На правах рукописи

МИЛКОВА МАРИЯ АЛЕКСАНДРОВНА

РАЗРАБОТКА И ЭКСПЕРИМЕНТАЛЬНАЯ АПРОБАЦИЯ МЕТОДИКИ ПОСТРОЕНИЯ ТЕМАТИЧЕСКОЙ МОДЕЛИ ПРИ РАБОТЕ С НАУЧНОЙ ИНФОРМАЦИЕЙ

Специальность 08.00.13 – «Математические и инструментальные методы экономики»

АВТОРЕФЕРАТ

диссертации на соискание учёной степени

кандидата экономических наук

Москва 2021

Работа выполнена в лаборатории экспериментальной экономики Федерального государственного бюджетного учреждения науки Центрального экономико-математического института Российской академии наук

НАУЧНЫЙ РУКОВОДИТЕЛЬ: **Козырев Анатолий Николаевич** доктор экономических наук, кандидат физико-математических наук, научный руководитель лаборатории экспериментальной экономики Федерального государственного бюджетного учреждения науки Центрального экономико-математического института Российской академии наук

ОФИЦИАЛЬНЫЕ ОППОНЕНТЫ: **Лугачев Михаил Иванович**, доктор экономических наук, профессор, заведующий кафедрой экономической информатики Московского государственного университета имени М.В. Ломоносова

Писарева Ольга Михайловна, кандидат экономических наук, доцент, заведующая кафедрой математических методов в экономике и управлении Федерального государственного бюджетного образовательного учреждения высшего образования «Государственный университет управления»

ВЕДУЩАЯ ОРГАНИЗАЦИЯ: Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН)

Защита состоится 20 декабря 2021 г. в 15-00 на заседании диссертационного совета Д 002.013.01, на базе Федерального государственного бюджетного учреждения науки Центральный экономико-математический институт Российской академии наук по адресу: 117418, Москва, Нахимовский проспект, д. 47, ауд. 518, 520.

С диссертацией можно ознакомиться в библиотеке ФГБУН ЦЭМИ РАН и на сайте ФГБУН ЦЭМИ РАН http://www.cemi.rssi.ru.

Сведения о защите и автореферат размещены на сайте Высшей аттестационной комиссии при Министерстве образования и науки Российской Федерации https://vak.minobrnauki.gov.ru/

Автореферат разослан «____» октября 2021 г.

Ученый секретарь диссертационного совета Д 002.013.01,
кандидат экономических наук А.И. Ставчиков

I. ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Как отмечал Герберт Саймон еще 50 лет назад, «информационное перенасыщение с одной стороны рождает дефицит внимания с другой». Сегодня это высказывание особенно актуально, поскольку с развитием цифровых технологий и информация, и внимание все чаще воспринимаются как товар. В эпоху цифровой трансформации темпы накопления информации настолько стремительны, что это порождает сложности ее восприятия. Задача извлечения необходимой информации из больших массивов с целью принятия на ее основе решений все чаще отдается на откуп информационно-коммуникационным технологиям. Здесь речь идет о развитии различных поисковых, рекомендательных, вопросно-ответных и других систем, снижающих затраты экономического агента на поиск информации. Однако важно различать информацию коммерческую и информацию научную, а также информацию, на основе которой предполагается формирование новых знаний. Алгоритмы, работающие с коммерческой информацией (Amazon, Alibaba и др.) нацелены на сбор информации о пользователе и на вывод тех индивидуальных рекомендаций, которые будут соответствовать его профилю и истории совершенных покупок. Алгоритмы, работающие с информацией научной, а также той, которая так или иначе призвана обогащать тезаурус человека, устроены в настоящее время схожим образом. Различные рекомендательные сервисы, в том числе и в электронных библиотеках (Cyberleninka и др.) выдают в рекомендациях схожие по смыслу публикации к уже прочитанным. Поисковые системы (Google, Яндекс) внедряют и совершенствуют систему выдачи готовых ответов.

Стоит отметить также, что человек, как лицо принимающее решение на основе информации, обладает ограниченной рациональностью. Несмотря на развитие сервисов, оказывающих информационные услуги, склонность к когнитивным искажениям, социальное ускорение, клиповое мышление и другие факторы приводят к существенному снижению порога, на котором принимается решение о прекращении поиска информации. Цифровые технологии не только преобразуют данные в информацию, но и ставят перед собой цель преподнесения ее в виде готовых знаний, стремятся исключить стимулы для самостоятельного анализа информации экономическим агентом.

С экономической точки зрения принципиальное значение здесь имеет вопрос о распределении наиболее дефицитного в настоящее время ресурса – внимания, дефицит которого растет вместе с ростом объема информации в новой цифровой реальности. Конкуренция за внимание приводит к развитию как новых платформ для привлечения и управления вниманием (социальные сети, личные блоги, видеоканалы), так и специальных технологий, нацеленных на максимизацию просмотров, рейтингов и т.п. Внимание колеблется, переключается сначала на стиль, а потом через него – к содержанию. В борьбе за внимание пользователей активно используются свойства человеческого

мозга (например, автоматическая конформность, доминирование эмоциональных областей), а также склонность к принятию быстрых решений на основе эвристик.

Актуальной является разработка инструментов, позволяющих оказывать помощь в выборе информации не на основе простоты ее потребления или схожести с потребленной ранее, а на основе ее ценности с точки зрения смысла. Здесь важно отметить, что спрос на информационные продукты, представленные в виде научно-технической информации устроен иначе, чем спрос на коммерческие информационные продукты. Основным отличием является стремление к полноте и ценности извлекаемой информации, вместо обращения внимания исключительно на быструю, популярную или продвигаемую информацию.

Важными в этом смысле являются как разработка самих инструментов, включая модели, алгоритмы и программный код, так и методики их настройки и применения к конкретным областям человеческой деятельности. Разработка соответствующего инструментария имеет прямое отношение к экономической науке, так как, во-первых, он может быть рассмотрен в рамках функционирования экономики информационных продуктов, а именно, представляет собой способ организации потребления информационных продуктов, помощи выбора из множества. Во-вторых, инструментарий может быть применен к анализу экономической литературы (в данной работе это показано на примере такого информационного продукта как научные публикации по поведенческой и экспериментальной экономике). В-третьих, позволяет вычленять смысл из больших текстовых коллекций для принятия экономических решений (в данной работе это показано на примере таких информационных продуктов как патентные документы, среди которых необходимо найти по смыслу только те, которые соответствуют реализации программы импортозамещения).

Степень научной разработанности проблемы

Степень научной разработанности темы рассматривается в рамках математических и инструментальных методов в экономике, принимая во внимание исследования, касающиеся обработки и поиска информации человеком, на стыке таких дисциплин как экономика информации, экономика знаний, поведенческая экономика, нейроэкономика. Собственно инструментальные методы лежат в области анализа текстов, анализа естественного языка, а именно, в области методов семантической компрессии информации (сжатого представления смысла). Одним из них является тематическое моделирование – метод выявления в больших текстовых коллекциях скрытых тем и их характеристик.

Научная разработанность проблемы описана в трёх ключевых направлениях: 1) ценность представления информации в надлежащем виде; 2) личностные препятствия к представлению и восприятию информации и 3) инструменты для извлечения ценной информации.

Вопросы экономической ценности информации, поиска информации как снижения неопределенности, затрат на сбор информации с учетом продолжительности поиска разрабатывались Дж. Стиглером, П. Нельсоном, М. Ротшильдом. Наделение информации определенными свойствами,

неполнота информации, асимметрия, недостатки сбора информации и другие особенности отмечались в работах К. Эрроу, Дж. Акерлофа, Дж. Стиглица и др. Модели принятия решений на основе информации разрабатывались А. Бенерджи (модель стадного поведения), С. Бикчендани (модель информационных каскадов), модели, определяющие оптимальное взаимодействие раскрытия информации и конкуренции (М. Генцков и Е. Каменика), неприятия неоднозначности (Ф. Микарони), рационального невнимания (К. Саймс, Р. Рейс). Проблемы принятия решений на основе информации нашли обширное применение в теории игр (Р. Зельтен, А. Рубинштейн, Дж. Хасаньи и многие другие). В современных исследованиях учитывается различная ценность информации для разных покупателей, изучаются оптимальные механизмы раскрытия информации, продажи информации (К. Шапиро, Х. Вэриан, Д. Бергеманн, М. Бабая, А. Смолин). Среди отечественных ученых, занимающихся экономикой информации, экономикой информационных продуктов следует выделить работы В.Л. Макарова, А.Н. Козырева, М.И. Лугачева, К.Г. Скрипкина, Ю.Е. Хохлова.

Экономические аспекты производства, распространения и управления знаниями изучаются в рамках экономики знаний (Ф. Махлуп), информационной экономики (М. Порат). Среди отечественных ученых – существенный вклад в развитие экономики знаний оказали работы В.Л. Макарова, Г.Б. Клейнера, А.Е. Варшавского.

Ограниченная рациональность человека при обработке информации, психологические аспекты принятия решений, а также особенности принятия решений в условиях неопределенности и риска, в том числе на основе эвристик, изучались Г. Саймоном, У. Эдвардсоном, М. Алле, Д. Эллсбергом, Д. Канеманом и А. Тверски, В. Смитом, Ф. Хайеком, Р. Хайнером, Ю. Эльстером, Г. Гигеренцером, Р. Талером, К. Санстейном и др. Поведенческие аспекты принятия решений применительно к анализу экономических организаций, рынков изучались О. Уильямсоном, Ж. Тиролем, Дж. Акерлофом, Р. Шиллером. Р. Талером, А. Шлейфером и др. Вопросы управления ограниченной рациональностью нашли отражения в теории подталкивания (Р. Талер, К. Санстейн, С. Бенартзи), изучению влияния медиа и социальных сетей (К. Санстейн, С. ДеллаВигна, Х. Эллкот, М. Генцков, Д. Рэнд и др.).

В области нейроэкономики существенный вклад в понимание особенностей принятия решений на основе информации внесли работы П. Глимчера, К. Камерера, Е. Фехра, А. Рустичини и др. Среди российских ученых – В.А. Ключарева.

Изучение внимания к информации как важного звена в принятии экономических решений нашло отражение в направлении, получившем название «экономика внимания». Будучи предложенной Г. Франком, М. Голдхабером, экономика внимания развивалась в работах Т. Давенпорта, Дж. Бека, Т. Ву, Р. Лэнхема, А. Фестре и П. Гарусте, К. Камерера, а также отечественными учеными - А.Н. Козыревым, Г.Г. Почепцовым.

Отечественные ученые, отмечающие важность расширения охвата (или, что то же, обращения внимания) накапливающихся знаний в контексте функционирования экономики знаний – В.Л. Макаров, Г.Б. Клейнер, А.Е. Варшавский; среди отечественных ученых, занимающихся вопросами повышения эффективности работы научного сообщества, стимулирования научно-исследовательской деятельности – Е.В. Балацкий, О.М. Писарева.

Работы, подчеркивающие и развивающие междисциплинарное сотрудничество экономики, психологии, социологии – Г.Б. Клейнер, В.М. Полтерович, Д.А. Жданов; среди поведенческих экономистов – А.В. Белянин, К. Паниди.

Вопросами автоматического понимания текстов занимается одна из наиболее динамично развивающихся областей искусственного интеллекта – анализ естественного языка (NLP). В настоящей работе представляют интерес методы семантической компрессии информации, реализованные в рамках тематического моделирования. В целом, развитию тематического моделирования способствовали работы Т. Хоффмана, Д. Блея. Отечественным ученым К.В. Воронцовым предложен подход к тематическому моделированию на основе аддитивной регуляризации (аддитивная регуляризация тематических моделей – АРТМ). АРТМ предполагает представление коллекции документов в виде матрицы частот слов и нахождение ее неотрицательного низкорангового матричного разложения (матрицы условных вероятностей слов в темах и матрицы условных вероятностей тем в документах), наилучшего по критерию максимума правдоподобия, с добавлением регуляризаторов. Задача решается с помощью ЕМ-алгоритма. АРТМ активно развивается К.В. Воронцовым и его учениками: А. Потапенко, М. Апишевым, А. Яниной, В.Г. Булатовым и др.

В российской науке методы семантического структурирования контента научных электронных библиотек, в том числе с использованием тематического моделирования, рассматривались в работах Паринова С.И., Когаловского М.В. Среди ученых, включающих анализ естественного языка в экономический дискурс, стоит отметить работы С. Бейкера, С.И. Паринова, И.О. Голощаповой, Д.О. Афанасьева, Е.А. Федоровой.

Вклад данной работы состоит как в расширении набора инструментов для анализа текстовой информации, так и в применении их к новым областям знаний. В том числе, расширяется область исследований, которые вводят в экономический дискурс анализ патентной информации, выходящей далеко за рамки патентной классификации. В качестве ключевой темы выбрано «импортозамещение», поскольку именно эта тема весьма актуальна в сегодняшней политической ситуации. Другая интересная с точки зрения применения разработанных инструментов сфера – поведенческая экономика, которая непосредственно связана с экономикой знаний, информационной экономикой, нейроэкономикой и психологией. Кроме того, именно для поведенческой экономики существуют исследования о том, кто и каким образом внес вклад в развитие данного направления (Аннотированная

библиография П. Уаккера[1]). Тем самым появляется возможность оценить результат работы программы, обучившейся без «учителя» с результатом работы квалифицированного специалиста. В том, что касается разработки собственно инструментов, вклад работы – это набор программ и методик для проведения экспериментов, представленных в открытом доступе в репозитории Github. Результаты применения методики выложены на специально созданном сайте.

Целью диссертации является адаптация тематического моделирования для выявления структуры научно-технической информации, способствующей эффективному получению знаний. В рамках работы необходимо разработать методику построения тематической модели (в концепции АРТМ), позволяющую сконфигурировать среду для восприятия научно-технической информации, так чтобы: 1) снизить затраты на поиск ключевой информации по исследуемой научной теме; 2) обеспечить полноту и ценность (с точки зрения смысла) извлекаемой информации 3) минимизировать сдвиги в восприятии (сместить фокус внимания с первой попавшейся, эмоциональной информации).

Задачи исследования:

1. Подготовить методическое обеспечение для непредвзятого анализа корпуса документов (в смысле эффектов искажённого восприятия) через формализацию процесса построения тематической модели в виде создания комплекса программ и рекомендаций для выявления тематической структуры корпуса научных публикаций. А именно, выбрать метаданные публикаций, определить стратегию регуляризации (порядок включения регуляризаторов, сетку для подбора коэффициентов), метрики качества модели, позволяющие получать интерпретируемые результаты моделирования.
2. Экспериментально проверить авторскую методику на примере научных статей по поведенческой и экспериментальной экономике. При проверке выявить ключевые темы данного направления, их ключевые термины, авторов, ключевые публикации.
3. Проанализировать чувствительность полученных результатов.
4. На основе результатов моделирования создать онлайн ресурс, позволяющий осуществлять навигацию по структуре коллекции научных публикаций по поведенческой и экспериментальной экономике.
5. Привести пример решения задачи другого типа, когда необходимо выявить не ключевую информацию внутри всей текстовой коллекции, а определить структуру только интересующего сложносоставного фрагмента. Более конкретно, в коллекции патентных документов выявить и структурировать только те, что соответствуют (по смыслу) пунктам планов по импортозамещению по 22 отраслям промышленности.

[1] Peter P. Wakker. Annotated Bibliography – URL: https://personal.eur.nl/wakker/refs/webrfrncs.pdf (дата обращения: 25.09.2021)

Объект и предмет исследования

Объект исследования – объединение отдельных исследователей, научных коллективов и потребителей научного знания, связанных общей тематикой исследований. *Предмет исследования* – информационное взаимодействие производителей и потребителей научного знания, объединенных единой тематикой исследования.

Научная новизна диссертационной работы заключается в разработке авторской интерпретации алгоритма тематического моделирования и комплекса вычислительных программ на его основе, позволяющих адаптировать применение тематического моделирования (в концепции АРТМ) к экономическим исследованиям теоретического и прикладного характера, в том числе в области экономики знаний, поведенческой экономики, а также при выполнении прикладных экономических исследований.

В частности, научную новизну представляют следующие научные положения, выносимые на защиту:

1. Выявлены особенности потребления информации, на их основе разработан и показан в экспериментах способ организации информационных продуктов, представленных в виде научных публикаций и патентных документов, помогающий в выборе ценных (с точки зрения вычлененного смысла) продуктов. Научно обосновано, что подход к организации информационных продуктов на основе тематического моделирования, в отличие от подхода на основе алгоритмов повсеместно распространенных поисковых систем, позволяет повысить эффективность работы научного сообщества (с точки зрения организации информационного взаимодействия производителей и потребителей научного знания).

2. Предложен формализованный подход к выявлению структуры больших коллекций научных публикаций: разработан комплекс программ, проведены эксперименты, представлен набор инструкций. Предложенная формализация алгоритма АРТМ, в отличие от общего алгоритма АРТМ, позволяет существенно снизить трудоемкость подбора параметров при построении тематической модели на базе научных публикаций.

3. На основе формализованного подхода АРТМ впервые выявлены машинным способом основные темы направления поведенческая и экспериментальная экономика (на базе научных публикаций), в том числе ключевые фразы, основные авторы и опорные публикации по каждой из тем. В сравнении с составляемой вручную аннотированной библиографией по поведенческой экономике, автоматизированный подход также выделяет релевантные темы, может уступать в точности выявленных основных фраз и авторов по темам, однако позволяет быстро определять адекватную структуру научной области там, где экспертное аннотирование отсутствует.

4. Предложен и апробирован подход к оценке результатов программы импортозамещения на основе патентных данных с помощью применения методов тематического моделирования.

Предложенный подход патентного поиска позволяет находить релевантные документы сразу по большому числу пунктов планов импортозамещения по различным отраслям промышленности, что невозможно осуществлять стандартными средствами поисковых и аналитических систем.

Теоретическая значимость исследования заключается в следующем:

1. В рамках проводимого исследования, работа систематизирует результаты не только экономических, физико-математических и технических наук, но и соответствует критериям междисциплинарности: опирается на достижения психологических, социологических, нейроэкономических наук.

2. Разработанный подход организации результатов научной деятельности вносит вклад в экономику информации, экономику информационных продуктов в части расширения формализованного и измеримого описания информационных продуктов, которые могут выступать объектом экономических отношений.

3. Формализация построения тематических моделей является методическим вкладом в развитие подходов к анализу больших объемов текстовой информации, актуальных для принятия решений в экономической науке.

Практическая значимость исследования заключается в следующем:

1. Предложенная формализованная процедура тематического моделирования на основе аддитивной регуляризации представляет собой значимый вклад в практику выделения ключевой информации из больших текстовых коллекций и может быть использована в различных областях экономической науки. Например, при принятии решений о выборе информационных продуктов, обладающих определенной ценностью, а также при проведении прикладных исследований, в том числе при заключении крупных международных контрактов.

2. Предложенная методика работы с научной информацией представляет практическую ценность при проведении исследований – получении представления о структуре направления, выявлении ключевых авторов и опорных публикаций, (внутри поведенческой и экспериментальной экономики). Подход позволяет повысить эффективность работы научного сообщества путем организации информационного взаимодействия производителей и потребителей научного знания.

3. Предложенная методика применения тематического моделирования к анализу программы импортозамещения позволяет получить представление о состоянии всех отраслей экономики в разрезе соответствия плану импортозамещения на основе патентных данных.

Методы исследования

В диссертационном исследовании применялись методы научного познания, как на теоретическом, так и на эмпирическом уровне.

На теоретическом уровне к ключевым методам исследования относились: анализ, моделирование, синтез, обобщение.

На эмпирическом уровне ключевым методами являлись:

1. Информационный поиск (с помощью составления регулярных выражений) для извлечения необходимых для анализа данных.

2. Метод лингвистической предобработки текстовых данных для подготовки их к анализу: лемматизация, фильтрация, удаление стоп-слов.

3. Метод машинного обучения – тематическое моделирование на базе аддитивной регуляризации тематических моделей – для проведения мягкой кластеризации больших текстовых коллекций, выявления скрытых тем, а также ключевых слов и словосочетаний с максимальной содержательной нагрузкой в выявленных темах.

4. Метод визуализации данных – для получения наглядного представления о получаемых результатах.

Информационная база исследования

1. Научные статьи (название и аннотация) по поведенческой и экспериментальной экономике, находящиеся в репозитории Semantic Scholar[2], относящиеся к областям экономики и бизнеса, до 2020 года включительно. Всего 37352 статьи.

2. Патентные документы за период январь 2016 - июнь 2019 гг. из базы Роспатента[3]. Всего 152718 документов; отраслевые планы импортозамещения Минпромторг для 22 отраслей промышленности[4].

Достоверность и апробация результатов работы

Результаты представления научных публикаций в области поведенческой и экспериментальной экономики верифицированы представителями предметной области.

Основные научные положения и результаты исследования были представлены на следующих конференциях и научных семинарах:

1. Милкова М.А. Инструментальное обеспечение поведенческой экономики. Научный семинар Института системного анализа РАН под рук. Лившица В.Н. – Январь 2020 г.

2. Милкова М.А. Инструментальное обеспечение экономики внимания. Научный семинар ЦЭМИ РАН под рук. ак. Макарова В.Л. – Март 2020 г.

3. Milkova M.A. Patent-based import substitution analysis with Additively Regularized Topic Models. X International Scientific and Practical Conference named after A. I. Kitov "Information Technologies and Mathematical Methods in Economics and Management" – October 2020.

---

[2] Semantic Scholar Open Research Corpus // Интернет платформа Semantic Scholar: https://www.semanticscholar.org – URL: http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/download/ (дата обращения 22.04.2020)

[3] Открытые реестры ФИПС // Федеральный институт промышленной собственности: https://www.fips.ru/ – URL: https://www.fips.ru/registers-web/ (дата обращения 22.04.2020)

[4] Отраслевые планы импортозамещения. – Текст: электронный // Государственная информационная система промышленности: https://gisp.gov.ru/ – URL: https://gisp.gov.ru/plan-import-change/ (дата обращения 22.04.2020)

4. Милкова М.А. Экономика внимания и ее инструментальное обеспечение. IV Российский экономический конгресс, тематическая конференция Поведенческая и экспериментальная экономика. Декабрь 2020 г.

5. Милкова М.А. Восприятие информации в период пандемии COVID-19. Международная научная школа-семинар «Системное моделирование социально-экономических процессов» им. С.С. Шаталина – Октябрь 2020 г.

Область исследования. Настоящее исследование соответствует паспорту научной специальности 08.00.13 – «Математические и инструментальные методы экономики» и соответствует требованиями следующих разделов:

2. Инструментальные средства:

2.6. Развитие теоретических основ методологии и инструментария проектирования, разработки и сопровождения информационных систем субъектов экономической деятельности: методы формализованного представления предметной области, программные средства, базы данных, корпоративные хранилища данных, базы знаний, коммуникационные технологии.

2.8. Развитие методов и средств аккумуляции знаний о развитии экономической системы и использование искусственного интеллекта при выработке управленческих решений.

Публикации основных результатов исследования. Основные положения и результаты диссертационного исследования отражены в 13 научных публикациях общим объемом 19,2 п.л. (в т.ч. авт. – 18 п.л.), в том числе в 3 статьях в изданиях, включенных в перечень ВАК РФ, в 1 статье в издании, индексируемом в базе данных Scopus, оставшиеся – в других научных изданиях.

Объем и структура работы

Диссертационная работа содержит введение, три главы, заключение и библиографию общим объемом 172 стр., включая 9 таблиц, 13 рисунков и 7 Приложений.

В *первой главе* систематизированы ключевые аспекты работы с информацией в рамках экономики информации, экономики знаний, поведенческой экономики, нейроэкономики, экономика внимания.

Вторая глава содержит описание подходов к семантической компрессии информации больших текстовых коллекций, позволяющих упорядочить характер работы с информацией. Описана концепция разведочного поиска информации, подход тематического моделирования, в частности – аддитивной регуляризации тематических моделей (АРТМ). Предложена авторская формализация алгоритма.

Третья глава содержит апробацию формализованного алгоритма аддитивной регуляризации для решения задачи выявления структуры коллекции научных публикаций в области поведенческой и экспериментальной экономики. Также на основе АРТМ решена задача поиска патентных

документов, соответствующих по смыслу пунктам двадцати двух отраслевых планов импортозамещения.

II. ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

1. Проведен мультидисциплинарный анализ литературных источников на стыке экономики информации, экономики знаний, поведенческой экономики, нейроэкономики, экономики внимания в разрезе восприятия информации, поиска информации и принятия экономических решений.

Рассмотрены ключевые аспекты работы с информацией в рамках экономики информации и экономики знаний. Информация в экономике информации играет ключевую роль в принятии экономических решений. Выделяют поиск информации как снижение неопределенности; затраты на поиск; определение оптимальной продолжительности поиска; разработку стратегии принятия решений о необходимости дополнительного сбора информации; модели информационных каскадов, стадного поведения; неприятия неоднозначности; рационального невнимания. Информация может быть наделена различными свойствами (асимметричная, неполная, зашумленная). К современным задачам в области экономики информации относятся исследования управления атрибутами и стоимостью информационных продуктов, с учетом различной ценности информации для разных покупателей, изучаются оптимальные механизмы раскрытия информации, продажи информации. Экономика знаний прежде всего исследует экономические аспекты производства, распространения и управления знаниями. Схожим понятием является экономика, основанная на знаниях.

Рассмотрены различные аспекты ограниченной рациональности человека при обработке информации. Показаны особенности принятия решений в условиях неопределенности и риска: подход эвристик и сдвигов, фрейминг, теория перспектив, подход быстрых и экономных эвристик и связанные с этими особенностями разнообразные эффекты. Поведенческие аспекты принятия решений показаны также применительно к анализу экономических организаций, рынков. Отклонения от рациональных стандартов могут использоваться как инструмент экономической политики, направленной на повышение благосостояния: используются меры подталкивания, в том числе цифрового подталкивания. В современных поведенческих исследованиях большое внимание уделено влиянию медиа, социальных сетей и сообществ.

Показаны ключевые характеристики принятия решений на стыке экономических, психологических и нейронаук. Так, нейронаука указывает, что в основе принятия экономических решений лежит взаимодействие между автоматическими и управляемыми процессами, между когнитивными и эмоциональными системами, причем роль «рационального» суждения серьезно переоценивается людьми.

Систематизированы исследования теории внимания в экономических науках, описано понятие экономики внимания. Выделено два направления развития: первое относится к теории информации, второе – к поведенческой экономике. В первом направлении внимание рассматривается как дефицитный экономический ресурс и обсуждаются последствия этого для пользователей информации (конкретизация информации, чтобы привлечь внимание пользователей и повысить их полезность, дискриминация цен, дискриминация продуктов и др.). Отдельно развивается теория рационального невнимания.

С другой стороны, ограниченное внимание – важное понятие поведенческой экономики. Внимание интерпретируется в рамках концепции ограниченной рациональности. Ограниченность внимания, склонность к когнитивным искажениям при восприятии информации, клиповый характер мышления современного человека объясняет актуальность разработки инструментов для эффективного распределения внимания в условиях информационного взрыва.

Инструментальные методы, используемые в данной работе, лежат в области анализа текстов, анализа естественного языка, а именно, в области методов семантической компрессии информации (сжатого представления смысла). Отмечено, что достижения в области анализа естественного языка в основном используются для внедрения в сторонние системы (рекомендательные, вопросно-ответные, поисковые и др.), которые позволяют снижать затраты пользователей на поиск информационных продуктов, способствуют решению проблем, связанных с дефицитом внимания, «подталкивают» к принятию быстрых решений. Однако стоит отметить, что спрос на коммерческие информационные продукты устроен иначе, чем спрос на научно-техническую информацию (к которой относятся научные публикации, патентные документы – они также представляют собой информационные продукты). Основным отличием при работе с научной информацией является стремление к полноте и ценности извлекаемой информации, вместо обращения внимания исключительно на быструю, популярную или продвигаемую информацию.

2. В качестве инструмента, позволяющего оказывать помощь в выборе ценных (с точки зрения вычлененного смысла) информационных продуктов (представленных в виде научных публикаций, патентных документов), предлагается использовать подход аддитивной регуляризации тематических моделей (АРТМ).

Постановка задачи тематического моделирования в концепции АРТМ (предложена К.В. Воронцовым)[5]:

Пусть есть коллекция документов $\mathcal{D} = \{d_1, \dots, d_N\}$, каждый из которых представляет собой последовательность n_d слов $(w_1, \dots, w_{n_d})$ из словаря $\mathcal{W}$. Предполагается, что существует конечное множество тем T, и каждое употребление слова w в каждом документе d связано с некоторой темой

[5] Воронцов, К.В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады академии наук. 2014. 3(456). DOI: 10.7868/S0869565214090096

t ∈ T. Слова w и документы d являются наблюдаемыми переменными, тема t ∈ T является латентной (скрытой) переменной. Задачей тематического моделирования является нахождение скрытых тем в документах, а также слов, характеризующих каждую из тем (в общем случае, в качестве характеристик тем могут выступать не только слова, но и любые другие характеристики). Иначе говоря, построение тематической модели рассматривается как задача одновременной кластеризации документов и слов по одному и тому же множеству кластеров – тем.

Подход аддитивной регуляризации представляет задачу тематического моделирования как задачу нахождения неотрицательного низкорангового матричного разложения заданной матрицы частот слов документах $F = (p_{dw})_{D\times W}$ в произведение двух неотрицательных нормированных матриц меньшего размера:

матрицы слов в темах $\Phi = (\varphi_{wt})_{W\times T}$, $\varphi_{wt} = p(w|t) = \frac{n_{wt}}{n_t}$

матрицы тем в документах $\Theta = (\theta_{td})_{T\times D}$, $\theta_{td} = p(t|d) = \frac{n_{td}}{n_d}$

$$p(w|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td} \qquad (1)$$

Матрицы F, Φ, Θ имеют неотрицательные нормированные столбцы, представляющие дискретные распределения.

Для оценивания параметров Φ, Θ тематической модели (1) по коллекции документов $\mathcal{D}$ максимизируется логарифм правдоподобия выборки при ограничениях неотрицательности и нормированности столбцов матриц Φ, Θ:

$$L(\Phi,\Theta) = ln \prod_{d\in D}\prod_{w\in d} p(w|d)^{n_{dw}} = \sum_{d\in D}\sum_{w\in d} n_{dw} ln \sum_{t\in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

$$\sum_{w\in W} \varphi_{wt} = 1, \quad \varphi_{wt} \geq 0, \qquad (2)$$

$$\sum_{t\in T} \theta_{td} = 1, \quad \theta_{td} \geq 0$$

Задача решается с помощью ЕМ-алгоритма. На первом этапе выбирается начальное приближение для $\varphi_{wt}, \theta_{td}$. На Е-шаге вычисляются вспомогательные переменные p_{tdw} (вероятность, что слово w из документа d принадлежит теме t)

$$p_{tdw} = p(t|d,w) = \frac{n_{tdw}}{n_{dw}}$$

По формуле Байеса вероятность p_{tdw} можно записать как:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}} \qquad (3)$$

На М-шаге вычисляются частотные оценки максимального правдоподобия для искомых условных вероятностей $\varphi_{wt}, \theta_{td}$.

$$\varphi_{wt} = \frac{n_{wt}}{n_t}, \quad n_{wt} = \sum_d n_{tdw} = \sum_d n_{dw} p_{tdw}, \; n_t = \sum_{w\in W} n_{wt} \qquad (4)$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \; n_{td} = \sum_w n_{tdw} = \sum_w n_{dw}\, p_{tdw}, \; n_d = \sum_{t\in T} n_{td} \qquad (5)$$

Вычисления (3)-(5) продолжаются в цикле до сходимости.

Помимо построения модели на основе текста, АРТМ подход позволяет строить так называемую мультимодальную тематическую модель. Под модальностями понимаются метаданные, так или иначе характеризующие тематику текста. К модальностям могут относиться: биграммы (n-граммы), авторы, цитируемые или цитирующие документы и т.п. В мультимодальной модели матрица Φ определяется для каждой модальности, а матрица Θ является общей. Оптимизационная задача представляет собой максимизацию взвешенной суммы лог-правдоподобий при условиях нормировки и неотрицательности столбцов матриц Φ^m, Θ (для каждой модальности задается ее вес).

Ключевым моментом является не единственность матричного разложения $\Phi\Theta$, определяемого с точностью до невырожденного преобразования $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$ (задача является некорректно поставленной). Выбор преобразования S никак не контролируется и зависит от случайного начального приближения. Согласно теории регуляризаци, решение такой задачи возможно доопределить и сделать устойчивым. Для этого к основному критерию добавляют дополнительный критерий — регуляризатор, учитывающий специфические особенности данной задачи и знания предметной области.

Таким образом, наряду с правдоподобием требуется максимизировать r критериев-регуляризаторов $R_i(\Phi, \Theta)$, i = 1,…r с коэффициентами регуляризации τ_i. Таким образом, оптимизационная задача имеет вид:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta} \quad (6)$$

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta),$$

$$\sum_{w \in W} \varphi_{wt} = 1, \quad \varphi_{wt} \geq 0,$$

$$\sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0$$

Решение задачи (6) строится на основе так называемого регуляризованного ЕМ-алгоритма, Е-шаг (7) которого аналогичен (2), а М-шаг заменяется регуляризованными уравнениями (8-9).

На первом этапе выбирается начальное приближение для φ_{wt}, θ_{td}. На Е-шаге вычисляются вспомогательные переменные p_{tdw}:

$$p_{tdw} = \mathrm{norm}_{t \in T}(\varphi_{wt} \theta_{td}), \quad (7)$$

где оператор нормировки norm преобразует произвольный вектор в вектор вероятностей дискретного распределения путем обнуления отрицательных элементов и нормировки.

На М-шаге вычисляются частотные оценки максимального правдоподобия для искомых условных вероятностей φ_{wt}, θ_{td}:

$$\varphi_{wt} = \mathrm{norm}_{w \in W} \left(n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right), \quad n_{wt} = \sum_d p_{tdw} n_{dw} \quad (8)$$

$$\theta_{td} = \mathrm{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_w n_{dw} p_{tdw} \quad (9)$$

Вычисления (7) – (9) продолжаются в цикле до сходимости.

В теории АРТМ предлагается конструировать регуляризаторы на основе дивергенции Кульбака-Лейблера. Вводятся регуляризаторы разреживания (формализует предположение о том, что каждый документ относится к небольшому числу тем; слово характеризует небольшое число тем) – необходимо максимизировать KL-дивергенцию между распределениями φ_t, θ_d и равномерными распределениями; регуляризаторы сглаживания (с целью выделения слов общей лексики) – необходимо минимизировать KL-дивергенцию между фоновыми компонентами распределений φ_t, θ_d и равномерными распределениями; регуляризаторы декоррелирования (формализации требования различности тем) – необходимо минимизировать ковариации между φ_t, φ_s.

Основная задача при построении АРТМ заключается в подборе траектории регуляризации – функции коэффициента регуляризации от номера итерации и критериев качества модели. Траектории регуляризации подбираются, анализируя их влияние на критерии качества модели в ходе итераций.

Измерение качества тематической модели ведется по совокупности критериев – оценивается *точность* и *интерпретируемость* модели.

В качестве критерия *точности* описания коллекции документов используется перплексия. Перплексия показывает, насколько хорошо модель приближает наблюдаемые частоты появления слов в документах. Точность модели тем выше, чем меньше перплексия.

$$perplexity(\mathcal{D}; p) = exp\left(-\frac{1}{n}L(\Phi, \Theta)\right) = exp\left(-\frac{1}{n}\sum_{d\in\mathcal{D}}\sum_{w\in d} n_{dw} \ln p\,(w|d)\right)$$

Интерпретируемость модели оценивается степенью разреженности матриц Φ и Θ (долей нулевых элементов в матрице), а также когерентностью.

Тема считается когерентной, если характеризующие ее наиболее вероятные слова неслучайно часто встречаются вместе. В ряде работ установлено, что именно когерентность коррелирует с человеческими оценками интерпретируемости тем.

Когерентность темы t (coherence) выражается через среднюю поточечную взаимную информацию (Pointwise Mutual Information, PMI) по всем парам k наиболее вероятных слов темы t:

$$coherence = \frac{2}{k(k-1)}\sum_{i=1}^{k-1}\sum_{j=1}^{k} PMI(w_i, w_j)$$

где w_i — i-й термин в порядке убывания φ_{wt}; $PMI(w_i, w_j)$ оценивает, насколько термы w_i, w_j не случайно встречаются рядом.

$$PMI(w_i, w_j) = ln\frac{p(w_i, w_j)}{p(w_i)p(w_j)} = ln\frac{n_{w_i,w_j}|D| + \varepsilon}{n_{w_i}n_{w_j}},$$

число наиболее вероятных слов в теме k обычно берется около 10-15;

ε вводится для избегания нуля под логарифмом, в данной работе $\varepsilon = 10^{-6}$.

Несмотря на широкое распространение подхода в области технических, физико-математических наук, наличие библиотеки для построения АРТМ (BigARTM, в среде Python), немалого числа решенных практических задач, отсутствие формализации алгоритма построения тематических моделей создает серьезный барьер для использования данного инструментария представителями экономических и других смежных наук. Ключевыми нерешенными проблемами являются определение стратегии регуляризации, обеспечение устойчивых и интерпретируемых результатов, анализ чувствительности результатов к изменениям параметров модели. Имеющиеся публикации, представляющие тематическую модель для той или иной задачи, не объясняют, как именно выбирается диапазон для перебора значений коэффициентов регуляризации. Всё это не позволяет сформировать стандарт исследовательской работы, который позволил бы различным научным группам сравнивать и воспроизводить результаты своих коллег.

Некоторым шагом вперед явилась разработка В.Г. Булатовым[6] относительных коэффициентов регуляризации (λ). Известно, что для регуляризаторов сглаживания и разреживания формула М-шага имеет вид: $\varphi_{wt} = norm_{w\epsilon W}(n_{wt} + \tau\beta_w)$, $\beta_w = (\frac{1}{|W|})$- равномерное распределение, $\tau > 0$ для регуляризатора сглаживания, $\tau < 0$ для регуляризатора разреживания.

$$\varphi_{wt} = \frac{n_{wt}+\tau\beta_w}{\sum_{w\in W} n_{wt}+\tau\beta_w} = \frac{n_{wt}+\tau\beta_w}{n_t+\tau} \qquad (10)$$

Влияние регуляризации можно описать как притягивание (в случае сглаживания) или отдаление (в случае разреживания) распределения n_{wt}/n_t , полученного как оценка максимального правдоподобия к равномерному распределению β_w с некоторым весом λ. φ_{wt} может быть записана как выпуклая комбинация этих двух распределений:

$$\varphi_{wt} = (1-\lambda)\frac{n_{wt}}{n_t} + \lambda\beta_w, \quad 0\leq \lambda \leq 1 \qquad (11)$$

Приравнивая (10) и (11), выражая τ: $\tau = \frac{n_t\lambda}{(1-\lambda)|W|}$

Таким образом, выражение М-шага имеет вид: $\varphi_{wt} = norm_{w\epsilon W}\left(n_{wt} + n_t\frac{\lambda}{(1-\lambda)}\beta_w\right)$.

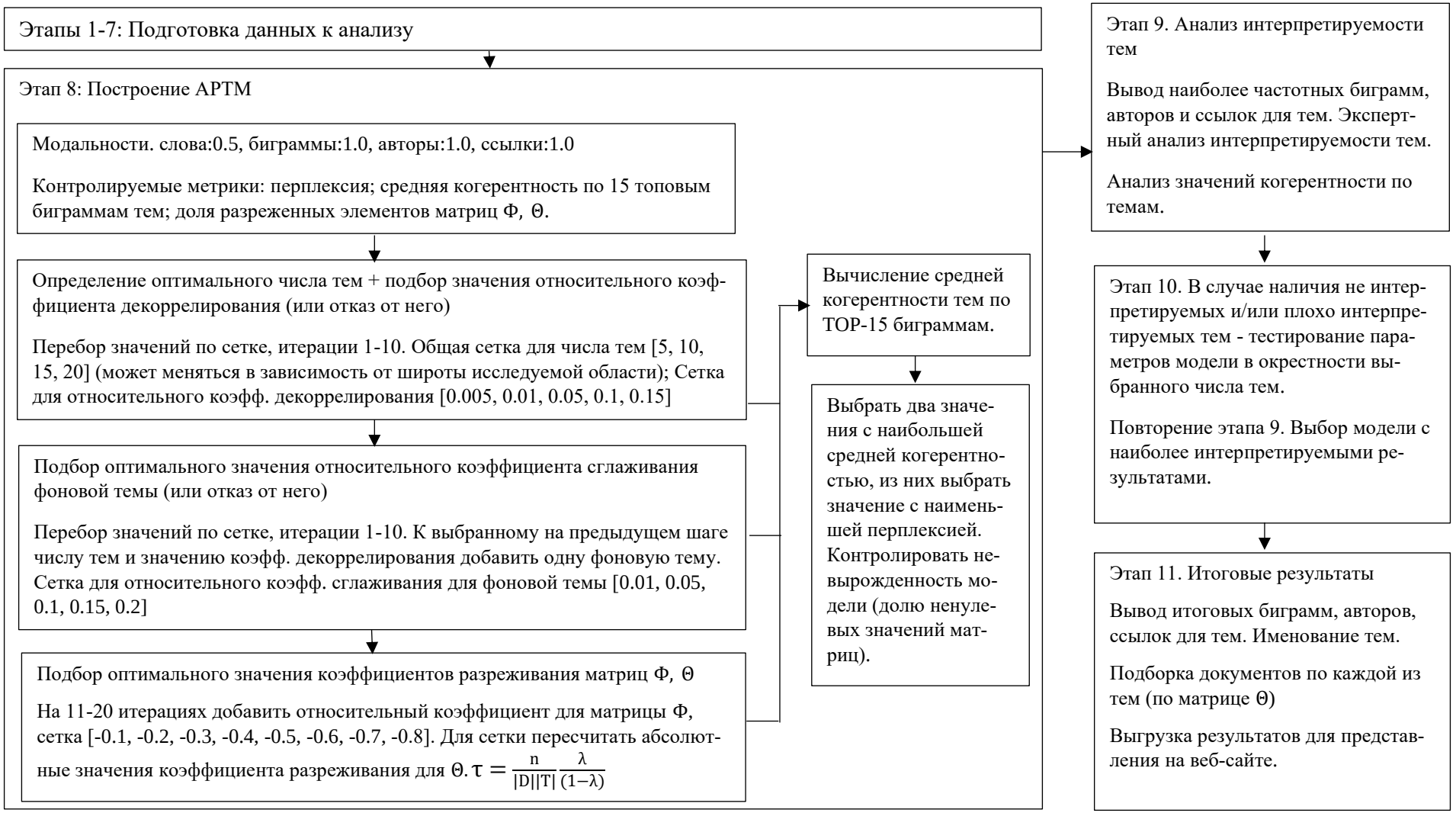
Величина $\frac{\lambda}{(1-\lambda)}$ определяет, во сколько раз регуляризатор влияет на оценку φ_{wt} больше, чем коллекция. Однако, чем больше значение n_t (число слов в теме), тем сильнее будет регуляризация. Возможно усреднение коэффициентов регуляризации по всем темам:

$$\varphi_{wt} = norm_{w\epsilon W}\left(n_{wt} + \frac{n}{|T|}\frac{\lambda}{(1-\lambda)}\beta_w\right).$$

[6] Булатов, В.Г. Методы оценивания качества и многокритериальной оптимизации тематических моделей в библиотеке TopicNet. Диссертация на соискание ученой степени канд. технич. наук: 05.13.18 / Булатов Виктор Геннадьевич — М., 2020. — 147 с.

3. Формализация алгоритма APTM для проведения мультимодальной мягкой кластеризации научных публикаций с целью получения информации о структуре больших текстовых коллекций.

Схема 1 – Формализованный алгоритм APTM для научных публикаций



4. Формализованный алгоритм АРТМ апробирован на базе научных статей по поведенческой и экспериментальной экономике

Апробация включает следующие этапы: (1) сбор релевантных данных; (2) проведение предобработки данных, приведение в пригодный для построения модели формат; (3) построение тематической модели на основе формализованного алгоритма аддитивной регуляризации, включающий выбор оптимального числа тем, оптимальной стратегии регуляризации (4) анализ чувствительности результатов; (5) верификация результатов у экспертного сообщества в области поведенческой экономики.

На основе коллекции англоязычных научных статей по поведенческой и экспериментальной экономике из базы Semantic Scholar (общее число рассматриваемых статей 37352) проведены расчеты, показывающие, что наилучшие результаты тематической модели (с точки зрения получения интерпретируемых результатов) достигаются на основе следующей метаинформации (включения следующих модальностей): 1) одиночных слов Названия и Аннотаций статей; 2) двусловных словосочетаний (биграмм) Названия и Аннотаций статей; 3) авторов статей; 4) списков использованной литературы. Число слов - 36485, число биграмм – 497933, число авторов – 54287, число ссылок – 461695.

Веса модальностей предлагается брать равными 0.5 для слов, 1.0 для биграмм, авторов и ссылок.

На основе формализованного алгоритма осуществлена мягкая кластеризация статей.

Выделено 15 тем. Итоговые параметры модели: 10 итераций без регуляризации, добавление на 11-20 итерациях (10 итераций) коэффициентов разреживания для матриц Ф ($\tau^{(d)}_{sparse\,\Phi}$) и Θ ($\tau^{(d)}_{sparse\,\Theta}$). Значения коэффициентов:

$\tau^{(d)}_{sparse\,\Phi} = -0.6$, $\tau^{(d)}_{sparse\,\Theta} = -2.81$ (в силу особенностей технической реализации для матрицы Θ приведены абсолютные значения коэффициента).

Значение средней когерентности $coh = -1.9052$; перплексия $perpl = 5.07e + 13$.

Для каждой из тем выделены ключевые биграммы, авторы и ссылки. Все кроме одной темы являются интерпретируемыми, совпадают с экспертным представлением о подтемах направления, устойчивы к изменению начальных инициализаций матриц Ф, Θ, поддаются экспертному именованию. Без какого-либо экспертного вмешательства в состав тем, на основе алгоритма выделены следующие подтемы (именованы вручную):

Ограниченная рациональность, теория перспектив; Проэкологическое поведение; Ограниченная рациональность, методы подталкивания и поведение потребителя; Общие вопросы поведенческой и экспериментальной экономики; Поведенческие факторы в корпоративном управлении; Экспериментальная экономика; Влияние фрейминга и других факторов на покупательную способность,

электронная коммерция; Влияние маркетинговой стратегии на потребителя; Поведенческая теория фирмы; Поведенческая экономика и налогообложение; Поведенческие финансы; Экспериментальные подходы к анализу готовности платить; Игровые подходы и теория перспектив в решении различных проблем; Политика в отношении современных проблем.

Результаты работы модели выгружены на специально созданный сайт http://behavioral.site.

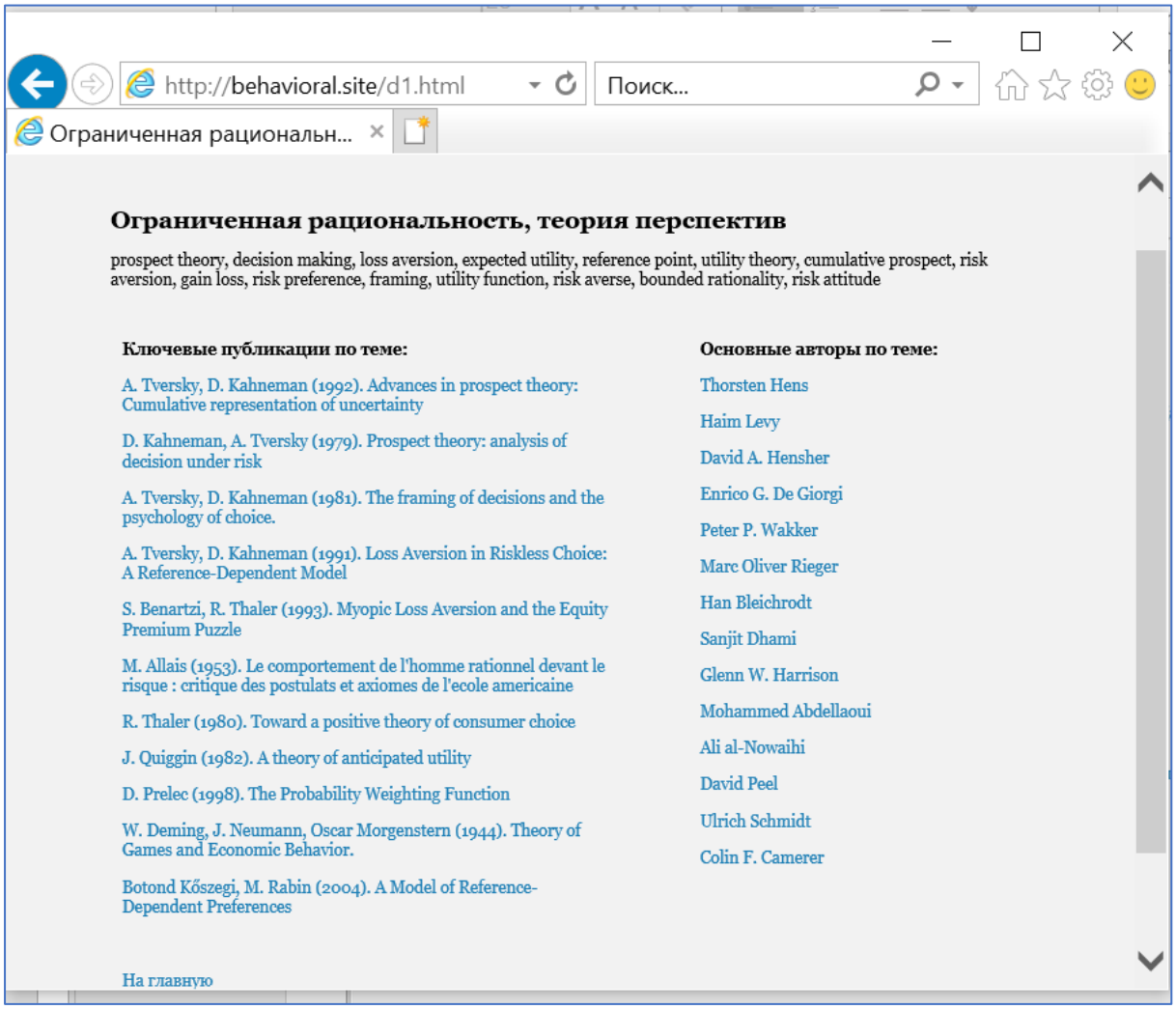


Рисунок 1 – Скриншот одной из страниц сайта. Приведены характеристики темы «Ограниченная рациональность, теория перспектив»: ключевые фразы, ключевые публикации по теме, основные авторы

5. Предложен и апробирован подход к оценке программы импортозамещения на основе патентных данных: в коллекции патентных документов выявлены и структурированы документы, соответствующие (по смыслу) пунктам планов по импортозамещению в 22 отраслях промышленности.

Задача получения представления о ключевых аспектах больших текстовых коллекций может быть сформулирована не только как задача мягкой кластеризации всей коллекции документов, но и

как задача кластеризации только определенной части документов по априорно задаваемым правилам. Остальные документы при этом не представляют интереса и отбрасываются. Пример решения задачи такого типа продемонстрирован применительно к анализу программы импортозамещения по 22 отраслям промышленности на основе патентных данных.

В каждой из отраслей составлен перечень продуктов/ технологий для импортозамещения. Общее число пунктов Планов равнялось 1553.

Исследование проводилось в преддверии окончания установленного для реализации Программы срока, и было важно представить некоторые результаты импортозамещения, основанные на анализе патентных данных. Несмотря на исключительную важность проведения подробного анализа по каждому из направлений развития, полезно иметь и общую структуру результатов. Подход, охватывающий сразу все отрасли, позволит как продемонстрировать результаты выполнения программы в целом, так и даст общее представление о состоянии различных отраслей экономики (на основе патентных данных).

Для построения тематической модели были собраны патенты на изобретения и полезные модели, выданные за 3,5-летний период (январь 2016 – июнь 2019 гг.) – всего 152718 документов: 120768 изобретений и 31950 полезных моделей. Тематическая модель строилась на основе Названий и Рефератов патентов, представленных в виде униграмм (т.е. одиночных слов). В отдельную модальность были выделены наиболее частотные биграммы (двусловные словосочетания с частотой встречаемости в Названии и Рефератов более или равной 2), оптимальный вес модальности биграмм определялся экспериментально и был выбран равным 5.

Были введены 22 регуляризатора сглаживания, по одному для каждой из тем. Коэффициент сглаживающих регуляризаторов (абсолютные значения) был выбран равным
$1e+7$. Общее число итераций: 40.

Ключевые метрики качества итоговой модели: доля разреженных элементов матриц униграмм $\Phi^1 = 0{,}994$, биграмм $\Phi^2 = 0{,}998$, $\Theta = 0{,}818$.

Итогом построенной модели стала кластеризация патентных документов: выбраны патентные документы по каждой из 22 отраслей в соответствии с темой, характеризуемой набором слов и словосочетаний из соответствующего Плана.

Помимо стандартных автоматически вычисляемых метрик, качество модели также оценивалась с помощью экспертов, определяющих, насколько релевантным является отобранный документ. Патентному документу ставилось в соответствие значение quality q=1 в случае, если патент точно соответствовал одному из заявленных в Плане пункту импортозамещения; значение q=0,5 присваивалось в случае, если патент связан с одним из пунктов Плана; q=0 – если не соответствовал ни одному из пунктов Плана.

Для документов со значениями q=1, q=0,5 была выделена ключевая фраза/слово, характеризующая как документ, так и его принадлежность к тому или иному пункту Плана. Какие именно позиции импортозамещались в каждой из отраслей наглядно показано на диаграммах Sankey. Например, для Черной металлургии:

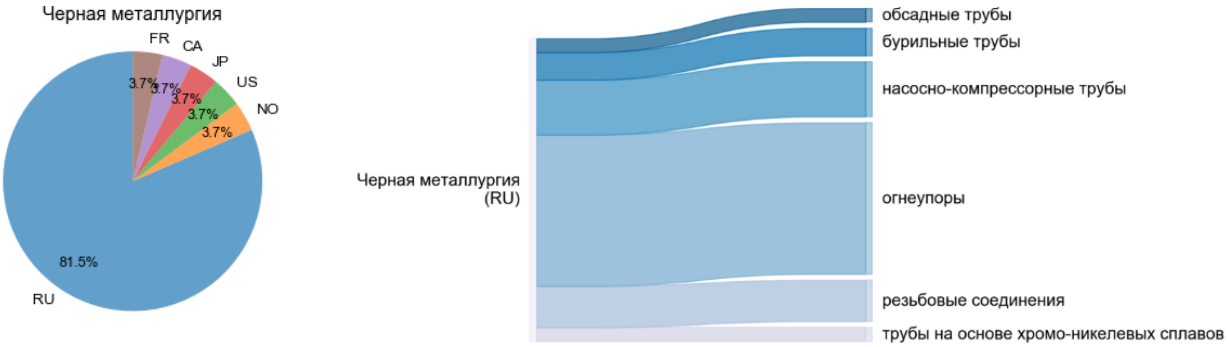


Рисунок 2 – Черная металлургия. Распределение по странам-патентообладателям и категориям патентования российскими патентообладателями

В Таблице 1 агрегированы результаты для всех отраслей.

Таблица 1 – Характеристики импортозамещения на основании патентных данных

Отрасль	Процент российских патентов	Категории импортозамещения (RU)	Число категорий, *k*	Средний балл, *q*	Суммарн. значение балла
Автомобильн. промышл.	74,1%	двигатель внутреннего сгорания	1	0,93	18,5
Гражданск. авиастроение		-	-	-	-
Детские товары	95,2%	мебель для детей; игры и игрушки; спортивные комплексы; детская одежда; детское творчество	5	0,95	19,0
Легкая промышленность	52,2%	нетканые материалы; защитная одежда; переработка шерсти	3	0,96	11,0
Лесопромышл. комплекс	12,5%	обработка целлюлозы; бумага, картон	2	0,75	3,0
Машиностр. для пищ. и перабат. промышл.	100,0%	обработка зерновых	1	0,83	2,5
Медицинская промышленность	41,7%	стерилизация и дезинфекция; эндоскопические аппараты; иглы инъекционные; имплантируемые насосы	4	0,90	4,5
Нефтегазовое машиностроение	78,3%	катализаторы гидроочистки; бурение скважин; катализаторы гидрокрекинга; переработка углеводородного сырья; гидроразрыв пласта;	6	1,00	18,0

Отрасль	Процент российских патентов	Категории импортозамещения (RU)	Число категорий, k	Средний балл, q	Суммарн. значение балла
		катализаторы каталитического крекинга			
Промышленность обычных вооруж.	82,4%	патроны; спортивное оружие	2	0,79	11,0
Радиоэлектронная промышленность		-	-	-	-
Сельскохоз. и лесное машиностр.	83,3%	подшипники; зерноуборочный комбайн; пресс подборщик	3	0,70	17,5
Станкоинструментальная промышл.	78,9%	фрезерный станок; токарный станок; расточный станок; шпиндели; финишное шлифование; гидроабразивная резка; станки чпу	7	0,93	14,0
Строит. материалы и строит. конструкц.	96,6%	керамическая масса для плитки; теплоизоляционные материалы; щебеночно-мастичные асфальтобетоны	3	0,68	19,0
Строительно-дорожн. техника*	93,3%	дорожное покрытие; гидравлическое оборудование; фронтальные погрузчики; бульдозеры; фронтальный погрузчик; экскаватор; прицеп и полуприцеп; крановое шасси; коммунальная техника	9	0,79	11,0
Судостроительная промышленность	92,9%	движитель; гребневой винт	2	0,50	6,5
Транспортн. машиностр.	62,7%	вагон-цисцерна; тормозная система; тележки вагона; крытый вагон	4	0,81	26,0
Тяжелое машиностр.	50,0%	крепь горная; холодильные установки	2	0,75	1,5
Фармацевтич. промышл.	56,3%	инозин + никотинамид + рибофлавин + янтарная кислота; висмут калий аммоний цитрат; дротаверин; йогексол; лопинавир + ритонавир; этилметилгидроксипиридина сукцинат; рокурония бромид; дигоксин; 1 карбамоилметил 4 фенил 2 пирролидон; фенспирид; изониазид; лаппаконитина гидробромид; иммуноглобулин стандартный; бромдигидрохлорфенил-бензодиазепин; десмопрессин; финголимод; анастрозол	17	0,94	17,0

Отрасль	Процент российских патентов	Категории импортозамещения (RU)	Число категорий, k	Средний балл, q	Суммарн. значение балла
Химическая промышленность	87,5%	лакокрасочные материалы; уплотнительные материалы; эпоксидный композит; клеевые материалы; полиэтилентерефталат; сверхвысокомолекулярный полиэтилена; полимерные композиты	7	0,79	11,0
Цветная металлургия	88,6%	алюминиевый сплав; алюминий, электролиз; алюминиевая лигатура; гидроксид алюминия; алюминиевый порошок; алюминиевая фольга; алюминиевые прутки; анодная масса	8	0,81	25,0
Черная металлургия	81,5%	огнеупоры; насосно-компрессорные трубы; резьбовые соединения; бурильные трубы; трубы на основе хромо-никелевых сплавов; обсадные трубы	6	0,89	19,5
Энергетическое машиностр.**	50,0%	трансформаторы тока	1	1,00	1,0

Подобный вариант кластеризации удобен, так как позволяет емко представить структуру коллекции патентных документов, сформированную уже на основе только тех слов и словосочетаний, которые представляют для нас интерес.

В зависимости от числа найденных релевантных патентных документов, степени их соответствия заявленному плану, а также общему количеству пунктов Плана, каждая из отраслей получила свой рейтинг: $\text{Score} = {\sum(q) \cdot k}/{N}$, где k – число категорий – различных позиций Плана, по которым были найдены релевантные патентные документы (q=1, q=0,5); N – общее число пунктов Плана. Результаты ранжирования отраслей представлены на Рисунке 3.

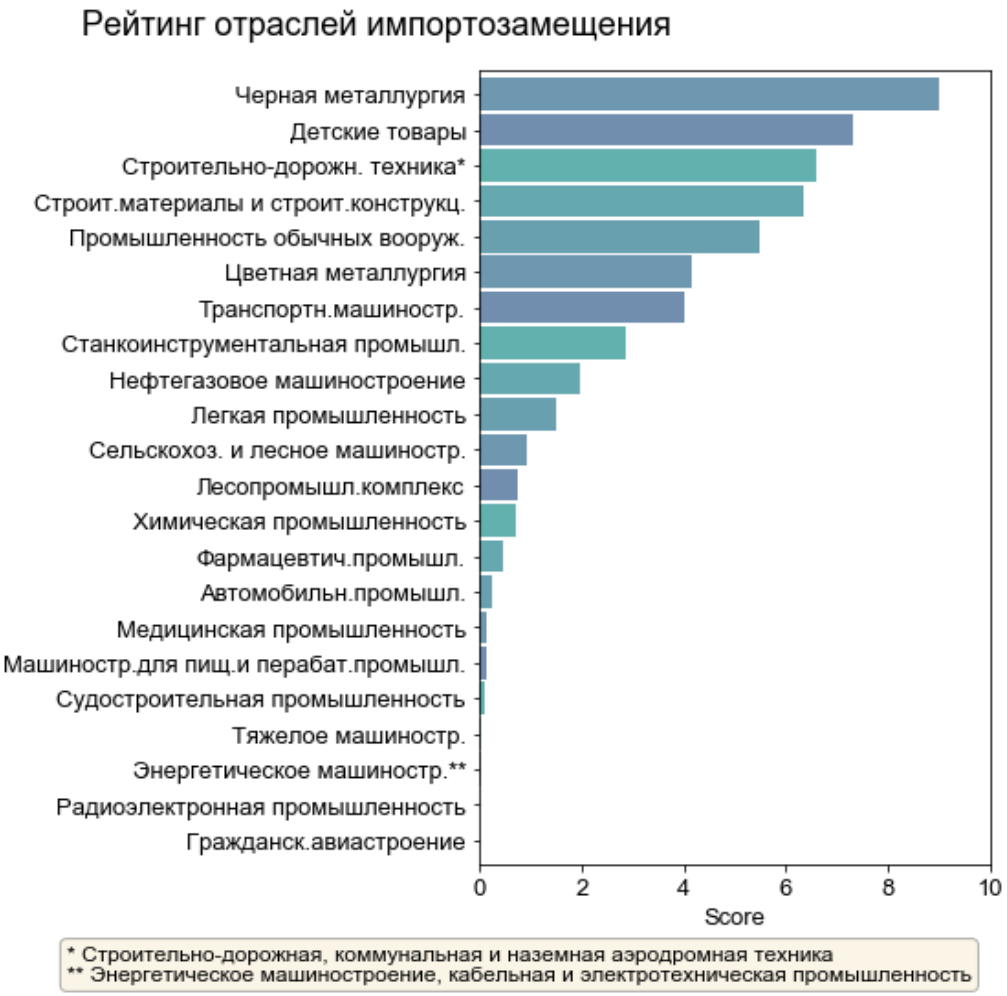


Рисунок 3 – Рейтинг отраслей импортозамещения на основании соответствия выданных патентов заявленному плану импортозамещения отрасли

III. РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. В данной работе предлагается способ организации потребления информационных продуктов, помощи выбора из множества. В основе лежат идеи (Г. Саймон, Ф. Хайек), что 1) внимание ограничено определенным классом результатов, которые мы ожидаем видеть, 2) внимание определяет ту узкую избирательную область, которая выделяется в общем проблемном пространстве, внутри которой осуществляется поиск решений, 3) знания имеют фрагментарную природу.

Дизайн цифрового мира влияет на поведение и принимаемые решения, и использование достижений поведенческих наук в сочетании с цифровыми технологиями способно оказывать существенное воздействие. Таким образом, учитывая особенности потребления информации экономическими агентами разработан и показан в экспериментах способ организации информационных продуктов, позволяющий сконфигурировать среду для восприятия информации, чтобы 1) снизить затраты на поиск ключевой информации по исследуемой научной теме; 2) обеспечить полноту и ценность (с точки зрения смысла) извлекаемой информации; 3) минимизировать сдвиги в восприятии

(сместить фокус внимания с первой попавшейся, эмоциональной информации). Применительно к текстовой информации, данным требованиям удовлетворяет подход аддитивной регуляризации тематических моделей. Таким образом, именно инструментарий тематического моделирования (для семантической компрессии текстовой информации) может оказать помощь как в управлении ограниченной рациональностью, так и в информационном взаимодействии производителей и потребителей научного знания.

Формализованный алгоритм аддитивной регуляризации тематических моделей представляет собой набор понятных инструкций, реализован в виде набора программ на языке Python. Применительно к решению задачи о представлении больших объемов научных публикаций предложено включать следующие модальности: слова, биграммы, авторы, ссылки (список статей, на которые ссылается публикация). Веса модальностей 0.5, 1.0, 1.0, 1.0 соответственно. На первом этапе осуществляется перебор по сетке оптимального числа тем вместе с относительным коэффициентом декоррелирования. На втором этапе тестируется значение коэффициента сглаживания фоновой темы, который добавляется к коэффициенту декоррелирования. Проводится 10 итераций. На 11-20 итерации добавляется регуляризатор разреживания для матриц. На каждом из этапов контролируется перплексия и средняя когерентность тем. Выбирается два значения с наибольшей когерентностью, из них выбирается значение, соответствующее наименьшей перплексии.

Созданный набор программ позволяет решать задачи тематической кластеризации для коллекций научных публикаций. Результаты выложены в открытом репозитории Github: https://github.com/behavioral-econ-codes/Publications

2. Авторская методика экспериментально проверена на примере научных публикаций по поведенческой и экспериментальной экономике. Выделены 15 тем, для каждой из которых представлены ключевые биграммы, авторы и ссылки. Все кроме одной темы являются интерпретируемыми, совпадают с экспертным представлением о темах направления, поддаются экспертному именованию. Без какого-либо экспертного вмешательства в состав тем, на основе алгоритма выделены следующие темы (именованы вручную): *Ограниченная рациональность, теория перспектив; Проэкологическое поведение; Ограниченная рациональность, методы подталкивания и поведение потребителя; Общие вопросы поведенческой и экспериментальной экономики; Поведенческие факторы в корпоративном управлении; Экспериментальная экономика; Влияние фрейминга и других факторов на покупательную способность, электронная коммерция; Влияние маркетинговой стратегии на потребителя; Поведенческая теория фирмы; Поведенческая экономика и налогообложение; Поведенческие финансы; Экспериментальные подходы к анализу готовности платить; Игровые подходы и теория перспектив в решении различных проблем; Политика в отношении современных проблем.*

Подход позволяет повысить эффективность работы научного сообщества путем организации информационного взаимодействия производителей и потребителей научного знания. Также, в свете обсуждений теории подталкивания, в особенности цифрового подталкивания, подход, основанный на использовании тематического моделирования текстовых коллекций документов, позволяет осуществлять своего рода «подталкивание к изучению», предоставляя структуру исследуемого направления и его ключевые положения. Это – первый шаг к получению знаний, не отменяющий дальнейшее углубление в рассматриваемую тему.

3. Проведен анализ чувствительности результатов тематического моделирования к изменению начальных инициализаций матриц. Показано, что в случае, если темы являются хорошо интерпретируемыми, их состав не меняется существенно в зависимости от начальных приближений.

4. Создан онлайн ресурс http://behavioral.site для осуществления пользовательской навигации по коллекции статей по поведенческой и экспериментальной экономике, полученной в результате работы формализованного алгоритма APTM.

5. Проведен анализ программы импортозамещения на основе патентных данных с использованием инструментария аддитивной регуляризации тематических моделей. Подход позволяет осуществлять поиск документов по смыслу (согласно пунктам двадцати двух отраслевых планов импортозамещения) и на выходе получать подборку релевантных документов по каждой из отраслей. Данный подход является своего рода «крупным планом» патентного поиска, который может служить как конечной целью, так и являться отправной точкой для более детального анализа.

6. Автором также решены схожие задачи, оставшиеся за рамками данной диссертационной работы, демонстрирующие применение методов семантической компрессии текстовой информации для быстрого получения представления об исследуемой области. Так, на основе графового алгоритма Textrank была решена задача выделения ключевых слов (и их семантических связей) из правительственных документов направления Цифровая экономика. На основе алгоритма bm25 была решена задача извлечения ключевых предложений из нормативных документов, регламентирующих вопросы продовольствия и питания.

Также, в рамках изучения ограниченной рациональности при принятии решений, автором была разработана система для поддержки принятия многокритериальных решений на основе метода аналитических сетей/иерархий.

IV. СПИСОК ПУБЛИКАЦИЙ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи, опубликованные в журналах из перечня ведущих рецензируемых научных изданий ВАК Минобрнауки РФ

1. Милкова М.А. Инновационный подход к поиску информации на примере патентного анализа плана импортозамещения // Экономическая наука современной России. 2020. № 1(88). С. 143-157. DOI: 10.33293/1609-1442-2020-1(88)-143-157 (1.7 п.л.).

2. Милкова М.А., Неволин И.В., Пигорев Д.П. Извлечение ключевой информации из нормативных документов о политике продовольствия и питания в России // Экономическая наука современной России. 2021. № 2(93). С. 101-114. DOI: 10.33293/1609-1442-2021-2(93)-101-114 (2.2 п.л., авт. – 1.8 п.л.).

3. Milkova, M., Andreichikova, O., Andreichikov, A. Decision making under uncertainty: a heuristics overview and the analytic network process // Psychology Journal of the Higher School of Economics. 2019. Vol. 16. N 4, pp. 730–751. DOI: 10.17323/1813-8918-2019-4-730-751 (2.8 п.л., авт. – 2.0 п.л.).

Статьи, опубликованные в научных изданиях, входящих в базу данных SCOPUS

4. Milkova M.A. Patent-based import substitution analysis with Additively Regularized Topic Models // Proceedings of the 10th International Scientific and Practical Conference named after A. I. Kitov "Information Technologies and Mathematical Methods in Economics and Management (IT&MM-2020)". Moscow, Russia, October 15-16, 2020. CEUR Workshop Proceedings. – 2021, Vol. 2830, pp. 16-27. urn:nbn:de:0074-2830-9 (1.0 п.л.).

Статьи, опубликованные в других научных журналах:

5. Милкова М.А. Тематическое моделирование: восприятие научной информации // Цифровая экономика. 2021. № 14(2). С. 31-36. DOI: 10.34706/DE-2021-02-04 (0.9 п.л.).

6. Милкова М.А. Информация и ограниченная рациональность выбора в цифровой экономике // Цифровая экономика. 2021. № 13(1). С. 69-88. DOI: 10.34706/DE-2021-01-08 (3.2 п.л.).

7. Милкова М.А. Феномен внимания в информационной среде: экономика внимания // Цифровая экономика. 2020. № 3(11). С. 73-87. DOI: 10.34706/DE-2020-03-08 (3.5 п.л.).

8. Милкова М.А. Тематическое моделирование патентных документов как инновационная технология управления вниманием в условиях перенасыщения информации. Современные вызовы и реалии экономического развития России: материалы VI Международной научно-практической конференции / под ред. Л.И. Ушвицкого, А.В. Савцовой. – Ставрополь: Издательско-информационный центр «Фабула», 2020. С. 117-120 (0.4 п.л.).

9. Милкова М.А. Экономика внимания и ее инструментальное обеспечение / IV Российский экономический конгресс «РЭК-2020». Том XIX. Тематическая конференция «Поведенческая и

экспериментальная экономика» (сборник материалов) / Составители А.В. Белянин, А.Д. Суворов. – М., 2020. С. 62-64 (0.3 п.л.).

10. Милкова М.А. Восприятие информации в период пандемии COVID-19 / Системное моделирование социально-экономических процессов: труды 43-й Международной научной школы-семинара, г. Воронеж / под ред. В.Г. Гребенникова, И.Н. Щепиной. – Воронеж: Изд-во «Истоки», 2020. С. 334-338 (0.3 п.л.).

11. Милкова М.А. (2019). Теория подталкивая и ее искажения в информационной среде // Цифровая экономика, 4(8). С. 21-26. DOI: 10.34706/DE-2019-04-02 (0.9 п.л.).

12. Милкова М.А. Тематические модели как инструмент «дальнего чтения» // Цифровая экономика. 2019. № 1(5). С. 57-70. DOI:10.34706/DE-2019-01-06 (1.2 п.л.).

13. Милкова М.А. Извлечение ключевых терминов направления «Цифровая экономика»: графоориентированный подход // Цифровая экономика. 2018. № 4(4). С. 57-65. DOI:10.34706/DE-2018-04-06 (0.8 п.л.).

Переводы, выполненные и опубликованные автором по теме диссертации:

Санстейн, К.Р. Пренебрежение вероятностью: эмоции, наихудшие случаи и право // Цифровая экономика. 2020. № 4(12). С. 49-74. DOI: 10.34706/DE-2020-04-06. Перевод с англ.: Cass R. Sunstein (2001). “Probability Neglect: Emotions, Worst Cases, and Law", The Yale Law Journal, 112(1), pp. 61-107.

Франк, Г. За пределами денег и информации: экономика внимания // Цифровая экономика. 2020. № 2(10). С.45-51. DOI: 10.34706/DE-2020-02-04. Перевод с нем.: Georg Franck (2007). “Jenseits von Geld und Information: Zur Ökonomie der Aufmerksamkeit”. Handbuch Unternehmenskommunikation, pp. 159-168.

Франк, Г. Оплата славой: как неэпистемологические мотивы способствовали феноменальному успеху современной науки // Цифровая экономика. 2020. № 1(9). С. 58-62. DOI: 10.34706/DE-2020-01-06. Перевод с англ.: Georg Franck (2015). “The Wage of Fame: How Non-Epistemic Motives Have Enabled the Phenomenal Success of Modern Science”, Gerontology, 61 (1), pp. 89-94.

Работы по смежным темам, опубликованные в том числе, в журналах из перечня ВАК, Scopus, WoS:

Milkova, M., Andreichikova, O., Andreichikov, A. (2019). At the junction of mathematics and psychology: cognitive orientation of the AHP/ANP and new perspectives of structuring complexity // International Journal of the Analytic Hierarchy Process, 11(1), DOI: 10.13033/ijahp.v11i1.611.

Milkova, M., Andreichikova, O., Andreichikov, A. Venture capitalists decision making: applying Analytic Network Process to the startups evaluation // International Journal of the Analytic Hierarchy Process. 2018. Vol. 10(1). DOI: 10.13033/ijahp.v10i1.511.

Milkova M., Andreichikova O. Software announcement: Multichoice as new software for decision making with Analytic Network Process // International Journal of the Analytic Hierarchy Process. 2016. Vol. 8(2). DOI: 10.13033/ijahp.v8i2.413.

Андрейчикова О., Милкова М. Применение метода аналитических сетей для сравнительной оценки деятельности молодых компаний-стартапов // Экономика и предпринимательство. 2016. № 3 ч.1. С. 785-792.

Свидетельства о государственной регистрации:

Свидетельство о регистрации программы для ЭВМ №2016616698, Multichoice / Милкова М.А., Андрейчикова О.Н. // Федеральная служба по интеллектуальной собственности, 17.06.2016.