*На правах рукописи*

Милкова Мария Александровна

**РАЗРАБОТКА И ЭКСПЕРИМЕНТАЛЬНАЯ АПРОБАЦИЯ МЕТОДИКИ ПОСТРОЕНИЯ ТЕМАТИЧЕСКОЙ МОДЕЛИ ПРИ РАБОТЕ С ПЕРЕИЗБЫТКОМ НАУЧНОЙ ИНФОРМАЦИИ**

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 Центральный экономико-математический институт Российской академии наук

|  |  |
| --- | --- |
| НАУЧНЫЙ РУКОВОДИТЕЛЬ: | **Козырев Анатолий Николаевич**  доктор экономических наук, кандидат физико-экономических наук, заведующей лаборатории экспериментальной экономики Федерального государственного бюджетного  учреждения науки Центральный экономикоматематический институт Российской академии наук |
| ОФИЦИАЛЬНЫЕ ОППОНЕНТЫ: |  |
|  |  |
| ВЕДУЩАЯ ОРГАНИЗАЦИЯ: |  |

Защита состоится \_\_\_\_\_\_\_\_\_\_\_ 2021 г. в \_\_\_\_ на заседании диссертационного совета Д 002.013.01 при Федеральном государственном бюджетном учреждении науки Центральный экономико-математический институт Российской академии наук по адресу: 117418, Москва, Нахимовский проспект, д. 47, ауд. 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 <http://vak.gov.edu>.

Автореферат разослан \_\_\_\_\_\_\_\_ 2021 г.

Ученый секретарь диссертационного совета Д 002.013.01,

кандидат экономических наук А.И. Ставчиков

1. **ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ**

**Актуальность темы исследования**

Проведение научных исследований, в том числе, в области экономики и других общественных наук предполагает использование инструментов для сбора и анализа информации различного характера, представленной в разных форматах. Это могут быть результаты экспериментов, статистические данные или тексты, как правило, в цифровом формате. Применительно к экономике и экономической науке в этом плане уместно говорить, с одной стороны, об информации, используемой экономическими агентами для принятия тех или иных решений, с другой стороны, об информации, используемой в научных исследованиях по экономике, в том числе, содержащейся в массиве научных текстов (статей, монографий, докладов) и документов. Также стоит отметить возможности для проведения научных экспериментов в экономике, ставших реальностью с появлением новых информационных технологий.

Меняется и само понимания эксперимента в экономической науке. Так, в 60х годах для тестирования предсказаний экономической теории и изучения поведения экономических агентов получили развитие методы лабораторного эксперимента, в недавнем времени – адаптирован метод рандомизированного контрольного исследования к тестированию эффективности политики экономического развития[[1]](#footnote-1).

Во всех отмеченных выше случаях речь идет об объемах информации, значительно превосходящих возможности человека по ее охвату, анализу и осмыслению без использования специальных инструментов. Актуальными 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. Также чрезвычайно важна интерпретируемость получаемых результатов. Это – отдельная проблема.

Если смотреть с обозначенной выше точки зрения на поведение людей, то нельзя не отметить следующее. В эпоху цифровой трансформации процесс накопления информации настолько стремительный, что это порождает новые сложности не только в осуществлении исследовательского поиска, обработки доступной информации и получения знаний, но и в повседневной жизни, в том числе в поведении экономических агентов, то есть в том, что изучает экономическая наука. Данные тенденции сочетаются с ограниченной рациональностью человека. Несмотря на доступность цифровой информации и развитие поисковых систем, социальное ускорение, клиповое мышление и другие факторы приводят к существенному снижению порога, на котором принимается решение о прекращении поиска информации, что имеет негативное долгосрочное влияние на развитие человеческого капитала. Глобальная смена принципов работы с информацией, направленная на предоставление готовых ответов и решений (поисковыми и другими системами), может привести к деградации способности анализировать и мыслить системно. Самым дефицитным ресурсом становится внимание.

Происходящие изменения принципов работы с информацией характеризуются кардинальной сменой распределения внимания, регулирующего взаимодействие людей с миром как на индивидуальном, так и на социальном уровне. То, как мы распределяем внимание, определяет, какие решения мы принимаем, что, в свою очередь, влияет на экономику, политику и этику. Возрастающая конкуренция за внимание привела к росту новых технологий для привлечения внимания, которые активно используют свойства человеческого мозга (например, автоматическую конформность, доминирование эмоциональных областей), а также склонность к принятию решений на основе эвристик.

В условиях перенасыщения информации условие полного информирования об исследуемой проблеме при проведении исследования или принятии решения является сложно выполнимым. Таким образом, актуальность приобретает разработка инструментов, позволяющих охватить больший объем информации, представить дорожную карту исследуемого вопроса, вместо обращения внимания на основную, популярную или продвигаемую информацию.

**Степень научной разработанности проблемы**

Степень научной разработанности темы стоит рассматривать 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 инструментальное обеспечение ограниченной рациональности в условиях перенасыщения цифровой информации, как оно трактуется в данной работе, лежит в области активно развиваемого в настоящее время направления – анализа естественного языка, а именно, различные методы семантической компрессии, к которым относится тематическое моделирование. Далее, проработанность темы описана в трёх ключевых направлениях: 1) ценность представления информации в надлежащем виде, 2) личностные препятствия к представлению и восприятию информации и 3) инструменты для повышения ценности представляемой 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, Г. 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. Симс, 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).

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, C. 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. Ключарева.

Ограниченное внимание как важное понятие поведенческой экономики выделялось К. Камерером, 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, а также отечественными учеными - А.Н. Козыревом, 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. Работы, подчеркивающие и развивающие междисциплинарное сотрудничество экономики, психологии, социологии - Г.Б. Клейнер, Д.А. Жданов; среди поведенческих экономистов – А.В. Белянин, К. Паниди.

Вопрос автоматического понимания текстов является достаточно проработанным в области компьютерных и технических наук. Более того, анализ естественного языка (NLP) – одна из наиболее динамично развивающихся областей искусственного интеллекта. В данной работе представляют интерес методы семантической компрессии информации как инструмент разведочного поиска. Концепция разведочного информационного поиска (exploratory search) развивалась, в противовес обычному итерационному, последнее десятилетие и предложена Г. Марчионини, Р. Уайтом, Р. Рохом. Подход к разведочному поиску как задаче построения тематической модели предложен отечественным ученым К.В. Воронцовым. В целом, развитию тематического моделирования как такового способствовали работы Т. Хоффмана, Д. Блея. К. Воронцовым предложен подход на основе аддитивной регуляризации тематических моделей (АРТМ), который активно развивается в настоящее время им и его учениками: А. Потапенко, М. Апишевым, А. Яниной, В.Г. Булатовым и др.

Среди отечественных ученых, применяющих методы семантического структурирования контента научных электронных библиотек, в том числе с использованием тематического моделирования, стоит выделить работы 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 сферы их применения. В том числе, расширяется область исследований, которые вводят в экономический дискурс анализ патентной информации, выходящей далеко за рамки патентной классификаций. 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 как направление экономической науки, тесно связанное с другими областями науки и другими направлениями экономической науки, включая экономику знаний, информационную экономику, нейроэкономику и психологию. Кроме того, именно для поведенческой экономики существуют исследования о том, кто и каким образом внес вклад в развитие данного направления (Аннотированная библиография П.Уаккера[[2]](#footnote-2)). 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 алгоритма аддитивной регуляризации тематических моделей для выявления структуры научно-технической информации, способствующей эффективному получению знаний. Разработать методику построения тематической модели (в концепции АРТМ), позволяющую сконфигурировать среду для восприятия научно-технической литературы, так чтобы: 1) снизить затраты на поиск ключевой информации по исследуемой научной теме; 2) минимизировать сдвиги в восприятии (сместить фокус внимания с первой попавшейся, эмоциональной информации); 3) повысить познавательные функции, избежать деградации аналитического мышления.

*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, позволяющих выявлять тематическую структуру корпуса научных публикаций.
2. Экспериментально проверить авторскую методику на примере научных статей по поведенческой и экспериментальной экономике. При проверке выявить ключевые подтемы 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 структуризовать только те, что соответствуют (по смыслу) пунктам планов по импортозамещению по 22 отраслям промышленности.

**Объект и предмет исследования**

*Объектом* исследования является научная и техническая информация, представляемая в виде научных публикаций (на примере публикаций по поведенческой и экспериментальной экономике), патентных документов, *предметом* –способы выявления в больших текстовых коллекциях ключевой информации, актуальной для формирования новых знаний, а также возможности и эффективного использования в экономических исследованиях.

**Научная новизна** диссертационной работы заключается в разработке комплекса вычислительных программ, позволяющих адаптировать применения тематического моделирования (в концепции АРТМ) для применения в экономических исследованиях теоретического и прикладного характера, в том числе в экономике знаний, поведенческой экономике, а также при выполнении прикладных экономических исследований.

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

1. Показана эффективность применения методов тематического моделирования при проведении исследований на стыке нескольких научных дисциплин или направлений. В частности, проведен анализ литературы на стыке экономики информации, поведенческой экономики и нейроэкономики, в результате которого показана фрагментарная природа формирования новых знаний, наличие когнитивных искажений при восприятии информации и других особенностей, замедляющих развитие указанных направлений. Результаты исследования свидетельствуют о целесообразности применения авторских наработок при анализе сложной предметной области.
2. В экспериментах показана эффективность двух ключевых подходов к семантической компрессии информации с помощью применения метода аддитивной регуляризации тематических моделей: 1) мягкой кластеризации документов с целью получения информации о структуре; 2) разведочного поиска релевантных документов на основе блоков априорно задаваемой информации.
3. Предложен формализованный подход к выявлению ключевой информации в коллекции научных публикаций по поведенческой и экспериментальной экономике: разработан комплекс программ, проведены эксперименты, представлен набор инструкций, позволяющих быстро выявлять структуру научных областей (на базе научных публикаций).
4. Предложен и апробирован подход к оценке результатов программы импортозамещения на основе патентных данных с помощью применения методов тематического моделирования.

**Теоретическая и практическая значимость диссертационного исследования**

**Теоретическая значимость** исследования заключается в следующем:

1. В рамках проводимого исследования, работа систематизирует результаты не только экономических, физико-математических наук и технических, но и соответствует критериям междисциплинарности: опирается на достижения психологических, социологических, нейроэкономических наук.
2. Разработанный подход представления научных публикаций вносит вклад в экономику знаний в части расширения формализованного и измеримого описания информационных объектов, которые могут выступать объектом экономических отношений.
3. Формализация построения тематических моделей является методическим вкладом в развитие теории исследования текстовой информации в экономической науке.

**Практическая значимость** исследования заключается в следующем:

1. Предложенная формализованная процедура тематического моделирования на основе аддитивной регуляризации тематических моделей представляет собой значимый вклад в практику выделения ключевых тем в текстовой информации и может быть использована в различных областях экономической науки, 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, анализе динамики развития научных областей (внутри поведенческой экономики). Подход позволяет повысить эффективность работы научного сообщества, что является важнейшим звеном функционирования экономики знаний.
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 диссертации использовались методы научного познания теоретического и эмпирического уровней.

На теоретическом уровне в качестве ключевых методов исследования применены: анализ, синтез, обобщение, моделирование.

На эмпирическом уровне основными методами являлись:

1. Метод информационного поиска на основе комбинирования регулярных выражений с целью выявления заданной исследователем информации в массивах текстовых 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 объемов текстовых данных, 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**

В качестве информационной базы для диссертации использовались:

1. Исследования отечественных и зарубежных авторов, материалы периодической печати, медиа-источники, представленные в сети Интернет
2. Научные статьи (название и аннотация) по поведенческой и экспериментальной экономике, находящиеся в репозитории Semantic Scholar[[3]](#footnote-3), относящиеся к областям экономики и бизнеса, до 2020 года включительно. Всего 37352 статьи.
3. Патентные документы за период январь 2016 - июнь 2019 гг. из базы Роспатента[[4]](#footnote-4). Всего 152718 документов; отраслевые планы импортозамещения Минпромторг для 22 отраслей промышленности[[5]](#footnote-5).

**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. Х International Scientific and Practical Conference named after A. I. Kitov "Information Technologies and Mathematical Methods in Economics and Management" – October 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.

**Структура диссертации**

Цель и задачи определили структуру настоящего исследования. Диссертационная работа содержит введение, три главы, заключение и библиографию общим объемом 171 стр., включая 8 таблиц, 12 рисунков и 7 Приложений.

В *первой главе* систематизированы ключевые особенности восприятия информации и препятствия принятия экономических решений человеком, а также лежащие в основе механизмы, исследованные в экономике информации, экономике знаний, поведенческой экономике, нейроэкономике.

*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 разведочного поиска по базе патентных документов на основе блоков задаваемой априорной информации для поиска (пункты планов для импортозамещения по 22 отраслям промышленности).

1. **ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ**
2. **Проведен мультидисциплинарный анализ литературных источников на стыке экономики информации, экономики знаний, поведенческой экономики, нейроэкономики 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.

Если экономика информации носит аналитический и математический характер, то экономика знаний является преимущественно эмпирической и часто описательной. 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 перенасыщения информации.

К инструментам поведенческой экономики относятся: метод контролируемого лабораторного эксперимента – для тестирования предсказаний экономической теории и изучения поведения экономических агентов; метод визуализации мозговой активности – для понимания нейробиологических принципов принятия решения; метод подталкивания – для управления принятием решений; анализ естественного языка (извлечение информации из текстов) – для проведения эмпирических исследований, демонстрирующих особенности формирования суждений, а также для автоматического извлечения из больших массивов текстов значимой для человека информации. Отмечено, что достижения в области анализа естественного языка в основном используются для внедрения в сторонние системы (рекомендательные, вопросно-ответные, поисковые и др.), которые, в свою очередь, призваны снижать затраты пользователей на поиск информации, способствовать решению проблем, связанных с дефицитом внимания.

В экономических исследованиях в настоящее время отведено место человеку, обладающему ограниченной рациональностью, учитывающему при принятии решений личные установки, общественные нормы и другие внешние факторы. Если моделированию ограниченной рациональности в экономических науках посвящен ряд работ, то развитию инструментов по управлению ограниченной рациональности уделено значительно меньше внимания. Инструмент подталкивания показывает положительные результаты во множестве исследований, однако применимость его значительно ограничена в том плане, что формирование привычки выбора опции по умолчанию имеет негативное долгосрочное воздействие на развитие человеческого капитала. Стоит отметить, что к подталкиванию, облегчающему выбор индивида, присоединяется и борьба за внимание как наиболее дефицитный ресурс.

**2.** **В качестве инструмента, который способствует эффективному получению знаний, предлагается использовать подход аддитивной регуляризации тематических моделей (АРТМ).**

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

Пусть есть коллекция документов , каждый из которых представляет собой последовательность терминов ( из словаря . Термин может повторяться в документе несколько раз.

Предполагается, что существует конечное множество тем T, и каждое употребление терма w в каждом документе d связано с некоторой темой t ∈ T. Термы w и документы d являются наблюдаемыми переменными, тема t ∈ T является латентной (скрытой) переменной.

Подход аддитивной регуляризации представляет задачу тематического моделирования как задачу поиска разложения матрицы в произведение двух неотрицательных нормированных матриц меньшего размера:

матрицы термов в темах ,

матрицы тем в документах ,

Для определения параметров модели Φ, Θ максимизируется правдоподобие.

Стохастическое матричное разложение определено не единственным образом, а с точностью до невырожденного преобразования , то есть задача является некорректно поставленной. Решение такой задачи можно доопределить и сделать устойчивым путем добавления дополнительного критерия — регуляризатора

Таким образом, наряду с правдоподобием требуется максимизировать r критериев-регуляризаторов , i = 1,…r с неотрицательными коэффициентами регуляризации . Таким образом, оптимизационная задача имеет вид:

(1)

,

,

,

Решение задачи (1) строится на основе так называемого регуляризованного EM-алгоритма. На первом этапе выбирается начальное приближение для , На E-шаге вычисляются вспомогательные переменные :

, (2)

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

На M-шаге вычисляются частотные оценки максимального правдоподобия для искомых условных вероятностей , :

, (3)

, (4)

Вычисления (2) – (4) 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), тэги, классы, авторы, цитируемые или цитирующие документы и т.п. В мультимодальной модели матрица определяется для каждой модальности, а матрица является общей. Оптимизационная задача представляет собой максимизацию взвешенной суммы лог-правдоподобий и r-регуляризаторов при условиях нормировки и неотрицательности столбцов матриц (для каждой модальности задается ее вес).

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. Всё это не позволяет сформировать стандарт исследовательской работы, который позволил бы различным научным группам сравнивать и воспроизводить результаты своих коллег. Таким образом, устранение данных препятствий позволяет сделать шаг в развитии экспериментальной экономики, усиливающий её доказательные способности.

Имеющиеся публикации, представляющие тематическую модель для той или иной задачи, не объясняют, как именно выбирается диапазон для перебора значений коэффициентов. Выбор оптимального коэффициента делается на основе значений критериев качества модели, однако неясно насколько сильно меняется состав топовых слов тем (для каждой модальности). Высокий барьер для входа в область тематического моделирования представителей смежных специалистов (экономистов, социологов, психологов) отмечается в ряде отечественных и зарубежных работ.

3. **Формализация алгоритма АРТМ для проведения мультимодальной мягкой кластеризации документов с целью получения информации о структуре больших текстовых коллекций.**

Формализация строится с использованием описанных в работе Булатов (2020) относительных критериев регуляризации (относительный коэффициент показывает, во сколько раз регуляризатор влияет на оценку сильнее, чем коллекция).

Таблица 1 - Формализация алгоритма построения АРТМ для научных публикаций

|  |  |  |  |
| --- | --- | --- | --- |
| Шаг | Описание шага | Контролируемые меры качества | Критерии отбора |
| 1 | Отбор числа тем.  Исходя из общего понимания задачи определяется диапазон возможного числа тем. Внутри диапазона с равным шагом выбирается 4-5 значений числа тем. Например: 5, 10, 15, 20 тем (для определения подтем направления); число тем значительно увеличивается в случае расширения анализируемой области знаний.  Делается 10 проходов (1-10) по коллекции без регуляризации и с включенным регуляризатором декоррелирования. Тестируются несколько значений относительного коэффициента: 0.005, 0.01, 0.05, 0.1, 0.15. | Ключевые показатели: Cредняя когерентность (по 15 наиболее частотным биграммам)  Перплексия  Дополнительно контролируемые показатели:  Разреженность матриц и . | Выбираются два эксперимента, по которым были получены результаты с наибольшей когерентностью, из них выбирается эксперимент с наименьшей перплексией.  Особенность: Когерентность обратно коррелирует с перплексией |
| 2 | Подбор коэффициента сглаживания для фоновой темы (или отказ от него).  Перебор значений по сетке, 10 итераций. К выбранному числу тем и значению коэфф. декоррелирования добавить одну фоновую тему. Сетка для относительного коэффициента сглаживания [0.01, 0.05, 0.1, 0.15, 0.2] |
| 3 | Подбор коэффициентов разреживания.  Делается дополнительно 10 проходов (11-21 итерации), тестируется коэффициент разреживания для матриц и . Тестируемые значения относительного коэффициента для матрицы в диапазоне от -0.1 до -0.8 с шагом 0.1. Для каждого значения рассчитывается по формуле соответствующее значение абсолютного коэфф. разреживания для матрицы . |
| 4 | Анализ когерентности для выбранной стратегии регуляризации | Значения когерентности по темам, интерпретируемость тем | Если существуют не интерпретируемые и/или плохо интерпретируемые темы, проводится тестирование для близкого числа тем (±2 темы) |
| 5 | Выбранная стратегия регуляризации тестируется в окрестности выбранного числа тем (±2 темы) | Аналогично шагам 1-3. | |

1. **Предложен и апробирован подход формализованного алгоритма аддитивной регуляризации тематических моделей для мягкой мультимодальной кластеризации научных статей по поведенческой и экспериментальной экономике с целью получения информации о структуре.**

Подход включает следующие этапы: (1) определение источника научных знаний, сбор релевантных данных; (2) проведение предобработки данных, приведение в пригодный для анализа 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 достигаются на основе следующей метаинформации (включения следующих модальностей): 1) двусловных словосочетаний (биграмм) Названий и Резюме статей; 2) авторов статей; 3) списков использованной литературы. Веса модальностей предлагается брать равными 0.5 для слов, 1.0 для биграмм, авторов и ссылок.

На основе формализованного алгоритма осуществлена мягкая кластеризация статей.

Число слов - 36485, число биграмм – 497933, число авторов – 54287, число ссылок – 461695.

Выделено 15 тем. Итоговые параметры модели: 10 итераций без регуляризации, добавление на 11-20 итерациях (10 итераций) коэффициентов разреживания для матриц () и (). Значения коэффициентов:

, (в силу особенностей технической реализации для матрицы приведены абсолютные значения коэффициента).

Значение средней когерентности ; перплексия .

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 , , поддаются экспертному именованию. Без какого-либо экспертного вмешательства в состав тем, на основе алгоритма выделены и именованы следующие подтемы: *Ограниченная рациональность, теория перспектив; Про-экологическое 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*.

1. **Предложен и апробирован подход разведочного поиска к оценке программы импортозамещения на основе патентных данных с помощью применения метода аддитивной регуляризации тематических моделей.**

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 .

Итоговая модель обладала следующими метриками качества: доля разреженных элементов матриц униграмм , биграмм , . Размер ядра равен 628, средняя чистота тем 0,992, средняя контрастность тем 0,976. Общее число итераций: 40.

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 (1 – точно соответствует одному из заявленных пунктов Плана; 0.5 – связан с одним из пунктов; 0 – не соответствует). Для релевантных документов 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:

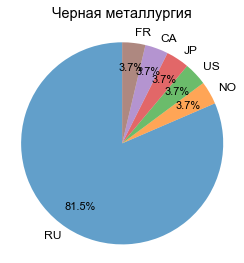
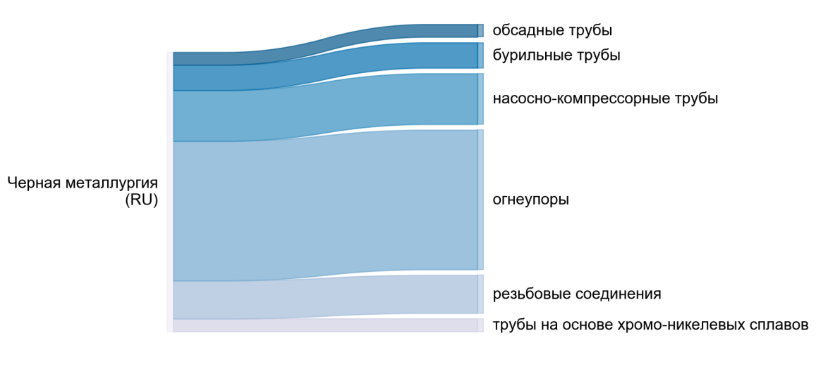


Рисунок 1 - Черная металлургия. Распределение по странам-патентообладателям и категориям патентования российскими патентообладателями

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: , 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 1.

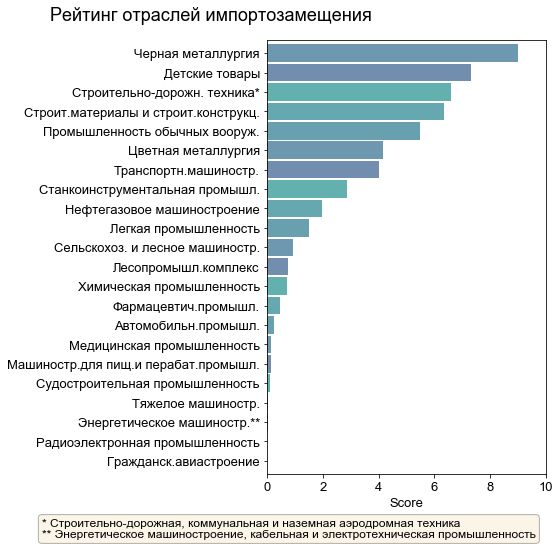


Рисунок 2 - Рейтинг отраслей импортозамещения на основании соответствия выданных патентов заявленному плану импортозамещения отрасли

1. **РЕЗУЛЬТАТЫ И ВЫВОДЫ**
2. В данной работе предлагается подход к подбору и анализу документов, противоположный доминирующему сегодня подходу (основан на предоставлении персонализированной информации, снижающий затраты на поиск, способствующий привлечению внимания). В основе лежат идеи (Г. Саймон, Ф. Хайек, К. Камерер), 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, 4) в целом, ограниченное внимание является важным понятием поведенческой экономики.

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 поведения индивидов, инструменты поведенческой экономики должны подсказывать, каким образом необходимо сконфигурировать среду для восприятия информации, чтобы 1) снизить затраты на поиск информации; 2) минимизировать сдвиги в восприятии (обращение внимания на продвигаемую, эмоциональную, первую попавшуюся информацию); 3) повысить познавательные функции, избежать деградации аналитического мышления. Применительно к текстовой информации, данным требованиям удовлетворяет подход аддитивной регуляризации тематических моделей. Таким образом, именно подход тематического моделирования как способ семантической компрессии текстовой информации может являться средством эффективного управления ограниченной рациональностью при получении знаний об исследуемой области.

1. Формализованный алгоритм аддитивной регуляризации тематических моделей представляет собой набор понятных инструкций, реализован в виде набора программ на языке Python. 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. Автоматизированное выделение подтем научной области позволяет 1) составить представление о структуре области; 2) дать представление о ключевых, опорных публикациях подтемы; 3) показать, какие авторы публикуются по каждой из подтем; 4) сделать тематическую подборку публикаций; 5) анализировать динамику развития подтем. Повышение познавательных функций исследователя оказывает существенный вклад в развитие и функционирование экономики знаний. 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 АРТМ.
5. Проведен анализ программы импортозамещения на основе патентных данных с использованием инструментария аддитивной регуляризации тематических моделей. Подход позволяет осуществлять поиск сразу по блокам априорно задаваемой информации (в данном случае – пункты сразу всех 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 релевантность применения методов семантической компрессии текстовой информации (с использованием графового алгоритма Textrank для выделения ключевых слов, выделения ключевых предложений на основе алгоритма bm25) для повышения эффективности управления вниманием с целью получения представления об исследуемой области (на примере анализа правительственных документов направления Цифровая экономика; нормативных документов, регламентирующих вопросы питания).

Более ранним исследованием, в рамках разработки инструментов по управлению ограниченной рациональностью, автором была разработана система для принятия многокритериальных решений на основе метода аналитических сетей/иерархий (Multichoice, свидетельство о регистрации программы для ЭВМ №2016616698).

**Возможные направления продолжения исследования**

Дальнейшее расширение практики использования формализованного алгоритма АРТМ в рамках экономики знаний (библиографические исследования, форсайт исследования), а также любых других исследований, ставящих перед собой задачу получения знаний из больших коллекций текстовых документов.

**Список работ, опубликованных по теме исследования**

**Работы, опубликованные в журналах Scopus, а также журналах из перечня ведущих периодических изданий ВАК Минобрнауки РФ по специальности 08.00.13 – Математические и инструментальные методы экономики**

Милкова М.А. (2020). Инновационный подход к поиску информации на примере патентного анализа плана импортозамещения // Экономическая наука современной России, 1(88). С. 143-157.

Milkova M.A. (2021). 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

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

**Работы, опубликованные в других журналах:**

Милкова М.А. (2021). Тематическое моделирование: восприятие научной информации // Цифровая экономика, 14(2), 31-36. DOI: 10.34706/DE-2021-02-04

Милкова М.А. (2021). Информация и ограниченная рациональность выбора в цифровой экономике // Цифровая экономика, 13(1), 69-88. DOI: 10.34706/DE-2021-01-08

Милкова М.А. (2020). 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. - c. 117-120

Милкова М.А. (2020). Феномен внимания в информационной среде: экономика внимания // Цифровая экономика, 3(11), с. 73-87. DOI: 10.34706/DE-2020-03-08

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

Милкова М.А. (2019). Тематические модели как инструмент «дальнего чтения» // Цифровая экономика, 1(5), с. 57-70. DOI:10.34706/DE-2019-01-06

Милкова М.А. (2018). Извлечение ключевых терминов направления «Цифровая экономика»: графоориентированный подход // Цифровая экономика, 4(4), 57-65.

**Переводы, выполненные и опубликованные автором по теме диссертации:**

Санстейн, К.Р. (2020). Пренебрежение вероятностью: эмоции, наихудшие случаи и право // Цифровая экономика, 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. (2020). За пределами денег и информации: экономика внимания // Цифровая экономика, 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. (2020). Оплата славой: как неэпистемологические мотивы способствовали феноменальному успеху современной науки // Цифровая экономика, 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:**

Милкова М.А., Неволин, И.В., Пигорев, Д.П. (2021). Извлечение ключевой информации из нормативных документов о политике продовольствия и питания в России // Экономическая наука современной России, forthcoming

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

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

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

Милкова М.А., Андрейчиоква О.Н. (2016). Свидетельство о регистрации программы для ЭВМ №2016616698. Multichoice

1. Нобелевская премия по экономике в 2019 году - А. Банерджи, Э. Дюфло и М. Кремер за работу по адаптации метода рандомизированных контрольных испытаний (РКИ) к области экономического развития. [↑](#footnote-ref-1)
2. Peter P. Wakker. Annotated Bibliography – URL: <https://personal.eur.nl/wakker/refs/webrfrncs.pdf> (дата обращения: 01.06.2021) [↑](#footnote-ref-2)
3. 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) [↑](#footnote-ref-3)
4. Открытые реестры ФИПС // Федеральный институт промышленной собственности: <https://www.fips.ru/> – URL: <https://www.fips.ru/registers-web/> (дата обращения 22.04.2020) [↑](#footnote-ref-4)
5. 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) [↑](#footnote-ref-5)