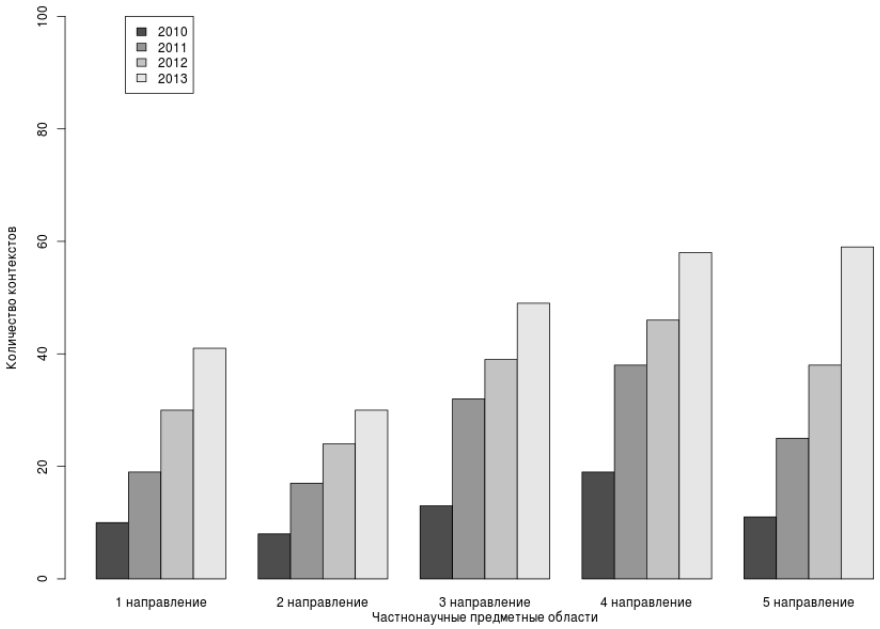

Рисунок 21 — Изменение мощности частнонаучных предметных областей

60, а среднее арифметическое за 2009-2013 гг. составляет 59.25:

$$N = \frac{1}{4} \sum_{i=2009}^{2013} n_i = 59.25,$$

где n_i – количество контекстов в i-ом году, N – количество прогнозируемых контекстов.

Количество прогнозируемых контекстов N было решено округлить до 60. Полученные в результате прогнозирования единичные предметные области были распределены по ранее выделенным частнонаучным предметным областям, результат приведён на рисунке 22.

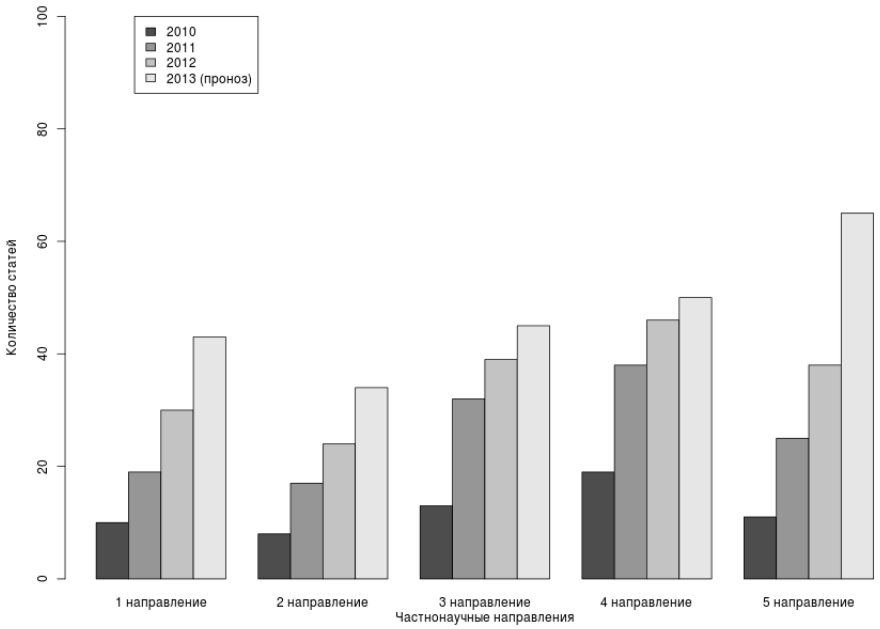

Рисунок 22 — Изменение мощности частнонаучных предметных областей

В качестве оценки ошибки прогноза использовалась распространённая мера MAPE (Mean Absolute Percentage Error, средняя абсолютная ошибка в процентах) [45]:

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\frac{\left|Z(t) - \widehat{Z}(t)\right|}{Z(t)}100\%,$$

где $Z(t)$ – реальное значение в момент времени t, $\widehat{Z}(t)$ – прогнозируемое значение в момент времени t.

Поскольку в рассматриваемой задаче состояние описывается семантической картой, т.е. матричной величиной, в качестве $Z(t)$ используется матричная норма. Обычно матричная норма выбирается как подчинённая ранее используемой векторной норме [63; 93]. Поскольку ранее в данном исследовании применялась Евклидова норма вектора $\|x\|_2$, при решении данной задачи была выбрана подчинённая её спектральная матричная норма $\|A\|_2$:

$$\|x\|_2 = \sqrt{\sum_i x^2},$$

$$\|A\|_2 = \sqrt{\lambda_{max}(A^\dagger A)},$$

где $A^{\dagger}$ – матрица, сопряжённая с A, $\lambda_{max}(A^{\dagger}A)$ – наибольшее собственное число матрицы $A^{\dagger}A$.

При прогнозе 60 контекстов за 2013 год был получен график MAPE, приведённый на рисунке 23.

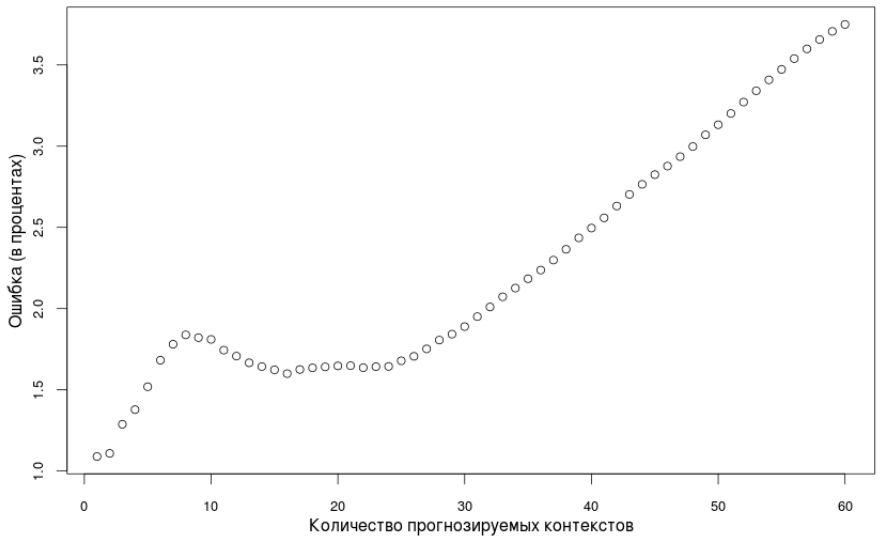

Рисунок 23 — Зависимость MAPE от числа прогнозируемых контекстов

Из рисунка 23 следует что средняя абсолютная ошибка в процентах вычисления семантической карты при прогнозировании исследовательской траектории на год для данной задачи не превышает 4%, следовательно, полученную модель можно считать адекватной.

На основе описанной методики был построен прогноз состояния изучаемой предметной области на 2014 год. В результате прогнозирования была получена графосемантическая модель, семантический граф которой приведён на рисунке 24.

Полученные в результате прогнозирования единичные предметные области были распределены по ранее выделенным частнонаучным предметным областям, результат приведён на рисунке 25.

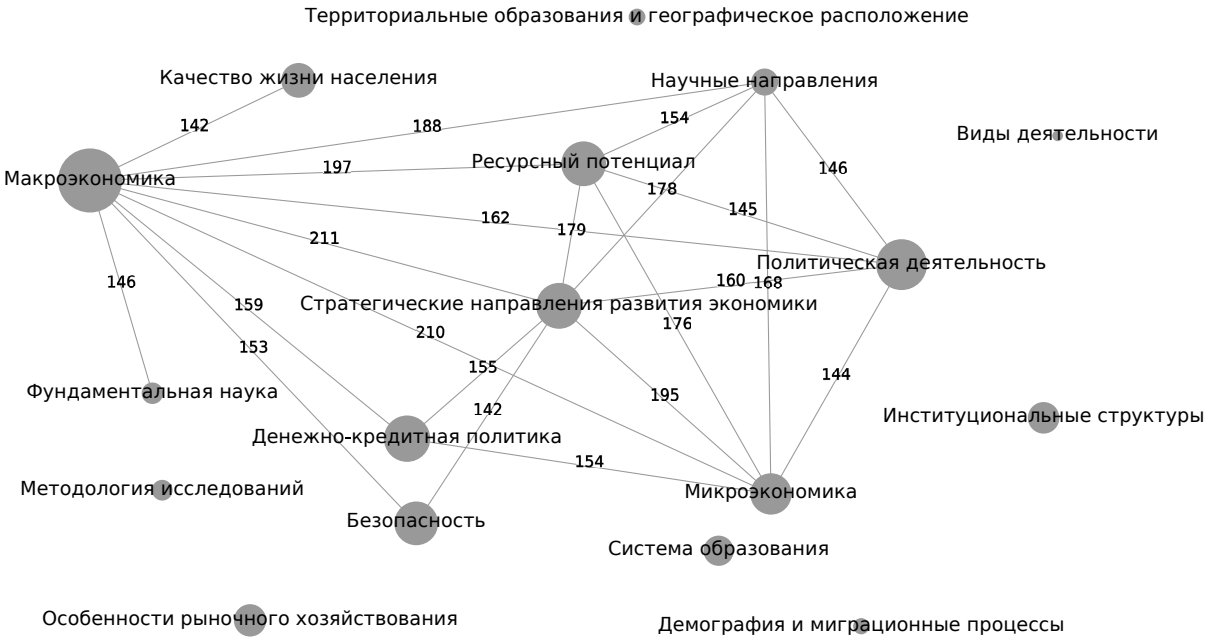

Рисунок 24 — Семантический граф прогнозируемой модели предметной области Ω_{rfh}

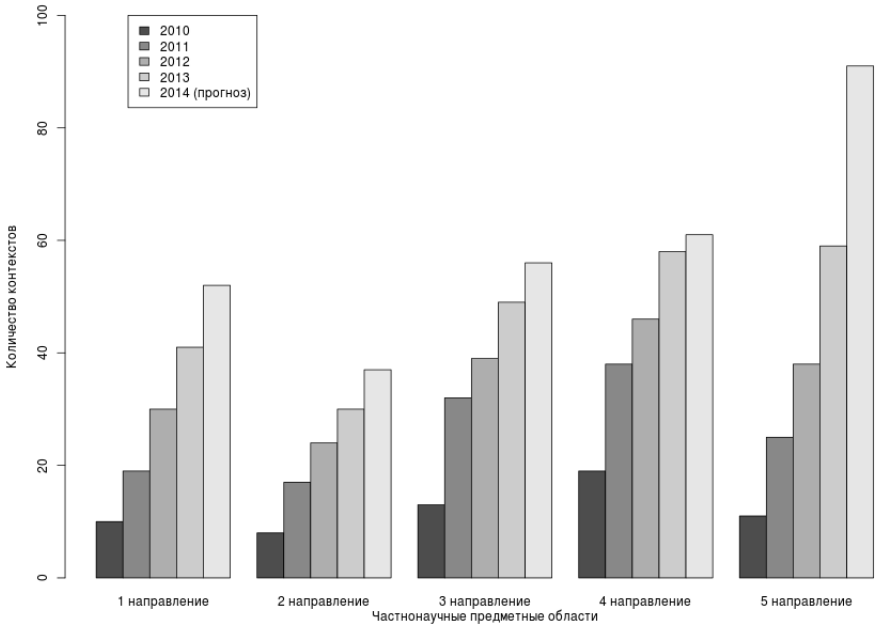

Рисунок 25 — Изменение мощности частнонаучных предметных областей

4.1.3 Оптимальная исследовательская траектория журнала «Вопросов экономики»

Для построения оптимальной исследовательской траектории научного журнала «Вопросов экономики» использовался вышеописанный алгоритм (77-79) с критерием качества управления (76). Экспертом были заданы коэффициенты значимости оценки и штрафов (69-72), значения приведены в таблице

15.

Таблица 15 — Коэффициенты значимости оценки и штрафов для ИТ ВЭ

i	Слагаемое	α_i
0	$\widetilde{I}$	0,9
1	R_1	0
2	R_2	0,03
3	R_3	0,03
4	R_4	0
5	R_5	0,04

Как видно из таблицы 15, соответствующий ограничению на отклонение от прогнозируемой траектории R_4 коэффициент α_4 равен 0, следовательно, в данной задаче нет необходимости строить прогноз.

С целью апробации разработанной методики предварительно был применён ретроспективный анализ. Для этого была построена оптимальная исследовательская траектория для журнала «Вопросы экономики» на 2013 год и сопоставлена с реальной исследовательской траекторией. Сопоставление производилось на основе оценки 76.

Для определения функции оценки качества предметной области $I(x)$ использовался показатель количества поддержанных РГНФ проектов в данной области. Значения данной функции представлены на рисунке 25. Согласно (73-74) была построена аппроксимирующая функция $\widetilde{I}(x)$ на основе ИНС.

В качестве исходного состояния была использована модель Ω_{ve}, построенная для среза контекстов за 2010-2012 года. Оценка исследовательской траектории за 2013 год составила 28,62672028. Реальная исследовательская траектория за 2013 состоит из 83 состояний, поэтому и количество состояний в искомой траектории T было установлено в это значение.

Для построения оптимальной исследовательской траектории использовались следующие параметры генетического алгоритма (77-79):

- размер популяции: 200 особей;
- максимальное количество итераций алгоритма: 50;
- допустимое максимальное количество итераций без улучшения результата: 15;

- вероятность скрещивания двух особей: 80%;
- вероятность мутации отдельной особи: 10%;
- доля наиболее приспособленных особей, отбираемых в следующую популяцию: 5%.
- функция скрещивания: случайное выбор отдельных хромосом из каждой особи-предка.

На рисунке 26 приведён график изменения целевой функции (функции приспособленности лучшей особи популяции) в процессе работы генетического алгоритма.

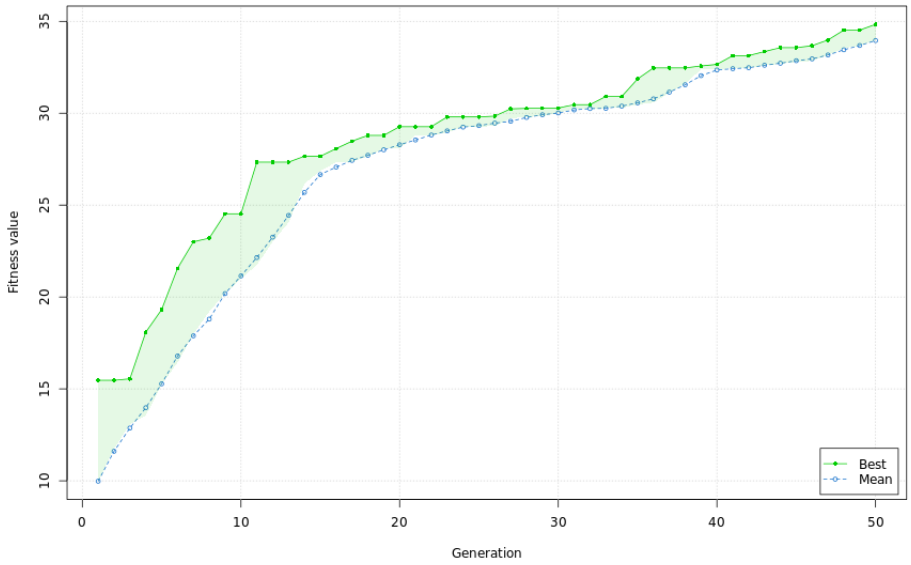

Рисунок 26 — График среднего и лучшего значений целевой функции популяции

Семантический граф последнего состояния полученной оптимальной исследовательской траектории журнала «Вопросы экономики» в 2013 году приведён на рисунке 27.

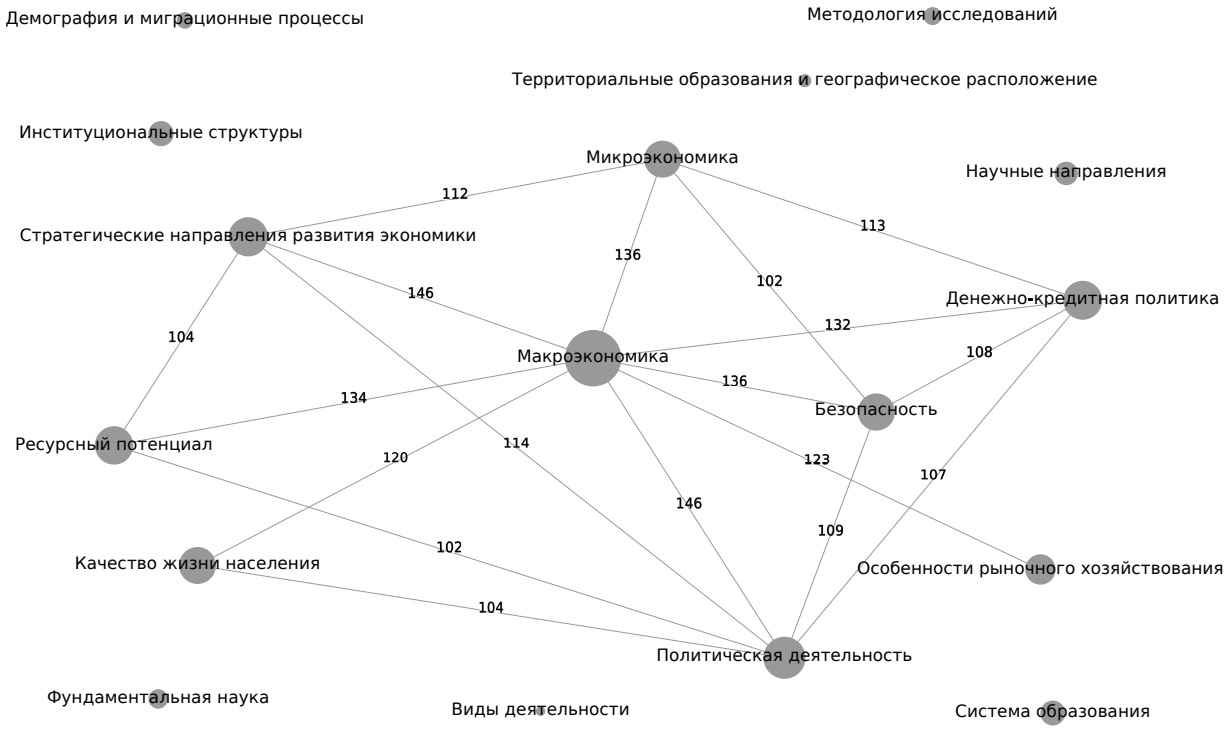

Рисунок 27 — Результирующий семантический граф оптимальной ИТ для Ω_{ve} в 2013 году

Оценка полученной траектории равна 34,83596204, что на 21,7% превышает оценку реальной траектории. Другими словами, следуя данной траектории в 2013, научный журнал мог бы достичь на 21,7% более высокой результативности согласно выбранной оценке.

Далее, была построена оптимальная исследовательская траектория научного журнала «Вопросы экономики» для 2014 года. Для генетического алгоритма (77-79) использовались следующие параметры:

- размер популяции: 10 особей;
- максимальное количество итераций алгоритма: 50;
- допустимое максимальное количество итераций без улучшения результата: 20;

Количество состояний в искомой исследовательской траектории T было установлено равным 85, как среднему значению данной величины в 2010-2013 годах.

Остальные параметры аналогичны предыдущему случаю. На рисунке 28 приведён график изменения целевой функции (функции приспособленности лучшей особи популяции) в процессе работы генетического алгоритма.

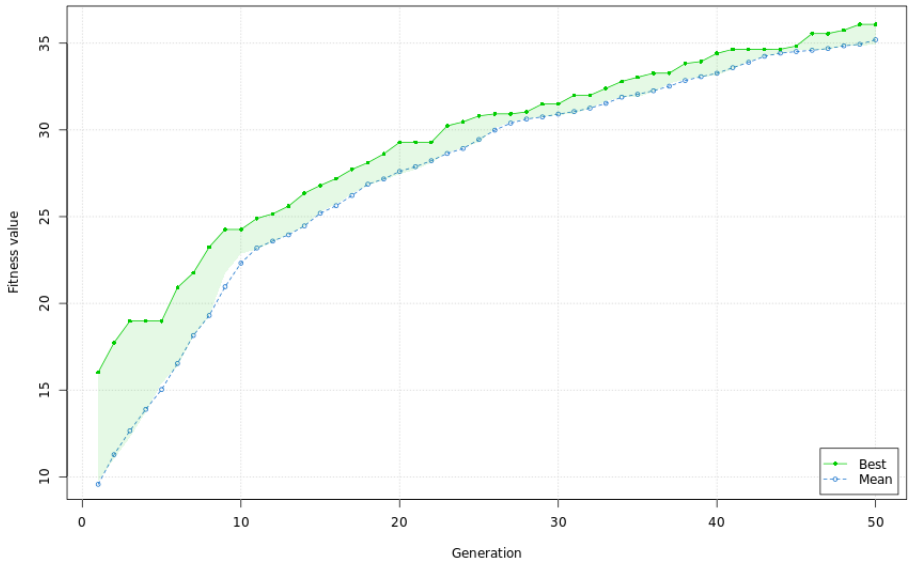

Рисунок 28 — График среднего и лучшего значений целевой функции популяции

Семантический граф последнего состояния полученной оптимальной исследовательской траектории научного журнала в 2014 году приведён на рисунке 29.

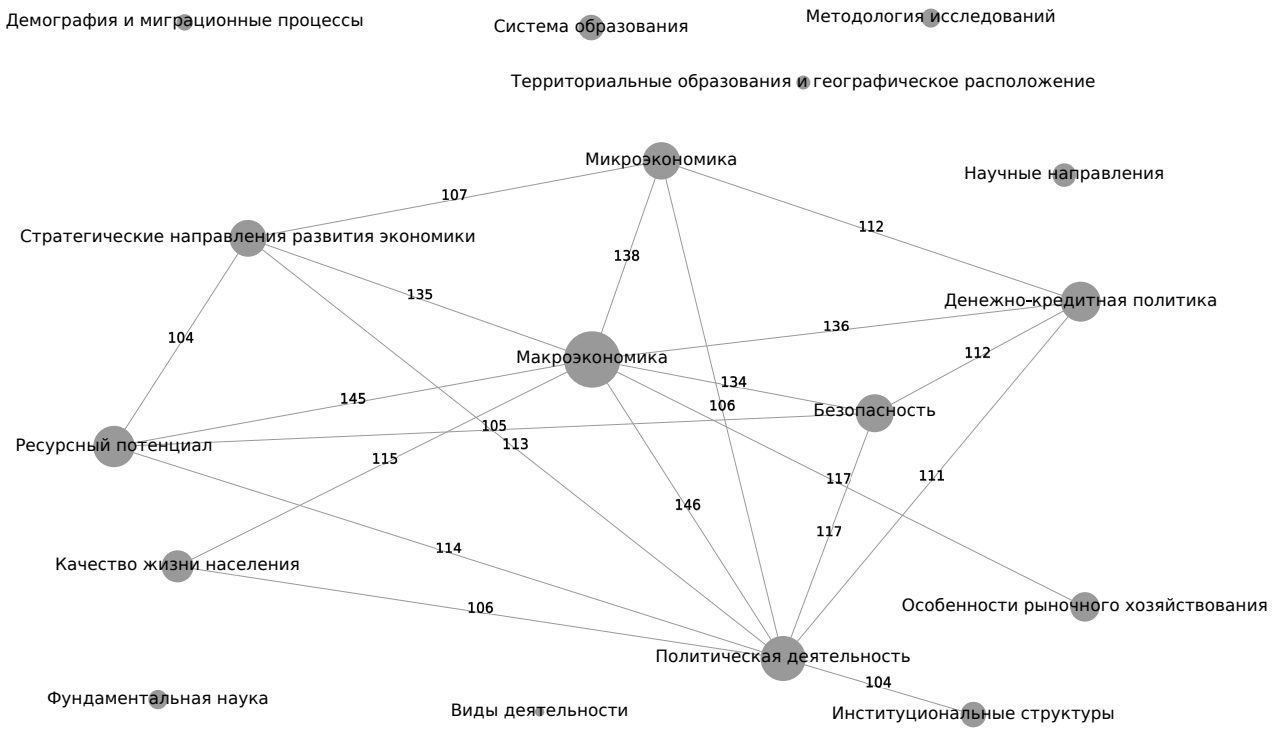

Рисунок 29 — Результирующий семантический граф оптимальной ИТ для Ω_{ve} в 2014 году

Оценка полученной траектории равна 36,08399301, что на 26,05% пре-

вышает аналогичный показатель реальной траектории за предыдущий год.

Следует заметить, что в обоих случаях, генетический алгоритм завершился по достижению максимального количества итераций, т.е. теоретические результат может быть улучшен при увеличении этого значения ценой существенного повышения временных затрат.

4.2 Построение оптимальной исследовательской траектории для научного коллектива

4.2.1 Графосемантическая модель научного коллектива

Не смотря на универсальность разработанных моделей, методик и алгоритмов, научный коллектив, как агент научного производства, является наиболее вероятным потребителем полученных результатов. Следовательно, важна апробация разработанных моделей, методик и алгоритмов применительно к реальному действующему научному коллективу. В данном исследовании описан процесс построения оптимальной исследовательской траектории на примере научной деятельности коллектива лаборатории прикладных и экспериментальных лингвистических исследований Пермской социопсихолингвистической школы за пятилетний период 2010-2014 гг. Границы научного коллектива устанавливались на основе принадлежности исследователя к коллективу лаборатории прикладных и экспериментальных лингвистических исследований Пермского государственного национального исследовательского университета (ПГНИУ).

В исходную выборку вошли 68 работ, опубликованных в период 2010-2014 гг. На основе данной выборки была построена графосемантическая модель Ω_{NK}. Как и в случае с научным журналом, описанном выше, множество компонентов было сформировано на основе ключевых слов, поля и связи поле-компонент определялись экспертами. Семантические поля, выделенные экспертами в ходе построения графосемантической модели Ω_{NK}, приведены в таблице 16 вместе с их частотностью (количеством контекстов, в которых присутствует данное поле).

Таблица 16 — Семантические поля графосемантической модели Ω_{NK}

№	Поле	Частотность
1	Социально-территориальная дифференциация языка (С-ТДЯ)	31
2	Социальные факторы	27
3	Социокультурные аспекты	26
4	Лексика	24
5	Структуры	24
6	Методы исследования	23
7	Национальные языки	20
8	Языковые контакты	19
9	Речь и текст	19
10	Язык	18
11	Семантика	16
12	Когниция	15
13	Ментальный лексикон	14
14	Теория социолингвистики	13
15	Языковые процессы	11
16	Сфера функционирования	11
17	Коммуникация	6
18	Грамматика	3
19	Фонетика	3

Распределение количества контекстов по количеству связанных семантических полей для модели Ω_{NK} приведено на рисунке 30. Как видно из данного рисунка, распределение семантических полей по контекстам отличается от такового для вышеописанной модели Ω_{rfh}, в частности, отсутствуют контексты без привязанных полей. В первую очередь, это обусловлено полностью ручной обработкой данных, что обеспечивает большую точность результирующей модели.

Полученная графосемантическая модель включает:

- 68 контекстов (множество Σ);
- 249 компонентов (множество C);
- 19 полей (множество F);
- 380 связей поле-компонент ($\Lambda(C, F)$).

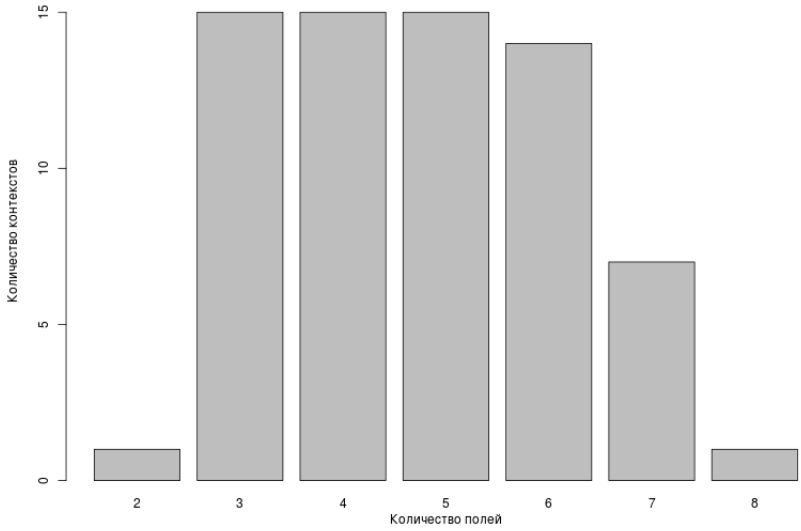

Рисунок 30 — Распределение количества контекстов по количеству связанных полей для модели Ω_{NK}

Семантический граф, построенный для модели Ω_{NK}, приведён на рисунке 31.

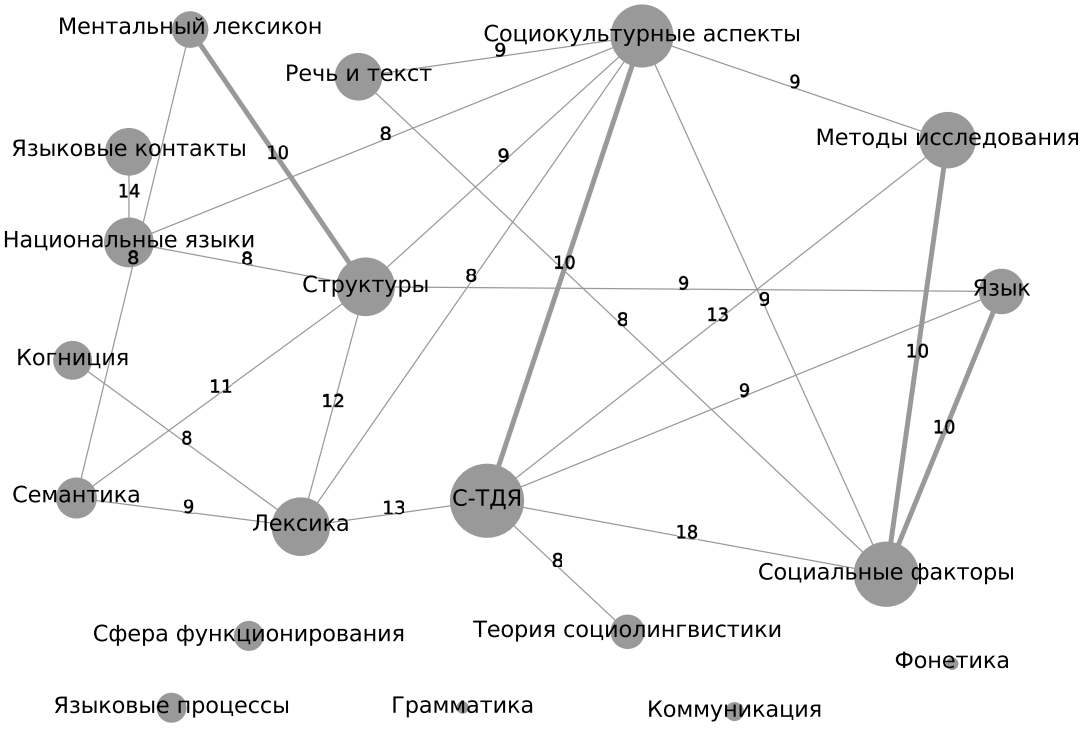

Рисунок 31 — Семантический граф $G_{\Omega_{NK}}$

Как и при исследовании научного журнала, на основе модели Ω_{NK} с помощью кластерного анализа были сформированы частнонаучные предметные области. Так же, как и в случае с научным журналом, в качестве метода кластерного анализа был применён алгоритм нечёткой кластеризации C-means

Таблица 17 — Описание частнонаучных предметных областей научного коллектива

№	Наборы полей	Наиболее репрезентативные статьи
1	1, 2, 3, 4	Словари городских субкультурных образований
2	6, 12, 14	Пермская школа социолингвистики: итоги работы и перспективы развития
3	1, 2, 3, 6, 9, 10	Социолингвистический подход при интерпретации речевой продукции говорящего
4	5, 7, 8, 9	Особенности структуры устных спонтанных текстов монолингвов и билингвов (на материале монологов русских, татар и коми-пермяков)
5	4, 5, 9, 10, 11, 12, 13	Фрейм «общественные отношения» в ментальном лексиконе носителей комипермяцкого, татарского и русского языков
6	2, 8, 16	Оценка доступа к научной информации для академических пользователей в интернете

(39-41). В результате анализа результатов, полученных с различными значениями параметров m (нечёткий параметр кластеризации) и C (количество кластеров, частнонаучных предметных областей), экспертами был выбран вариант, оптимально описывающий предметную область, в которой работает рассматриваемый научный коллектив. Этот вариант включает 6 частнонаучных предметных областей, а параметр $m = 2$. Полученные частнонаучные предметные области приведены в таблице 17.

4.2.2 Оценка качества предметной области

При работе с научным коллективом особую особую важность приобретают различного рода экспертные оценки, поскольку эксперты могут быть выбраны из состава научного коллектива. Как правило, такие эксперты хорошо представляют основные направления научной деятельности научного коллектива и осведомлены о состоянии исследований в этих и смежных направлениях. Поэтому в большинстве случаев при построении ИТ для научного коллектива наиболее очевидным выбором является формирование оценки

качества ПрО с привлечением экспертов из состава самого коллектива.

Однако, остаются актуальными озвученные выше проблемы, в частности, высокая трудоёмкость оценки множества предметных областей отдельными экспертами с последующим свёрткой и аппроксимацией полученной функции оценки $I(x)$. Кроме того, использование в качестве исходного материала лишь предыдущих работ исследуемого научного коллектива может привести к излишней субъективности экспертных оценок и «замыкании» исследовательской траектории в рамках предыдущего опыта научного коллектива.

«Замыкание» происходит вследствие отсутствия в исходной выборке работ из смежных предметных областей (и не только, если рассматривать возможность появления междисциплинарных предметных областей), в результате данным предметным областям будет присвоена нулевая оценка, что сведёт вероятность их появления в результирующей ИТ к минимуму. Фактически, в данном случае задача сводится к экстраполяции функции $I(x)$. В качестве простейшего решения данной проблемы может быть предложено введение коэффициента дисконтирования, с помощью которого некоторая оценка будет равномерна распределена между возможными предметными областями, не представленными в работах из исходной выборки. К недостаткам такого решения относится высокая неопределённость результата – любая из этих предметных областей может быть выбрана с одинаковой вероятностью.

Ещё одно решение указанной проблемы описано в параграфе, посвящённым построению ИТ для научного журнала, оно подразумевает построение макромодели предметной области на основе работ, близких к исходной модели по некоторым параметрам. Это могут быть работы, опубликованные в тех же журналах и сборниках, что и работы исходной выборки, работы из одной дисциплины или рубрики, работы близкие по тематике и т.д. Так, при построении ИТ для научного журнала «Вопросы экономики» макромодель была построена на основе проектов, поддержанных РГНФ в области знаний «02 - Экономические науки». Главным недостатком такого решения является высокая трудоёмкость сбора исходного материала для построения макромодели предметной области и последующая оценка отдельных работ (проблема, упоминавшаяся ранее). В качестве решения, в значительной мере лишённого недостатков вышеописанных способов, был предложен следующий подход.

В основе предлагаемого подхода лежит оценка предварительно выделенных частнонаучных предметных областей агента научного производства. Очевидно, задача оценки частнонаучных предметных областей менее трудоёмка по сравнению с вышеописанной задачей, поскольку количество частнонаучных предметных областей, как правило, значительно меньше мощности множества X (73). При этом так же пропадает необходимость в аппроксимации функции оценки, т.к. она может быть определена экспертами (или другим способом) для каждой частнонаучной предметной области.

Пусть $\widehat{I}(a)$ – оценка частнонаучной предметной области a, тогда оценка единичной предметной области x может быть получена как взвешенная сумма оценок частнонаучных предметных областей:

$$I(x) = \sum_{j=1}^{C} r_j(x)\widehat{I}(a_j) \tag{80}$$

где C – количество частнонаучных предметных областей, $\widehat{I}(a_j), j = \overline{1, C}$ – оценка j-ой частнонаучной предметной области, $r_j(x)$ – степень принадлежности единичной оцениваемой предметной области x к j-ой частнонаучной предметной области.

К преимуществам данного подхода относится не только меньшая трудоёмкость, но и отсутствие «замыкания» исследовательской траектории, поскольку для любой единичной предметной области x может быть вычислена величина $r_j(x)$ и, следовательно, оценка (80).

Для получения оценки $\widehat{I}(a)$ может быть использован один из методов принятия решений. В данном исследовании использовался метод коллективной экспертной оценки частнонаучных предметных областей по нескольким критериям с непосредственным оцениванием. При таком подходе m экспертов оценивают n объектов по l критериям по некоторой шкале. Для получения групповой оценки (ранга) каждой частнонаучной предметной области может быть вычислено среднее взвешенное значение экспертных оценок:

$$\widehat{I}(y_i) = \sum_{k=1}^{l}\sum_{j=1}^{m} \alpha_k \beta_j y_{ij}^k, i = \overline{1, n} \tag{81}$$

где y_{ij}^k – непосредственная оценка i-ой частнонаучной предметной области j-ым экспертом по k-му критерию, α_k – коэффициент относительной важности k-го критерия, β_j – коэффициент компетентности j–го эксперта, причём:

$$\sum_{k=1}^{l} \alpha_k = 1, \sum_{j=1}^{m} \beta_j = 1.$$

Для получения коэффициентов относительной важности критериев так же может быть применена непосредственная оценка. Пусть $t_i, i = \overline{1,l}$ – оценка i–го критерия (в данном случае подразумевается оценка одним экспертов, например руководителем научного коллектива), тогда коэффициенты относительной важности критериев могут быть получены следующим образом:

$$\alpha_i = \frac{t_i}{\sum_{j=1}^{l} t_i}.$$

Экспертами в данном исследовании выступили четверо ($l = 4$) докторов и кандидатов филологических наук, имеющих значительный опыт работы в рассматриваемой предметной области. Эксперты считаются одинаково компетентными в данной предметной области, поэтому коэффициенты компетентности экспертов $\beta_j = \frac{1}{l} = \frac{1}{4} = 0,25$, $j = \overline{1,m}$. В ходе исследования экспертами были выделены 5 критериев ($m = 5$) оценки частнонаучных предметных областей:

1. Сложность дизайна исследования – критерий, обозначающий уровень сложности общего плана организации и проведения исследования на всех его этапах.

2. Сложность сбора данных – критерий, определяющий сложность получения первичного лингвистического материала для осуществления исследования. Большую сложность сбора данных имеют полевые исследования живой речи; меньшую - исследования, основанные на работе с материалами, представленными в электронном формате.

3. Сложность обработки данных – критерий, определяющий сложность

анализа первичных данных, построение на их основе языковых и социокультурных моделей.

4. Сложность интерпретации – критерий, связанный с профессиональными компетенциями, позволяющими рассматривать полученные в ходе исследования языковые и социокультурные модели в контексте знаниевых форм, характерных для анализируемого фрагмента языковой и социокультурной реальности.

5. Междисциплинарность – критерий, определяющий сложность исследования в аспекте задействованных в нем предметных областей науки и связанных с ними компетенций.

Данные критерии были оценены по пятибалльной шкале и были вычислены коэффициенты их относительной значимости $\alpha_i, i = \overline{1,n}$, результат приведён в таблице 18.

Таблица 18 — Оценки критериев T_i и коэффициенты их относительной важности

i	Критерий T_i	Оценка	α_i
1	Сложность дизайна исследования	3	0,15
2	Сложность сбора данных	4	0,2
3	Сложность обработки данных	5	0,25
4	Сложность интерпретации	3	0,15
5	Междисциплинарность	5	0,25

Так же экспертами была произведена непосредственная оценка частнонаучных предметных областей (приведённых в таблице 17, $n = 6$) по полученным критериям (по пятибалльной шкале). Полученные оценки приведены в таблице 19.

Таблица 19 — Экспертные оценки частнонаучных предметных областей по критериям T_i

Эксперт	№ ЧПрО	T_1	T_2	T_3	T_4	T_5
1	1	4	5	3	4	5
	2	5	3	5	5	5
	3	4	4	4	4	4
	4	4	5	4	4	5
	5	4	5	4	4	5
	6	3	4	4	4	5
2	1	4	5	3	4	5
	2	5	3	5	5	5
	3	4	4	4	4	4
	4	4	5	4	4	5
	5	4	5	4	4	5
	6	3	4	4	4	5
3	1	4	3	5	5	5
	2	5	4	5	4	4
	3	5	5	4	5	4
	4	4	5	4	3	4
	5	5	5	5	4	4
	6	4	3	4	4	4
4	1	4	5	4	4	3
	2	4	3	3	5	3
	3	4	5	4	4	4
	4	4	5	5	4	3
	5	5	5	4	4	4
	6	3	3	4	3	4

По формуле 81 были вычислены средние оценки каждой частнонаучной предметной области, результаты приведены в таблице 20.

Таблица 20 — Средневзвешенные экспертные оценки частнонаучных предметных областей

№ ЧПрО	1	2	3	4	5	6
Оценка $\widehat{I}(y_i)$	4,2	4,2625	4,175	4,2875	4,4625	3,875

Для оценки согласованности экспертов обычно используют коэффициент ранговой корреляции Спирмена либо коэффициент конкордации Кенделла. При этом коэффициент корреляции Спирмена определяется для 2 выборок, тогда как коэффициент конкордации Кенделла может быть определён для про-

извольного количества выборок более 1. Поскольку в данной задаче используются 4 выборки (4 эксперта) по каждому из критериев, предпочтительнее использовать коэффициент конкордации Кенделла.

Коэффициент конкордации Кенделла определяется следующим образом: пусть заданы l $(l \geqslant 2)$, выборок:

$$x_1 = \left(x_1^1, \ldots, x_n^1\right),$$
$$\ldots$$
$$x_l = \left(x_1^l, \ldots, x_n^l\right),$$

тогда коэффициент конкордации Кенделла может быть вычислен следующим способом:

$$W = \frac{12}{l^2(n^3 - n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{l} R_{ij} - \frac{l(n+1)}{2} \right)^2, \tag{82}$$

где $R_{ij} \in \{1, \ldots, n\}$ – ранг элемента x_i^j в j-ой выборке.

Так же коэффициент конкордации Кенделла может быть определён через коэффициент корреляции Спирмена:

$$W = \frac{l-1}{l} \frac{2}{l(l-1)} \sum_{i<j} \rho_{x_i x_j} + \frac{1}{l}, \tag{83}$$

где $\rho_{x_i x_j}$ – коэффициент корреляции Спирмена между i-ой и j-ой выборками:

$$\rho_{x_i x_j} = 1 - \frac{6}{n(n-1)(n+1)} \sum_{k=1}^{n} (R_{ki} - R_{kj})^2. \tag{84}$$

Коэффициенты конкордации, рассчитанные по формулам (83,84) для каждого критерия оценки частнонаучных предметных областей, приведены в таблице 21.

Таблица 21 — Коэффициенты конкордации критериев T_i

i	Критерий T_i	W_i
1	Сложность дизайна исследования	0,954
2	Сложность сбора данных	0,918
3	Сложность обработки данных	0,896
4	Сложность интерпретации	0,946
5	Междисциплинарность	0,85

На основе вычисленных коэффициенты конкордации, согласованность оценки экспертов была признана удовлетворительной для данной задачи.

4.2.3 Оптимальная исследовательская траектория научного коллектива

Как и в задаче с «Вопросами экономики», для построения оптимальной исследовательской траектории научного коллектива использовался вышеописанный алгоритм (77-79) с критерием качества управления (76). Экспертами были заданы коэффициенты значимости оценки и штрафов (69-72), значения приведены в таблице 22.

Таблица 22 — Коэффициенты значимости оценки и штрафов для ИТ НК

i	Слагаемое	α_i
0	$\widetilde{I}$	0,5
1	R_1	0,2
2	R_2	0,1
3	R_3	0,1
4	R_4	0,05
5	R_5	0,05

Построение оптимальной исследовательской траектории научного коллектива было разделено на два этапа. На первом этапе был использован ретроспективный анализ: траектория строилась для 2014 года по данным 2010-2013 годов, затем оценки (76) полученной и реальной траектории были сопоставлены. На втором этапе была построена оптимальная исследовательская траектория для 2015 года по данным 2010-2014 годов.

Предварительно была оценена реальная исследовательская траектория научного коллектива за 2014 год, она составила 26,10715.

Поскольку для критерия (72) необходима прогнозируемая исследовательская траектория научного коллектива, на первом этапе с помощью вышеописанной методики был построен прогноз состояния изучаемой предметной области на 2014 год. Реальная исследовательская траектория научного коллектива в 2014 году включает 14 состояний, поэтому прогноз и оптимальная исследовательская траектории были так же построены для 14 состояний. Результирующий семантический граф для полученного прогноза приведён на рисунке 32.

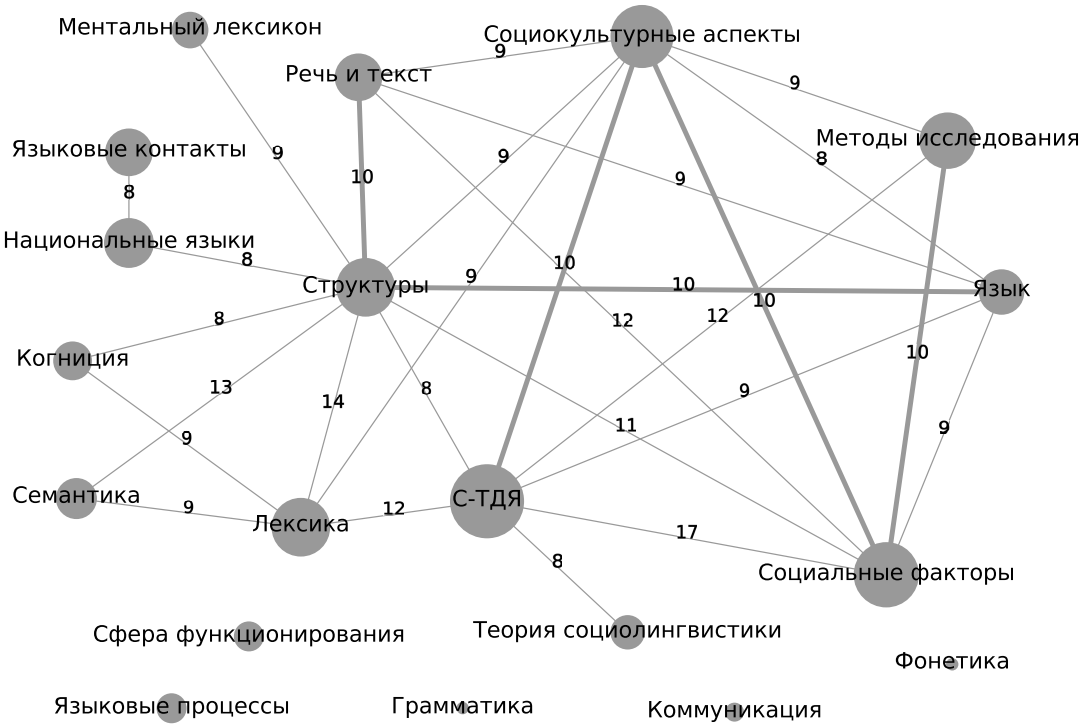

Рисунок 32 — Результирующий семантический граф прогнозируемой ИТ для Ω_{NK} в 2014 году

Оценка полученной прогнозируемой траектории равна 26,16573. Таким образом, согласно оценке (76), прогнозируемая траектория превосходит реальную на 2,24%. Очевидно, это обусловлено штрафом на отклонение от прогнозируемой траектории (72), который в равен нулю во всех состояниях прогнозируемой траектории.

С учётом полученного прогноза была построена оптимальная исследовательская траектория научного коллектива для 2014 года. При этом использовались следующие параметры генетического алгоритма (77-79):

- размер популяции: 200 особей;

- максимальное количество итераций алгоритма: 1000;
- допустимое максимальное количество итераций без улучшения результата: 20;
- вероятность скрещивания двух особей: 80%;
- вероятность мутации отдельной особи: 10%;
- доля наиболее приспособленных особей, отбираемых в следующую популяцию: 5%.
- функция скрещивания: случайное выбор отдельных хромосом из каждой особи-предка.

На рисунке 33 приведён график изменения целевой функции (функции приспособленности лучшей особи популяции) в процессе работы генетического алгоритма. Как видно из графика 33, генетический алгоритм не достигает 250 итерации, это происходит в результате выполнения правила об остановке алгоритма при отсутствии улучшения результата в течении 20 итераций.

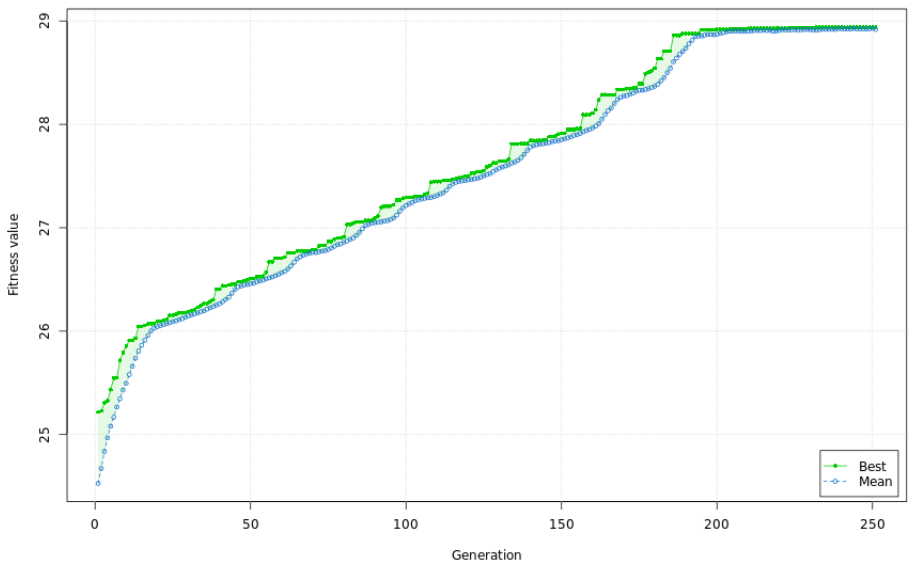

Рисунок 33 — График среднего и лучшего значений целевой функции популяции

Семантический граф последнего состояния полученной оптимальной исследовательской траектории научного коллектива в 2014 году приведён на рисунке 34.

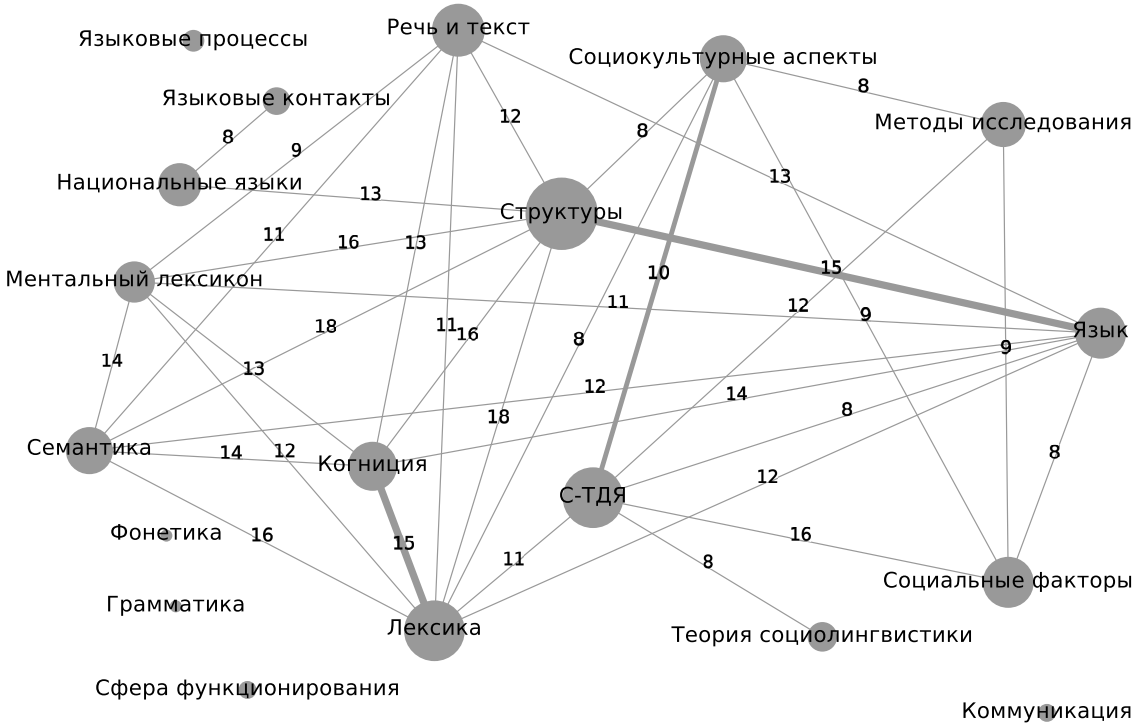

Рисунок 34 — Результирующий семантический граф оптимальной ИТ для Ω_{NK} в 2014 году

Оценка полученной траектории равна 28,9396, что на 10,85% превышает оценку реальной траектории и на 10,6% превышает оценку прогнозируемой траектории.

На втором этапе была построена оптимальная исследовательская траектория научного коллектива для 2015 года. Для этого предварительно был построен прогноз исследовательской траектории из 14 состояний на 2015 год. Семантический граф последнего состояние полученного прогноза приведён на рисунке 35. Оценка полученной прогнозируемой исследовательской траектории равна 28,04838.

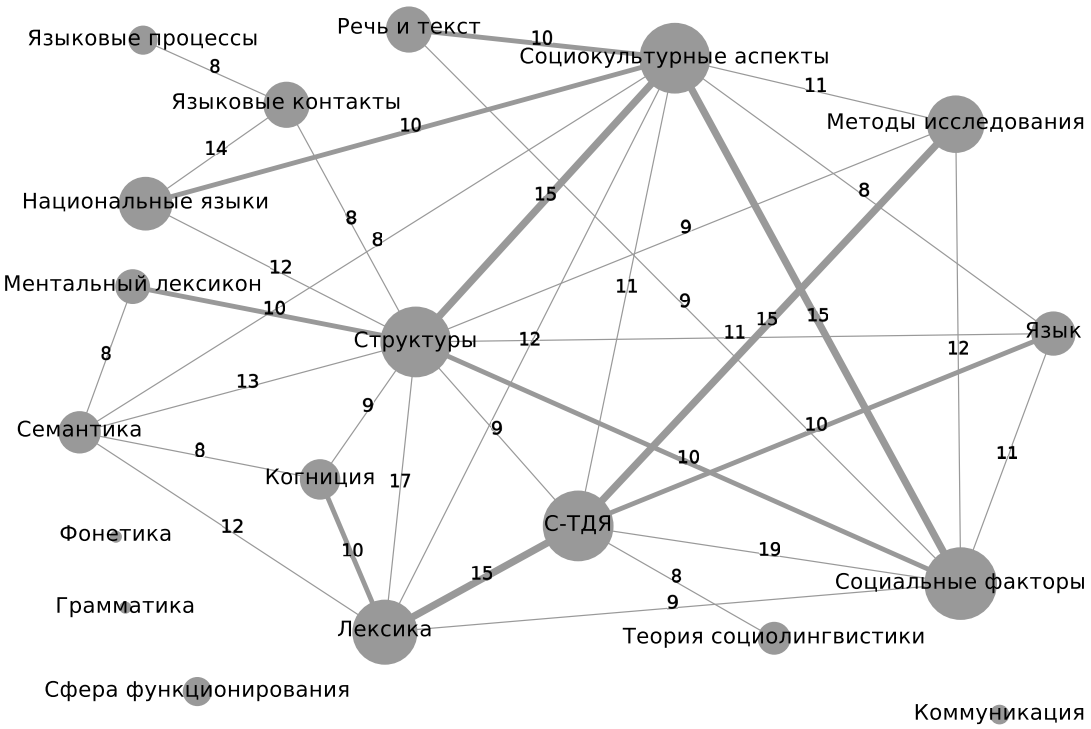

Рисунок 35 — Результирующий семантический граф прогнозируемой ИТ для Ω_{NK} в 2015 году

Для генетического алгоритма (77-79) использовались следующие параметры:

- размер популяции: 200 особей;
- максимальное количество итераций алгоритма: 1000;
- допустимое максимальное количество итераций без улучшения результата: 20;

Остальные параметры аналогичны предыдущему случаю. На рисунке 36 приведён график изменения целевой функции (функции приспособленности лучшей особи популяции) в процессе работы генетического алгоритма.

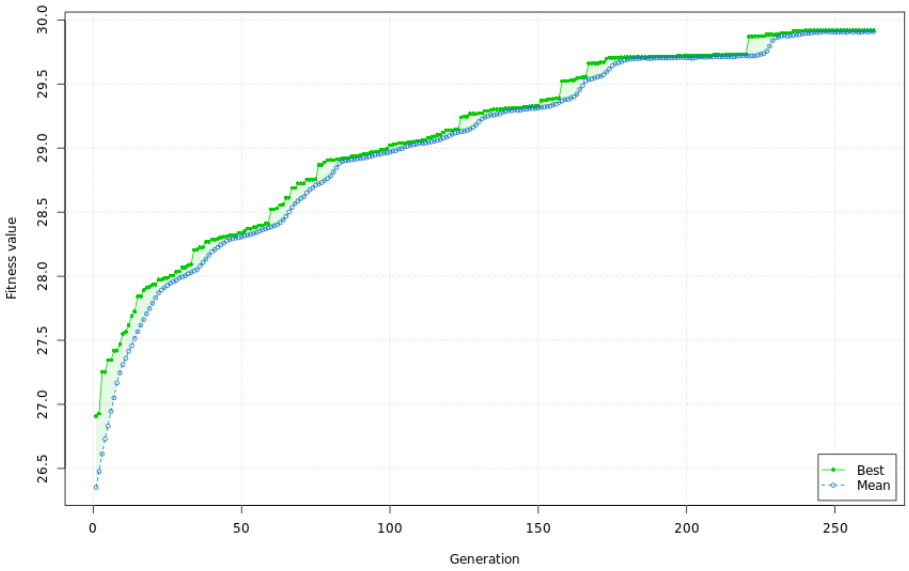

Рисунок 36 — График среднего и лучшего значений целевой функции популяции

Семантический граф последнего состояния полученной оптимальной исследовательской траектории научного коллектива в 2015 году приведён на рисунке 37.

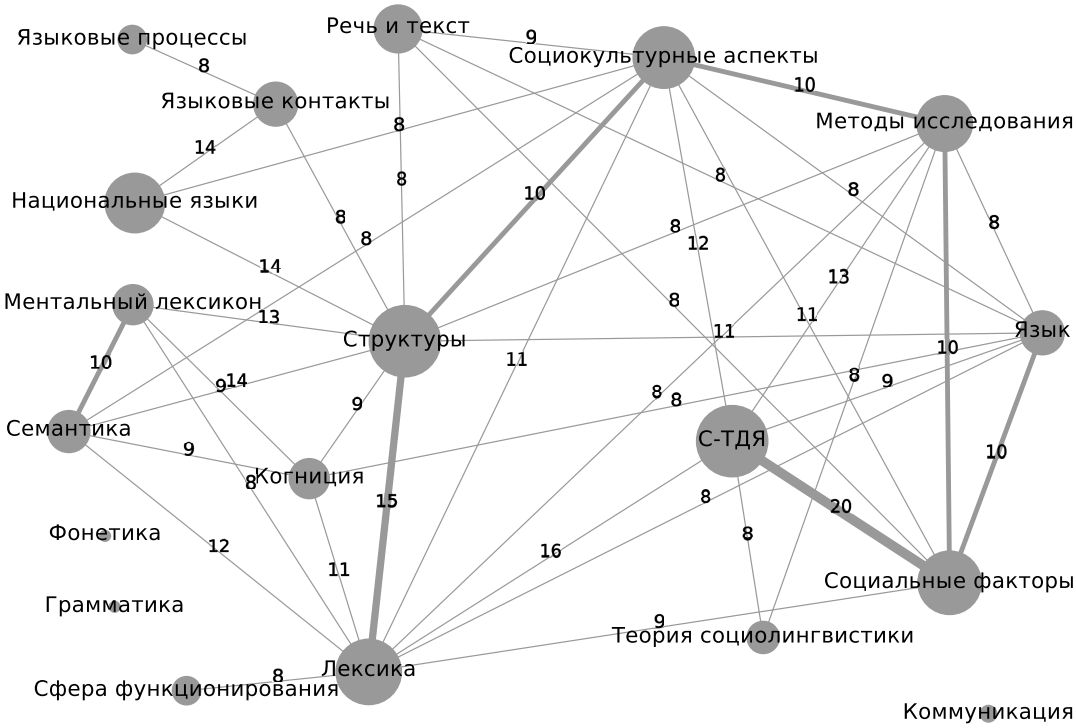

Рисунок 37 — Результирующий семантический граф оптимальной ИТ для Ω_{NK} в 2015 году

Оценка полученной траектории равна 29,91864, что на 14,6% превышает

аналогичный показатель реальной траектории за предыдущий год и на 6,7% превышает оценку прогнозируемой траектории на 2015 год.

4.3 Выводы по главе

1. Оценена погрешность разработанной методики прогнозирования на основе имитационного моделирования. Оценка произведена на основе ретроспективного анализа модели предметной области проектов, поддержанных РГНФ в 2013 году на основе данных за 2009-2012 годы. Погрешность, оценённая посредством метрики MAPE, не превысила 4% в 2013 году (60 контекстов).

2. Построены оптимальные исследовательские траектории для двух различных агентов научного производства: научного журнала «Вопросы экономики» (на 2014 год) и научного коллектива Пермской социопсихолингвистической школы (на 2015 год). Полученные результаты предоставлены агентам научного производства в качестве рекомендаций для повышения собственной эффективности.

3. В ходе решения задач построения оптимальных исследовательских траекторий была показана эффективность разработанной методики: согласно проведённому ретроспективному анализу, оценки полученных оптимальных исследовательских траекторий за последний рассматриваемый год научной деятельности не превышает таковые для реальных на 21,7% и 10,85% для научного журнала и научного коллектива соответственно.

Заключение

В работе поставлена и достигнута актуальная для широкого круга агентов научного производства цель разработки математических моделей, алгоритмического и программного обеспечения для решения задач управления научной деятельностью. В ходе выполнения исследований получены следующие результаты:

1. Проведён анализ существующих подходов, методов и математических моделей, используемых при решении задач управления научной деятельностью. Показана востребованность решений, позволяющих повысить эффективность деятельности агентов научного производства. Описаны недостатки существующих методов и моделей, обоснована необходимость разработки новой методики и сформулированы основные требования к ней. Поставлены цель исследования и задачи, решение которых необходимо для её достижения.

2. Разработана графосемантическая модель предметной области агента научного производства. В данной модели структура предметной области представляется в виде взвешенного неориентированного графа, в вершинах которого находятся семантические поля, объединяющие семантические компоненты, в качестве которых используются ключевые слова, описывающие контексты (научные публикации, проекты и т.д.). При этом состояние предметной области описывается $\frac{n(n-1)}{2}$ значениями, где n – количество полей (уникальные значения семантической карты – матрицы смежности графа).

3. Разработана математическая модель исследовательской траектории агента научного производства, в основе которой лежит графосемантическая модель предметной области и вероятностная графосемантическая модель. На основе полученной модели разработана методика прогнозирования исследовательских траекторий на основе имитационного моделирования. С помощью ретроспективного анализа показано, что при прогнозировании исследовательской траектории для научных проектов,

поддержанных Российским гуманитарным научным фондом в 2014 году, оценка погрешности MAPE не превышает 4%.

4. Разработана методика построения оптимальных исследовательских траекторий. В основе лежит решение дискретной задачи оптимального управления. Разработан численный метод решения дискретной задачи оптимального управления исследовательской траекторией на основе модифицированного метода динамического программирования Беллмана и генетического алгоритма. Кроме того, предложен аддитивный критерий оценки качества управления деятельностью агента научного производства, допускающий гибкую настройку под конкретного агента научного производства посредством задания весовых коэффициентов. Метод реализован на языке программирования R с применением технологии параллельных вычислений OpenCL, что позволило достичь высокой скорости работы. Так, время выполнения алгоритма составляет 2 часа для задачи с 17 полями и исследовательской траектории из 40 шагов и 100 итераций (на центральном процессоре Core i7-3770). Следует отметить, что траектории, получаемые в результате применения разработанной методики, носят рекомендательный характер, поскольку не учитывают всех возможных факторов, влияющих на формирование исследовательской траектории (например социальных и психологических).

5. Разработано программное обеспечение для моделирования, оценки и оптимизации исследовательских траекторий агентов научного производства в составе информационной системы «Семограф» и набора модулей на языке R для реализации трудоёмких численных методов. Разработанная система зарегистрирована в Федеральной службе по интеллектуальной собственности, патентам и товарным знакам.

6. Апробация производилась на двух агентах научного производства: научном журнале «Вопросы экономики» и научном коллективе лаборатории прикладных и экспериментальных лингвистических исследований Пермской социопсихолингвистической школы. В ходе решения задач построения оптимальных исследовательских траекторий была показана эффективность разработанной методики: согласно проведённому ретро-

спективному анализу, оценки полученных оптимальных исследовательских траекторий за последний рассматриваемый год научной деятельности превышают таковые для реальных траекторий на 21,7% и 10,85% для научного журнала и научного коллектива соответственно. Полученные результаты предоставлены агентам научного производства в качестве рекомендаций для повышения собственной эффективности.

Список литературы

1. *Adler R.*, *Ewing J.*, *Taylor P.* Citation Statistics: тех. отч. ; IMU, ICIAM, IMS. — 2008.
2. *Arnold D. N.*, *Fowler K. K.* Nefarious Numbers // Notices of the American Mathematical Society. — 2011. — Т. 58, № 3. — С. 434—437.
3. arXiv. — URL: `http://arxiv.org` (дата обр. 24.08.2014).
4. *Bishop C. M.* Pattern recognition and machine learning. — Springer, 2007. — ISBN 9780387310732.
5. *Burrell Q. L.* A simple model for linked informetric processes // Information Processing and Management. — 1992. — Т. 28. — С. 637—645.
6. *Burrell Q. L.* Hirsch's h-index: A stochastic model // Journal of Informetrics. — 2007. — Т. 1, № 1. — С. 16—25. — ISSN 17511577.
7. *Burrell Q. L.* On the h-index, the size of the Hirsch core and Jin's A-index // Journal of Informetrics. — 2007. — Т. 1, № 2. — С. 170—177. — ISSN 17511577.
8. *Campbell P.* Escape from the impact factor // Ethics in Science and Environmental Politics. — 2008. — Т. 8. — С. 5—7. — ISSN 1863-5415. — DOI: `10.3354/esep00078`.
9. CiteSeerX. — URL: `http://citeseerx.ist.psu.edu` (дата обр. 24.08.2014).
10. *Cohen R.*, *Havlin S.* Complex Networks: Structure, Robustness and Function. — Cambridge University Press, 2010.
11. *Costas R.*, *Bordons M.* Do age and professional rank influence the order of authorship in scientific publications? Some evidence from a micro-level perspective // Scientometrics. — 2011. — Т. 88, № 1. — С. 145—161.
12. *Costas R.*, *Leeuwen T. N. van*, *Bordons M.* A bibliometric classificatory approach for the study and assessment of research performance at the individual level: the effects of age on productivity and impact // Journal of the American Society for Information Science and Technology. — 2010. — Т. 61, № 8. — С. 1564—1581.

13. *Cowell R. G.* Probabilistic networks and expert systems. — Berlin : Springer, 1999.

14. *Dobrescu E. M.*, *Dumitrescu G.*, *Dobre E.-m.* About Hirsch index // Business & Leadership. — 2012. — № 2. — С. 7—14.

15. *Egghe L.* Note on a possible decomposition of the h-Index // Journal of the American Society for Information Science and Technology. — 2013. — Т. 64, № 4. — С. 871.

16. *Egghe L.* The Hirsch index and related impact measures // Annual Review of Information Science and Technology. — 2010. — Т. 44, № 1. — С. 65—114.

17. *Egghe L.* Theory and practise of the g-index // Scientometrics. — 2006. — Т. 69, № 1. — С. 131—152.

18. Elsevier Scirus. — URL: `http : / / www . elsevier . com / government/scirus` (дата обр. 24.08.2014).

19. *Freeman P.* Paper P-321G // R and D Management Research. — 1905.

20. *Gao X.*, *Guan J.* Network model of knowledge diffusion // Scientometrics. — 2012. — Т. 90, № 3. — С. 749—762.

21. *Garfield E.*, *Sher I. H.* New Factors in the Evaluation of Scienti ic Literature through Citation Indexing // American Documentation. — 1963. — Т. 14, № 3. — С. 195—201.

22. Google Scholar. — URL: `http://scholar.google.ru` (дата обр. 24.08.2014).

23. *Hirsch J. E.* An index to quantify an individual's scientific research output // PNAS. — 2005. — Т. 102, № 46. — С. 16569—16572.

24. *Huang M.-h.*, *Chi P.-s.* A Comparative Analysis of the Application of // Journal of Library and Information Studies. — 2010. — Т. 8, № 2. — С. 1—10.

25. InCities. — URL: `http://incites.isiknowledge.com` (дата обр. 25.08.2014).

26. *Jensen F.* An introduction to Bayesian networks. — Berlin : Springer, 1996.

27. *Jin B. H.* An evaluation indicator proposed by scientist // Science Focus. — 2006. — Т. 1, № 1. — С. 8—9.

28. *Klavans R.*, *Boyack K. W.*, *Strategies S.* Using global mapping to create more accurate document-level maps of research fields // Journal of the American Society for Information Science and Technology. — 2010. — Т. 62, № 1. — С. 1—18.

29. *Kosmulski M.* A new Hirsch-type index saves time and works equally well as the original h-index // ISSI Newsletter. — 2006. — Т. 2, № 3. — С. 4—6.

30. *Lawrence P. A.* Lost in publication: how measurement harms science // Ethics in Science and Environmental Politics. — 2008. — Т. 8. — С. 9—11. — ISSN 1863-5415. — DOI: 10.3354/esep00079.

31. *Leeuwen T. N. van* Critical comments on Institute for Scientific Information impact factors: a sample of inorganic molecular chemistry journals // Journal of Information Science. — 1999. — Дек. — Т. 25, № 6. — С. 489—498. — ISSN 0165-5515. — DOI: 10.1177/016555159902500605.

32. Map of Science. — URL: http://www.mapofscience.com/ (дата обр. 02.10.2014).

33. Mapping the backbone of science / K. W. Boyack [и др.] // Scientometrics. — 2005. — Т. 64, № 3. — С. 351—374.

34. Mapping the Semantic Structure of Cognitive Neuroscience / E. Beam [и др.] // Journal of Cognitive Neuroscience. — 2014. — Т. 26, № 9. — С. 1949—1965.

35. MARS. — URL: https://arbicon.ru/projects/MARS (дата обр. 24.08.2014).

36. Measuring knowledge transfer between fields of science / E. J. Rinia [и др.] // Scientometrics. — 2002. — Т. 54, № 3. — С. 347—362.

37. *Newman M. E. J.* The Structure and Function of Complex Networks // SIAM Review. — 2003. — Т. 45, № 2. — С. 167—256. — ISSN 0036-1445. — DOI: 10.1137/S003614450342480.

38. PLOS ONE. — URL: http://www.plosone.org/ (дата обр. 24.09.2014).

39. PubMed. — URL: http://www.ncbi.nlm.nih.gov/pubmed (дата обр. 24.08.2014).

40. *Roebuck K.* Object-Relational Mapping (Orm): High-Impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. — Emereo Pty Limited, 2011.

41. SCImago Journal & Country Rank. — URL: http://www.scimagojr.com (дата обр. 29.09.2014).

42. SciVal. — URL: https://scival.com (дата обр. 25.08.2014).

43. Scopus. — URL: http://www.scopus.com (дата обр. 24.08.2014).

44. *Small H.*, *Boyack K. W.*, *Klavans R.* Identifying emerging topics in science and technology // Research Policy. — 2014. — Т. 43, № 8. — С. 1450—1467. — ISSN 00487333. — DOI: 10.1016/j.respol.2014.02.005.

45. *Swanson D. A.*, *Tayman J.*, *Bryan T. M.* MAPE-R: a rescaled measure of accuracy for cross-sectional forecasts // Journal of Population Research. — 2011. — Т. 28, 2-3. — С. 225—243.

46. The R- and AR-indices: Complementing the h-index / B. Jin [и др.] // Chinese Science Bulletin. — 2007. — № 52. — С. 855—863. — ISSN 1001-6538.

47. Web of Science. — URL: http://thomsonreuters.com/thomson-reuters-web-of-science (дата обр. 24.08.2014).

48. *Webber J.*, *Parastatidis S.*, *Robinson I.* REST in Practice: Hypermedia and Systems Architecture. — O'Reilly Media, 2010.

49. *Zulueta M. A.*, *Bordons M.* A global approach to the study of teams in multidisciplinary research areas through bibliometric indicators // Research Evaluation. — 1999. — Т. 8, № 2. — С. 111—118.

50. *Абдеев Р. Ф.* Философия информационной цивилизации. — М. : ВЛАДОС, 1994.

51. *Авдулов А. Н.*, *Кулькин А. М.* Власть, наука, общество. Система государственной поддержки научно технической деятельности: опыт США. — Ин-т науч. информации по общественным наукам РАН, 1994.

52. *Аллахвердян А. Г.* Науковедение и новые тенденции в развитии российской науки. — М. : ООО "Издательская группа "Логос", 2005.

53. *Аллахвердян А. Г.*, *Семенова Н. Н.*, *Юревич А. В.* Наука в условиях глобализации. — М. : ООО "Издательская группа "Логос", 2009.

54. *Балаян Г. Г.*, *Жарикова Г. Г.*, *Комков Н. И.* Информационно-логические модели научных исследований. — М. : Наука, 1978.

55. *Баранов Д. А.* Математическая формализация метода графосемантического моделирования: Техника и технология: новые перспективы развития ; Материалы VIII Международной научно-практической конференции. — М., 2013. — С. 70—78.

56. *Белоусов К. И.* Теория и методология полиструктурного синтеза текста. — М. : Флинта, 2009.

57. *Блынский Л. Г.*, *Курганов В. Ю.* Моделирование иерархических структур в реляционных базах данных // Приборы и системы. Управление, контроль, диагностика. — 2003. — № 9.

58. *Бородин А. Н.* Элементарный курс теории вероятностей и математической статистики. — СПб. : Лань, 1999. — С. 224.

59. *Бочаров П. П.*, *Печинкин А. В.* Теория вероятностей. Математическая статистика. — 2-е изд. — М. : ФИЗМАТЛИТ, 2005.

60. *Булярская С. А.*, *Булярский С. В.*, *Савельева О. Г.* Эффективность деятельности виртуальных научных коллективов // Вестник ОГУ. — 2009. — Т. 9, № 103. — С. 36—38.

61. *Булярская С. А.*, *Булярский С. В.*, *Синицын А. О.* Формирование виртуальных научных коллективов в виде консорциумов // Вестник ОГУ. — 2009. — № 104. — С. 52—57.

62. *Бурков В. Н.*, *Новиков Д. А.* Как управлять проектами. — М. : Синтег, 1997.

63. *Волков Е. А.* Численные методы. — М. : Наука, 1987.

64. *Володарская Е.*, *Лебедев С.* Управление научной деятельностью (социально-психологические аспекты) // Высшее образование в России. — 2001. — № 1. — С. 85—94.

65. *Герасимова И. Б.* Когнитивная модель структуры личности как участника работы над научным проектом // Вестник уфимского государственного авиационного технического университета. — 2010. — Т. 2, № 37. — С. 228—232.

66. *Гладков М.*, *Шибанов С.* Сложные структуры в реляционных базах данных // Открытые системы. — 2004. — № 02.

67. *Глухов В. В.*, *Коробко С. Б.*, *Маринина Т. В.* Экономика знаний. — СПб. : Питер, 2003.

68. *Гохберг Л. М.*, *Сагиева Г. С.* Российская наука: библиометрические индикаторы // Форсайт. — 2007. — № 1. — С. 44—53.

69. *Григорьев Е.* Представления идентифицируемых сложных объектов в реляционной базе данных // Открытые системы. — 2000. — 01-02.

70. *Гринев М.* Системы управления полуструктурированными данными // Открытые системы. — 1999. — 05-06.

71. *Касьянов В. Н.*, *Евстигнеев В. А.* Графы в программировании: обработка, визуализация и применение. — СПб. : БХВ-Петербург, 2003.

72. *Качанов Ю. Л.*, *Шматко Н. А.* Эффективность управления научно-исследовательским коллективом. Т. 9. — М. : Университет-ская книга, 2010. — ISBN 9785986991351.

73. *Кислякова Ю.* Нарушение прав интеллектуальной собственности и охрана прав интеллектуальной собственности // Ученые записки орловского государственного университета. серия: гуманитарные и социальные науки. — 2009. — № 3. — С. 206—208.

74. Концепции самоорганизации. Становление нового образа научного мышления / А. Печенкин [и др.]. — М. : Академический научно-издательский, производственно-полиграфический и книгораспространительский центр РАН "Издательство "Наука", 1994. — ISBN 5-02-013587-9.

75. *Кормен Т. М.* Часть VI. Алгоритмы для работы с графами // Алгоритмы: построение и анализ = Introduction to Algorithms. — 2-е изд. — М. : Вильямс, 2006.

76. *Коцемир М. Н.* Публикационная активность российских ученых в ведущих мировых журналах // Acta Naturae. — 2012. — Т. 13, № 2. — С. 15—35.

77. *Кульберт М. Я.*, *Сухов Ю. М.* Марковские цепи как отправная точка теории случайных процессов и их приложения. — М. : МЦНМО, 2009. — ISBN 9785940572527.

78. *Леонов Н. И.* Принципы и подходы в управлении научной и инновационной деятельностью (опыт исследовательского университета) // Высшее образование в России. — 2011. — № 11. — С. 19—28.

79. *Матвеев А.*, *Новиков Д.*, *Цветков А.* Модели и методы управления портфелями проектов. — М. : ПМСОФТ, 2005.

80. *Момот А. И.*, *Леньков Р. В.*, *Романкова Л. И.* Концептуальные и методические основы мониторинга научной деятельности по проблемам профессионального образования в системе координационного управления // Научно-исследовательская деятельность в высшей школе: Аналитические обзоры по основным направлениям развития высшего образования. — М. : Научно-Исследовательский Институт Высшего Образования, 1998. — С. 64.

81. Научная электронная библиотека. — URL: `http://elibrary.ru` (дата обр. 24.08.2014).

82. *Новиков Д. А.* Стимулирование в организационных системах. — М. : Синтег, 2003.

83. *Новиков Д. А.*, *Суханов А. Л.* Модели и механизмы управления научными проектами в ВУЗах. — М. : Институт управления образованием РАО, 2005. — С. 80.

84. *Оре О.* Теория графов. — М. : Наука, 1968.

85. *Осовский С.* Нейронные сети для обработки информации. — М. : Финансы и статистика, 2002.

86. *Палей Д.* Моделирование квазиструктурированных данных // Открытые системы. — 2002. — № 09.

87. Приемы объектно-ориентированного проектирования. Паттерны проектирования / Э. Гамма [и др.]. — СПб. : Питер, 2001.

88. Проблемы оценки мирового уровня конкурентоспособности российской науки на примере национальной клинической медицины / В. И. Стародубов [и др.] // Научно-техническая информация. — 2012. — № 8. — С. 139—152.

89. Проекты, поддержанные РГНФ. — URL: `http://www.rfh.ru/index.php/en/grant/isintegr` (дата обр. 08.04.2014).

90. Публикационная активность российской медицинской науки в фокусе актуальной научной политики: оценка достижимости целевых показателей / В. И. Стародубов [и др.] // Вестник РАМН. — 2012. — № 6. — С. 27—35.

91. *Пчелина Ю. С.* Интеллектуальная собственность как основа для инноваций // Сборник научных трудов SWORLD. — 2012. — Т. 23, № 2. — С. 3—8.

92. *Райгородский А. М.* Модели случайных графов. — М. : МЦНМО, 2011.

93. *Самарский А. А.*, *Гулин А. В.* Численные методы. — М. : Наука, 1989.

94. *Седжвик Р.* Фундаментальные алгоритмы на С. Части 1 - 5. Анализ. Структуры данных. Сортировка. Поиск. Алгоритмы на графах. — ДиаСофтЮП, 2003.

95. Семограф. — URL: `http://semograph.com` (дата обр. 06.08.2014).

96. Система графосемантического моделирования / Д. А. Баранов [и др.]. — М. : Свидетельство о государственной регистрации в Федеральной службе по интеллектуальной собственности, патентам и товарным знакам. Зарегистрировано в Реестре программ для ЭВМ № 20111617192 от 15.09.2011.

97. *Таненбаум Э.*, *ван Стеен М.* Распределенные системы. Принципы и парадигмы. — СПб. : Питер, 2003.

98. *Терепищий С. А.* Стратегия инкрементального развития образования и науки // Профессиональное образование в современном мире. — 2014. — Т. 13, № 2. — С. 42—48.

99. *Тулупьев А. Л.*, *Николенко С. И.*, *Сироткин А. В.* Байесовские сети, логико-вероятностный подход. — СПб. : Наука, 2006.

100. *Федотова Е. И.* Особенности государственной политики россии в области вовлечения объектов интеллектуальной собственности в мировой рынок интеллектуальной собственности // Terra Economicus. — 2009. — Т. 7, 1-2. — С. 191—196.

101. *Феллер В.* Введение в теорию вероятностей и ее приложения, пер. с англ. — М., 1967.

102. *Фримен Э.*, *Сьерра К.*, *Бейтс Б.* Паттерны проектирования. — СПб. : Питер, 2011.

103. *Харари Ф.* Теория графов. — М. : Мир, 1973.

104. *Чернозуб С. П.* Образ науки как фактор самоорганизации научного сообщества // Общественные науки и современность. — 2007. — № 6. — С. 140—147.

105. *Штовба С. Д.*, *Штовба Е. В.* Индекс цитирования, учитывающий скрытую диффузию научных знаний // Научно-техническая информация. — 2013. — № 7. — С. 28—31.

Приложение 1

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2011617192

«Система графосемантического моделирования»

Правообладатель(ли): ***Баранов Дмитрий Александрович (RU), Белоусов Константин Игоревич (RU), Влацкая Ирина Валерьевна (RU), Зелянская Наталья Львовна (RU)***

Автор(ы): ***Баранов Дмитрий Александрович, Белоусов Константин Игоревич, Влацкая Ирина Валерьевна, Зелянская Наталья Львовна (RU)***

Заявка № **2011613149**

Дата поступления **3 мая 2011 г.**

Зарегистрировано в Реестре программ для ЭВМ

15 сентября 2011 г.

Руководитель Федеральной службы по интеллектуальной собственности, патентам и товарным знакам

Б.П. Симонов