УДК 519.711:001 DOI: http://doi.org/10.25728/pu.2024.3.4

ИНФОРМАЦИОННАЯ СИСТЕМА АНАЛИЗА НАУЧНОЙ ДЕЯТЕЛЬНОСТИ (ИСАНД) В ОБЛАСТИ ТЕОРИИ УПРАВЛЕНИЯ

Д. А. Губанов*, О. П. Кузнецов**, Е. А. Курако***, Д. В. Лемтюжникова****, Д. А. Новиков*****, А. Г. Чхартишвили******

*–******Институт проблем управления им. В.А. Трапезникова РАН, г. Москва,
****МАИ (национальный исследовательский университет), г. Москва

*✉ dmitry.a.g@gmail.com, **✉ olpkuz@yandex.ru, ***✉ kea@ipu.ru, ****✉ darabbt@gmail.com, *****✉ novikov@ipu.ru, ******✉ sandro_ch@mail.ru

Аннотация. Представлено описание подходов, лежащих в основе разрабатываемой в ИПУ РАН Информационной системы анализа научной деятельности (ИСАНД) в области теории управления. Описана онтология ИСАНД, ориентированная на представление и сбор знаний в области теории управления: как научного знания (онтология теории управления), так и знаний, связанных с научной деятельностью агентов в данной области (организаций, журналов, конференций и отдельных исследователей). Дана схема построенной на основе онтологии архитектуры ИСАНД как сложного программного комплекса, обеспечивающего сбор, хранение и анализ публикаций и их метаинформации, которые поступают из внешних источников. Описан алгоритм построения тематических профилей научных объектов (публикаций, ученых, организаций, журналов, конференций), описаны осуществляемые при помощи ИСАНД процессы обработки текстов и возможности сетевого анализа. Описаны основные возможности использования ИСАНД.

Ключевые слова: анализ научной деятельности, теория управления, информационная система, классификация, онтология, тематический профиль, тематическое пространство, термин, обработка текстов, анализ сетей.

ВВЕДЕНИЕ

Бурный рост числа научных публикаций, продолжающийся в течение последних десятилетий, породил спрос на разработку компьютерных систем, автоматизирующих работу с большими массивами публикаций. Любая такая система должна содержать базу публикаций и, соответственно, средства пополнения и сопровождения этой базы. Однако состав аналитических инструментов работы с публикациями в разных системах различен и определяется целями, которые ставят перед собой создатели таких систем. Хорошо известны базы Web of Science, Scopus, РИНЦ, Google Scholar, ResearchGate, OpenAlex и др., главная цель которых – анализ цитируемости публикаций, на основе которого вычисляются наукометрические оценки публикаций и их авторов (индекс Хирша), а также научных журналов (импакт-фактор).

Более сложными и менее исследованными являются задачи анализа содержания научных текстов. Систем такого рода гораздо меньше. Можно отметить американскую систему Semantic Scholar (www.semanticscholar.org), специализирующуюся на компьютерных науках и медицине, а также разработку НИУ ВШЭ iFORA (https://issek.hse.ru/ifora), которая, впрочем, работает не только с научными публикациями, но и с патентами, рыночной аналитикой и др.

Разрабатываемая в Институте проблем управления им. В. А. Трапезникова РАН (ИПУ РАН) *Информационная система анализа научной деятельности* (ИСАНД), тестовая версия которой доступна по адресу https://isand.ipu.ru, представляет

собой систему, ориентированную на анализ содержания научных публикаций по теории управления. База публикаций системы содержит массивы публикаций сотрудников ИПУ РАН с 2005 г., статьи журналов «Проблемы управления» за 2003–2023 гг., «Advances in Systems Science and Applications» (ASSA) за 2017–2020, 2022–2023 гг. и др., доклады конференций «Управление большими системами» за 2009 г., 2011–2018, 2021–2023 гг., «Управление развитием крупномасштабных систем» (MLSD) за 2007–2023 гг. и др.; в дальнейшем предполагается как существенное расширение этой базы за счет ретроспективной информации из других источников, так и регулярное ее пополнение и поддержание в актуальном состоянии.

В основе большинства задач, связанных с анализом содержания научных текстов, лежит позиционирование текстов в *тематическом пространстве*. Традиционные методы структурирования тематического пространства – *универсальные классификаторы* типа УДК [1], международного классификатора OECD [2], классификатора Российского научного фонда [3], ГРНТИ [4] и др. – не вполне соответствуют задачам, решаемым системой, по двум причинам. Первая причина – универсальность, достоинства которой оборачиваются недостатком: слишком крупным членением научных направлений и, соответственно, слишком слабой дифференциацией разделов этих направлений. Вторая причина – одномерность, связанная со строгим соблюдением таксономического принципа: каждый объект классификации должен характеризоваться ровно одной вершиной дерева классификатора. Это требование, во-первых, затрудняет классификацию междисциплинарных работ, а во-вторых, не дает возможности описать тот факт, что, например, два специалиста, работающих в одной и той же области, но использующих разный математический аппарат, по существу, имеют разные компетенции, которые в тематическом пространстве должны быть позиционированы поразному. Поэтому *классификатор ИСАНД*, разработанный в системе ИСАНД, является многомерным и основан на современных принципах построения онтологий. В его основе лежит предложенная в работе [5] трехмерная онтология наук об управлении. Подробное описание классификатора ИСАНД содержится в § 4.

Наличие классификатора, структурирующего тематическое пространство, позволяет характеризовать в терминах этого пространства основные *объекты* научной деятельности: публикации, научных сотрудников (исследователей), научные журналы, научные и научно-образовательные учреждения, научные конференции. Эти характеристики называются *профилями*, которым посвящен раздел «Тематические профили научных объектов». На основе построенных профилей в системе ИСАНД решаются конкретные задачи анализа, связанные с указанными объектами научной деятельности. Например, исследователя интересует поиск публикаций по заданной тематике; руководству учреждения может понадобиться поиск специалистов с указанными компетенциями, редакции научного журнала или организаторам конференции требуется компетентный рецензент для данного научного текста и т. д. Примеры таких задач приведены в § 6.

Отдельные разделы статьи посвящены описанию *архитектуры* системы ИСАНД, а также интеллектуальным *методам анализа текстов* и *сетей*, возникающих на базе объектов ИСАНД.

1. РАЗРАБОТКА ОНТОЛОГИИ ИСАНД

Онтология – формальная спецификация согласованного описания (концептуализации) предметной области (по Т. Груберу [6, 7]), разрабатываемая группой экспертов и интерпретируемая как машиной, так и человеком. Иными словами, онтология представляет собой формализованное описание согласованных экспертами понятий в определенной предметной области, разработанное для однозначного понимания как людьми, так и машинами. Web Ontology Language (OWL) [8, 9] – язык, предложенный консорциумом World Wide Web Consortium (W3C), который служит практическим средством для создания конкретных структурированных онтологий, позволяющих формализовать знания в некоторой предметной области с использованием классов, отношений, индивидов и логических ограничений. Онтологии OWL упрощают обмен информацией (как между людьми, так и между программными агентами), обеспечивают возможность повторного использования знаний, поддерживают вывод новых знаний и являются основой для построения баз знаний информационных систем, основанных на знаниях.

Разработка предметно-ориентированной OWL-онтологии в рассматриваемом случае включает в себя следующие этапы:

1. Анализ требований и сценариев использования информационной системы.

2. Создание основных классов и их атрибутов, отношений между классами, определение логических ограничений на классы и свойства.

3. Формализация в рамках выбранного языка OWL.

4. Валидация и тестирование онтологии.

5. Развертывание и интеграция онтологии.

6. Поддержка и обновление онтологии.

Разработка онтологии ИСАНД (как и непосредственно информационной системы) мотивирована запросами следующих потенциальных пользователей (*агентов*):

- *исследователи*,
- *редакции научных журналов* и *организаторы научных конференций*,
- *руководители* научных и образовательных организаций, подразделений и коллективов,
- *организаторы науки*.

Исследователям важна содержательная поддержка их научной работы, в том числе анализ актуальных научных направлений, изучение ключевых понятий и идентификация влиятельных агентов (исследователей, журналов, конференций, организаций) и научных публикаций.

Редакции журналов и организаторы конференций стремятся обеспечить соответствие представленных материалов журналу либо конференции, найти квалифицированных рецензентов и потенциальных участников конференции.

Руководители заинтересованы в поиске новых сотрудников и участников проектов, в анализе актуальности научных направлений внутри научной или образовательной организации.

Для организаторов науки актуальны вопросы, связанные с организационными структурами (организации, подразделения, научные коллективы и исследователи), с прогнозом и оценкой перспективности направлений научных исследований и эффективности деятельности агентов.

Таким образом, онтология ИСАНД ориентирована на представление и сбор знаний в области теории управления: как *научного знания* (онтология теории управления), так и знаний, связанных с *научной деятельностью* агентов в данной области (организаций, журналов, конференций и отдельных исследователей).

Рассмотрим онтологию научного знания и онтологию научной деятельности в теории управления.

1.1. Онтология научного знания (онтология теории управления)

Онтология научного знания предназначена для систематизации и классификации знаний в области теории управления. Предлагаемый классификатор (см. описание в работе [10]) является «системой координат» тематического пространства, позволяющей реализовать взгляд на совокупность научных направлений с определенной точки зрения, а также отразить возможную многотемность научного документа или многообразие компетенций ученого. Характеристикой объекта в этом пространстве является вектор, называемый профилем (см. § 3). Отметим, что этот классификатор частично отражен в публикации [11] (см. также работу [12]) и учитывает более ранние публикации по терминологии теории управления [13–15].

Онтология научного знания ИСАНД представляет собой существенно расширенную версию онтологии теории управления, предложенной в работе [5]. Она имеет четырехуровневую структуру, уровни которой (кроме нижнего) представляют собой дерево. Уровни нумеруются числами от 0 до 3. Нулевой уровень содержит четыре фиксированных вершины «Общенаучная проблематика», «Математический аппарат», «Предметная область», «Сфера применения». Предполагается, что при возможных расширениях онтологии этот уровень не изменяется. Он отражает не конкретные темы теории управления, а различные аспекты научных исследований: математический аппарат, используемый в исследовании (теорию игр, теорию вероятностей, …), предметную область, т. е. некоторую прикладную теорию (теорию автоматического управления, анализ данных, теорию управления в организационных системах, …) и конкретную сферу применения (подвижные объекты, производство, энергетику, финансы, медицину, …). Вершины нулевого уровня назовем *метафакторами*.

Каждая из вершин нулевого уровня является корнем тематического поддерева, раскрывающего ее содержание. Например, поддерево «Математический аппарат» содержит вершину «Теория игр» (первый уровень) и детализирующие ее вершины второго уровня: «Теория некооперативных игр» и др. Соответственно, поддерево «Предметная область» среди прочих содержит вершину «Теория управления в организационных системах» и детализирующие ее вершины второго уровня, напри

мер, «Механизмы планирования», а поддерево «Сфера применения» содержит вершины «Энергетика» (первый уровень) и «Атомная энергетика» (второй уровень). Каждый фактор нижнего (второго) уровня характеризуется фиксированным набором терминов.

Классификатор построен экспертами ИПУ РАН для теории управления, на данный момент он включает в себя четыре фактора нулевого уровня,

53 фактора первого уровня, 161 фактор второго уровня, более трех тысяч терминов (см. https://www.ipu.ru/sites/default/files/page_file/ClassifierCS.xlsx). Фрагмент графа классификатора приведен на рис. 1.

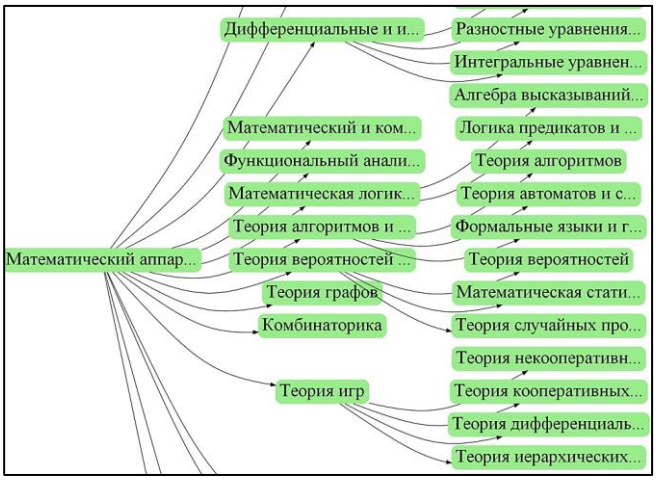


Рис. 1. Фрагмент графа классификатора

Ключевым терминам теории управления (около 1000 терминов) были даны определения и описания, в которых были поставлены гиперссылки на другие термины данной системы понятий – см. работы [11, 12] и https://www.ipu.ru/education/glossary.

1.2. Онтология научной деятельности

Онтология научной деятельности предназначена для описания агентов (организаций, сообществ, персон) и результатов их действий. Текущая версия онтологии включает в общей сложности 45 классов в таксономии (например, «Публикация»), 23 объектных свойства (например, «влияет на»), 37 простых свойств (например, «название»).

Онтология включает в себя девять классов верхнего уровня, в том числе классы «Агент», «Действие», «Результат», «Категория», «Роль» и «Профиль». На рис. 2 показана небольшая часть онтологии с основными классами и связями верхнего уровня (более полная версия – см. https://www.ipu.ru/sites/default/files/page_file/isand_ra_ontology.pdf). Класс «Агент» моделирует индивидуальные и коллективные сущности (организация, подразделение, научный коллектив и т. д.). Эти сущности выполняют действия (класс «Действие»), производя те или иные результаты (класс «Результат»). Подклассами класса «Агент» являются «Индивид» («Персона» и «ПрограммныйАгент»), «Организация», «Сообщество»,

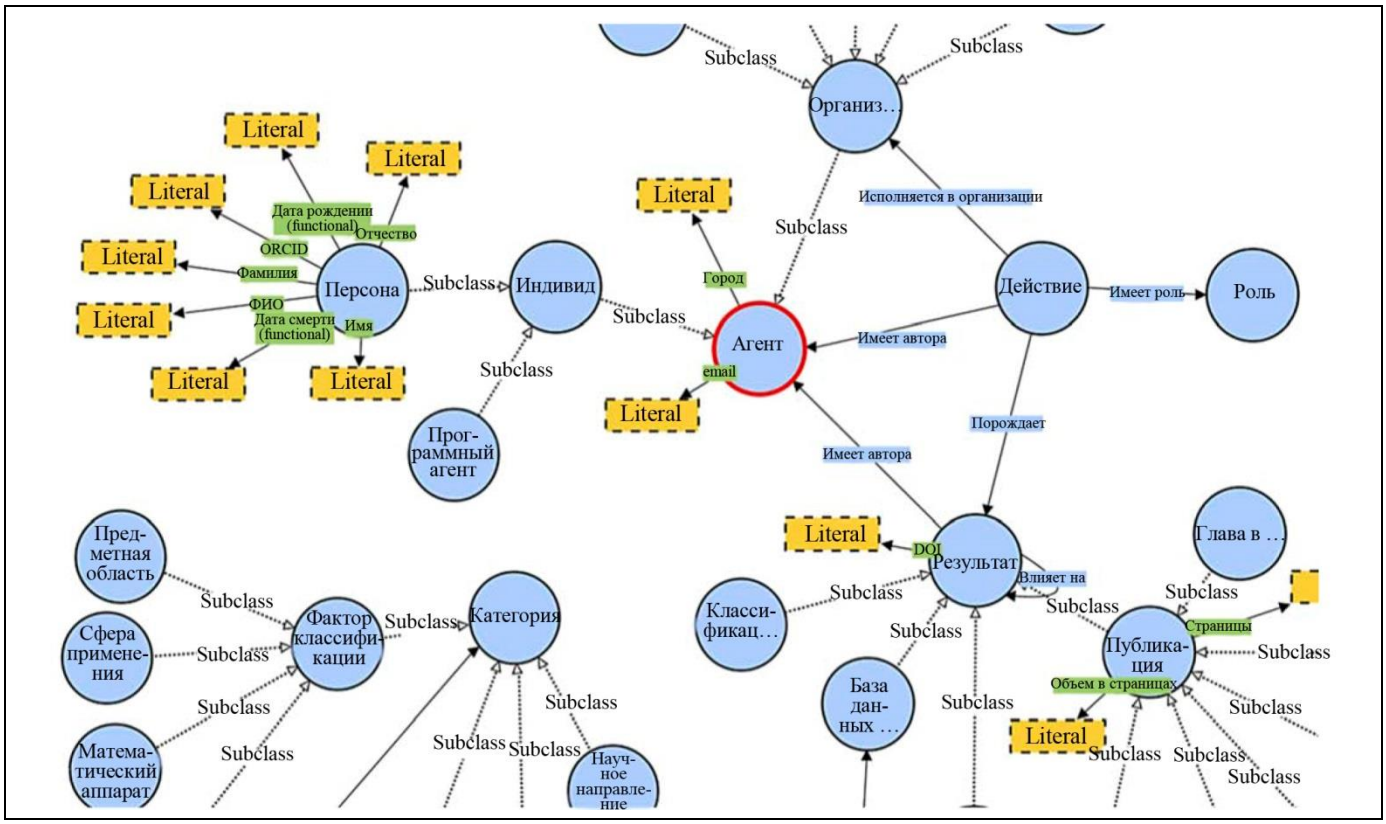


Рис. 2. Основные классы онтологии научной деятельности

«НаучныйКоллектив» и т. д. Класс «Организация» участвует в различных отношениях с другими типами сущностей, в частности класс связан через отношение «являетсяЧленом» с классом «Персона». Класс «Категория» позволяет тематически классифицировать другие сущности онтологии (например, результаты действий).

В целом классы в онтологии и связи между ними соответствуют методологии научной деятельности (см. монографию [16]).

1.3. Онтологии ИСАНД и разработка информационной системы

Онтологии являются основой базы знаний ИСАНД, реализованной в виде RDF-хранилища (*Resource Description Framework*). RDF – это стандарт W3C для описания метаданных ресурсов в сети интернет, позволяющий интегрировать и управлять данными из различных источников. Хранилище поддерживает обработку запросов на языке SPARQL, созданном специально для работы с RDF-данными и обеспечивающем гибкое управление семантикой и метаданными. Веб-приложения, использующие такое хранилище, могут легко адаптироваться к изменениям онтологической модели. Описание структуры данных и сами данные о конкретных экземплярах могут быть извлечены из хранилища одинаково эффективно. Эта база знаний занимает центральное место в архитектуре ИСАНД, описание которой приведено в следующем параграфе.

2. АРХИТЕКТУРА ИСАНД

ИСАНД представляет собой сложный программный комплекс, обеспечивающий сбор, хранение и анализ публикаций и их метаинформации, которые поступают из внешних источников.

Система работает с двумя внешними потоками информации. Первый поток – это загрузка из источников данных, таких как журналы, конференции, издательства и электронные библиотеки. Они предоставляют свою базу публикаций для загрузки в ИСАНД. В зависимости от возможностей источника получение данных может быть как разовым, так и регулярным. Второй поток – это взаимодействие пользователей с системой. Они могут добавлять и корректировать данные о своих публикациях, а также получать результаты поиска и анализа информации. Для этого предоставляются соответствующие методы сайта. Кроме того, предоставляется возможность воспользоваться электронными сервисами для получения данных для своих информационных систем.

Архитектура системы является многоуровневой и использует шаблон Request–Response для организации взаимодействия между компонентами. Основные подсистемы и связи между ними представлены на рис 3.

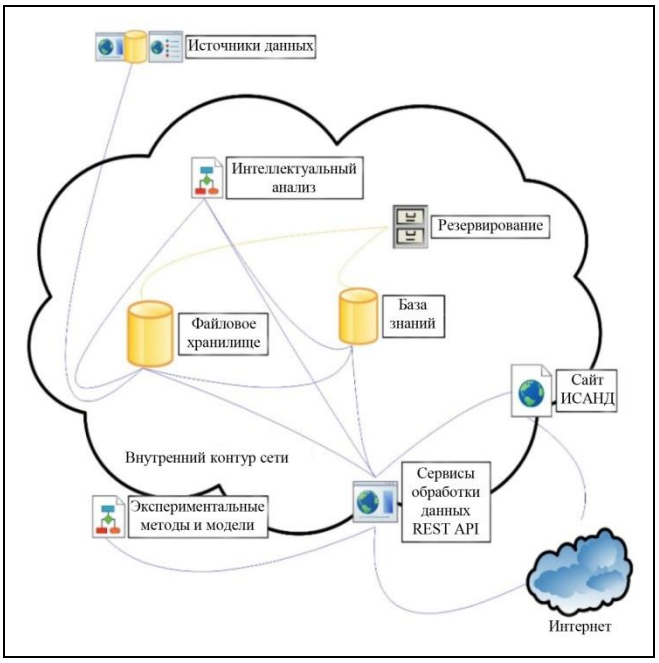


Рис. 3. Архитектура ИСАНД

В рамках внутреннего контура ИСАНД (см. рис. 3) можно выделить несколько основных подсистем:

- файловое хранилище и подсистема загрузки данных,
- база знаний,
- подсистема интеллектуального анализа информации,
- подсистема экспериментальных методов и моделей,
- подсистема резервирования,
- сервисы обработки данных,
- сайт ИСАНД.

Все файлы публикаций или файлы с метаинформацией о публикациях, в том числе и сборные архивы, поступают в подсистему загрузки данных, где они сохраняются в своем исходном виде и копируются в блок структурированных данных, где каждая публикация индексируется и хранится отдельно. В случае архива они автоматически разбираются на отдельные файлы. Публикации из блока

структурированных данных доступны остальным частям системы только для чтения.

Для каждого файла подсистема загрузки применяет методы из блока интеллектуального анализа информации, которые проверяют его целостность, язык и кодировку, а затем сегментируют публикацию на отдельные блоки. Такие методы можно назвать первичными. Из важных сегментов публикации можно отметить следующие: название, ключевые слова, авторы и библиографические ссылки. Дополнительно первичные методы анализа определяют состав и количество терминов из используемой онтологии научного знания (см. п. 1.1). Результаты работы данных методов сохраняются в структурированном блоке и также доступны остальным компонентам системы. Необходимо добавить, что ИСАНД – развивающаяся система, в которой эволюционируют модели анализа информации и происходит уточнение онтологии научного знания в части терминов и факторов классификации. После таких изменений первичные методы анализа системы будут выполнены повторно для всего структурированного архива.

Следует отметить, что часть источников данных предоставляет информацию не в виде файла с текстом публикации, а уже в сегментированном виде. Так как форматов данных много и представленные сегменты в них различны, то применяется другая часть первичных методов, которая для универсальности последующей обработки преобразует их в единый внутренний формат.

База знаний (БЗ) хранит информацию в графовом виде согласно модели представления данных RDF. Для определения сущностей и отношений между ними в БЗ используются онтологии научного знания и научной деятельности (см. пп. 1.1 и 1.2 соответственно). Для пополнения информации в БЗ используются результаты первичных методов обработки, расположенные в файловом хранилище во внутреннем формате. Обеспечивающее обновление базы ПО регулярно обращается в подсистему файлового хранилища и получает список новых поступлений. Так как данные поступают из разных источников, то периодически возникает ситуация, когда появляются дубликаты статей, загруженных ранее. Перед загрузкой происходит комплексная проверка идентичности поступающих сущностей уже существующим в БЗ, но точное сопоставление не может быть выполнено автоматически. Происходит это из-за неполноты получаемых данных. Если рассмотреть сущность «Персона», то ее экземпляры нельзя сопоставить только по фамилии, имени и отчеству, требуется дополнительная информация, например, электронная почта или идентификаторы внешних систем (в том числе ORCID iD, Scopus Author ID, ResearcherID). Часто дополнительной информации о «Персоне» в публикациях нет или она противоречива, поэтому в результате загрузки возникают дублирующие записи об одной и той же сущности. В связи с этими проблемами БЗ построена так, что все поступающие данные хранятся с учетом источника их поступления. В то же время одной из задач методов анализа информации является выявление дубликатов и объединение сущностей.

Подсистема анализа информации является ядром функционирования системы ИСАНД и помимо первичной загрузки и выявления дубликатов реализует вычисление тематических профилей (см. § 3), интеллектуальную обработку данных (см. § 4) и интеллектуальный анализ сетей (см. § 5). Реализован данный блок в виде сервисов, располагающихся на нескольких серверах.

Одной из важных задач ИСАНД является предоставление исследователям возможности реализовать свои новые модели и методы. Для этого существует специальный блок – подсистема экспериментальных методов и моделей. Данный блок имеет доступ к непубличным данным и методам, но работает в изолированной среде, т. е. результаты применения этих методов не попадают в основную систему. При этом в случае успешной реализации они могут быть перенесены уже в блок интеллектуального анализа информации.

Естественно, для сохранения накопленной обработанной информации реализована система резервного копирования. Информация из файлового хранилища и БЗ дублируется на отдельном сервере.

Для обеспечения общего доступа к данным системы и методам обработки информации реализован стандартизированный программный интерфейс API (*application programming interface*) на основе технологии REST (*representational state transfer*). Обращение к сервисам из внутреннего контура происходит по протоколу HTTP, а из внешнего по протоколу HTTPS.

Также через REST API работает сайт ИСАНД, который предоставляет пользователю удобный графический доступ в систему. Сайт позволяет просматривать метаинформацию о публикациях и результаты работы методов анализа, а также редактировать информацию о публикациях в зависимости от прав, полученных при регистрации.

3. ТЕМАТИЧЕСКИЕ ПРОФИЛИ НАУЧНЫХ ОБЪЕКТОВ

Классификатором в системе ИСАНД является онтология научного знания теории управления, описанная ранее в § 1, которая позволяет отразить возможную многотемность *научных объектов*. Под научными объектами в данном разделе будем иметь в виду либо агентов (ученых, организации, журналы, конференции), либо публикации. Характеристикой каждого из этих объектов в тематическом пространстве является вектор, называемый (тематическим) *профилем*.

Напомним (см. § 1), что онтология теории управления имеет четырехуровневую структуру (метафакторы, факторы, подфакторы, термины). Уровни нумеруются числами от 0 до 3.

Множество вершин первого уровня, называемых факторами, обозначим через $V = \{v_1, \ldots, v_n\}$. При этом i-я вершина первого уровня связана с множеством $V_i = \{v_{i1}, \ldots, v_{in_i}\}$ вершин второго уровня – подфакторов. Обозначим через m общее число подфакторов: $m = \sum_{i \in N} n_i$.

Третий уровень – это вершины-термины, характеризующие подфакторы. Каждый термин, как правило, характеризует один подфактор (т. е. в некоторых случаях древовидность онтологии может нарушаться).

Приведем далее **алгоритм расчета профилей научных объектов** в соответствии с работой [10]. Обозначим:

K – множество ученых;

L – множество публикаций;

Δ_{lij} – сумму числа вхождений в l-ю публикацию базовых терминов ij-го подфактора;

$$\omega(k, l) = \begin{cases} 1, \text{ если } k\text{-й ученый является автором} \\ \quad l\text{-й публикации;} \\ 0, \text{ в противном случае;} \end{cases}$$

$r(l)$ – количество авторов l-й публикации.

В соответствии с алгоритмом, описанным в работе [10], определим *профиль второго уровня публикации* l:

$$x_l = \left(x_{l1}, \ldots, x_{lij}, \ldots, x_{lnm}\right),$$

где $x_{lij} = \dfrac{\Delta_{lij}}{\sum_{i \in N} \sum_{j \in N_i} \Delta_{lij}}$, $l \in L$, $j \in N_i$, $i \in N$.

Очевидно, что этот вектор является стохастическим, т. е. $\sum_{i,j} x_{lij} = 1$.

Замечание. В дальнейшем возможно рассмотрение более сложных методов определения профиля (в том числе основанных на сетевых связях (ссылках) публикаций).

Для нахождения *профиля первого уровня публикации* l просуммируем для каждого фактора значения компонент профиля второго уровня, отвечающих связанным с ним подфакторам:

$$X_l = \left(X_{l1}, \ldots, X_{li}, \ldots, X_{ln}\right),$$

где $X_{li} = \sum_{j \in N_i} x_{lij}$, $l \in L$, $i \in N$.

Наконец, для нахождения *профиля нулевого уровня публикации* просуммируем для каждой из трех вершин нулевого уровня значения компонент профиля первого уровня, отвечающих связанным с ней вершинам первого уровня.

Таким образом, каждая публикация характеризуется трехмерным вектором профиля нулевого уровня, n-мерным вектором профиля первого уровня, m-мерным вектором профиля второго уровня. Все три вектора являются стохастическими.

На основании профилей публикаций можно определить профили других научных объектов, связанных с публикациями.

На основе аддитивного принципа агрегирования определим *профили второго и первого уровня* k-го *ученого*, используя массив его публикаций

$$y_{ij}^k = \frac{\sum_{l \in L} \omega(k, l) \dfrac{x_{lij}}{r(l)}}{\sum_{i \in N} \sum_{j \in N_i} \sum_{l \in L} \omega(k, l) \dfrac{x_{lij}}{r(l)}}, \quad k \in K, \quad j \in N_i, \quad i \in N,$$

$$Y_i^k = \sum_{j \in N_i} y_{ij}^k, \quad k \in K, \quad i \in N.$$

Профиль нулевого уровня определяется суммированием для каждой из трех вершин нулевого уровня значений компонент профиля первого уровня, отвечающих связанным с ней вершинам первого уровня.

Далее определим профили *журнала*, в котором опубликованы работы ученых. Пусть $U \subset L$ – множество работ, опубликованных в журнале $p \in P$, где P – множество журналов. Тогда профили определяются по формулам

$$w_{ij}^p = \frac{\sum_{l \in U} x_{lij}}{\sum_{i \in N} \sum_{j \in N_i} \sum_{l \in U} x_{lij}}, p \in P, \quad j \in N_i, \quad i \in N,$$

$$W_i^p = \sum_{j \in N_i} w_{ij}^p, \quad p \in P, \quad i \in N.$$

Размерности профилей журнала также равны m (для профиля второго уровня) и n (для профиля первого уровня).

Наряду с тематическими профилями важной характеристикой журнала является количество опубликованных в нем работ, т. е. количество элементов в множестве U.

Замечание. Аналогично профилю журнала можно рассчитать профиль *научной конференции*.

Поскольку профили публикации, ученого, организации, журнала, конференции могут быть заданы в виде стохастических векторов, можно единообразно рассчитывать степень близости между этими научными объектами. Предлагается применять следующее расстояние между двумя профилями, задаваемыми стохастическими векторами $\alpha = (\alpha_1, \ldots, \alpha_n)$ и $q = (\beta_1, \ldots, \beta_n)$:

$$d(\alpha, \beta) = 1 - \sum_{j=1}^{n} \min(\alpha_j, \beta_j) = \frac{1}{2} \sum_{j=1}^{n} |\alpha_j - \beta_j|.$$

Отметим, что эта метрика является частным случаем хорошо известного в теории вероятностей расстояния общей вариации и принимает значения от 0 до 1 включительно.

Замечание. Легко убедиться (см. работу [10]), что в данной метрике расстояние между профилями первого уровня всегда не больше расстояния между профилями второго уровня и не меньше расстояния между профилями нулевого уровня тех же объектов.

4. ИНТЕЛЛЕКТУАЛЬНАЯ ОБРАБОТКА ТЕКСТОВ

4.1. Выделение структуры научных публикаций

На данный момент функционирование ИСАНД связано с двумя процессами обработки текстов: выделение метаинформации из текста публикации и предобработка текстового слоя для вычисления профилей (см. § 3).

Задача автоматического выделения структуры научных публикаций возникает при необходимости систематизировать и нормализовать накопленные данные с разными целями: формирование базы данных публикаций с возможностью поиска по ней, построение графов цитирования по библиографическим ссылкам, использование размеченных данных для обучения языковых моделей. Одна из главных проблем для решения этой задачи – большая разрозненность в структуре самих публикаций. Это может быть и разная последовательность структурных элементов, и отсутствие каких-либо из них, и разный формат написания в рамках одного структурного элемента (к структурным элементам относят идентификаторы, заголовок статьи, авторов, аннотацию и т. д.).

Методы выделения структуры текста могут основываться на традиционных алгоритмах OCR (*Optical Character Recognition*). Они представляют из себя механическое или электронное преобразование документов в пригодные для редактуры и поиска данные. К этим алгоритмам относятся: метод шаблонов, граничный анализ, зональная сегментация, структурный метод. Однако данные подходы не являются автоматическими и требуют значительного вмешательства для настройки и корректировки под различные форматы научных публикаций.

Методы автоматического выделения структуры в большинстве случаев ориентированы исключительно на использование технологий машинного обучения, поскольку для обеспечения высокой эффективности эвристических методов требуется разработка множества правил, учитывающих все возможные особенности каждого типа структурных элементов. Важно отметить, что эти подходы не всегда обеспечивают высокую точность результата, которая также может зависеть от языка, на котором написана соответствующая публикация [17, 18].

В работе [19] авторы представили подход по извлечению метаданных из заголовков документов с кириллическими символами. Подход включает в себя: создание датасета CORE, извлечение текста из PDF-файла с помощью утилиты pdfMiner с последующей токенизацией и обучение моделей GROBID (*GeneRatiOn of BIbliographic Data*) и BiLSTM (*Bidirectional Long Short-Term Memory*) для сравнения результатов. Набор данных CORE предоставляет данные научных публикаций. Он состоит из метаданных и полных текстов в машинно-обрабатываемом формате. Датасет, основанный на ресурсах PubMed Central Open Access Subset, CiteSeer и Cora-ref [20], состоит из 15 553 документов, полученных после фильтрации всех кириллических исходных данных по языку, удаления дубликатов, отбрасывания тех документов, которые не являются научными публикациями.

Далее рассмотрим, как структура научных публикаций определяется в ИСАНД. Здесь в качестве источников данных выступают публикации из хранилища. Общее количество статей в нем на момент написания настоящей работы составляло

26 335, исключая дубликаты и подозрительные статьи. После исключения файлов с шифрованием, поврежденной кодировкой и отсутствием текстового слоя осталось 22 532 статьи, включая 21 520 статей на русском языке.

Автоматическое извлечение структурных элементов научных публикаций основывается на инструменте с открытым исходным кодом GROBID, который представляет собой свободно распространяемую библиотеку, обученную на англоязычных публикациях для автоматического извлечения структурных элементов, и имеет возможность дообучения на корпусах текста других языков. GROBID использует каскад моделей разметки последовательностей для анализа документа. Такой модульный подход позволяет адаптировать обучающие данные, функции, текстовые представления и модели к различным иерархическим структурам документа. Предлагаемая модель является расширением задачи распознавания именованных сущностей [21]. По умолчанию решение задачи выполняется с помощью стандартных методов «плоского» машинного обучения, базирующихся на линейном цепном методе условных случайных полей (англ. *Conditional Random Field*, CRF). Однако в GROBID можно использовать модели маркировки последовательностей глубокого обучения, обученные с помощью библиотеки Deep Learning Framework for Text (DeLFT). DeLFT — это платформа Keras и TensorFlow для обработки текста, ориентированная на маркировку последовательностей и классификацию текста, которая реализует стандартные современные архитектуры глубокого обучения, соответствующие задачам обработки текста.

Доступные нейронные модели включают комбинацию методов типа CRF и глубокой двунаправленной нейронной сети долгой краткосрочной памяти BidLSTM. Данная комбинация методов BidLSTM–CRF используется со встраиванием модели глобальных векторов для представления слов (англ. *Global Vectors for Word Representation*, GloVe), с дополнительным каналом функций, с эмбедингами из языковой модели (ELMo) и точно настроенными архитектурами на основе трансформеров со слоем активации CRF или без него, которые могут использоваться в качестве альтернативы линейного цепного метода CRF.

В настоящее время для полнотекстовых моделей не существует нейронной модели, поскольку входные последовательности для этой модели слишком велики для поддерживаемых в настоящее время архитектур глубокого обучения. Для этой задачи постановку проблемы необходимо изменить или использовать альтернативные архитектуры глубокого обучения (со скользящим окном и т. д.).

Для разбора научной публикации GROBID использует каскад моделей маркировки последовательностей, представленный на рис. 4. Архитектура и параметры структурных элементов зависят от используемых меток, от объема доступных обучающих данных, от времени выполнения, ограничений на память и точность и т. д.

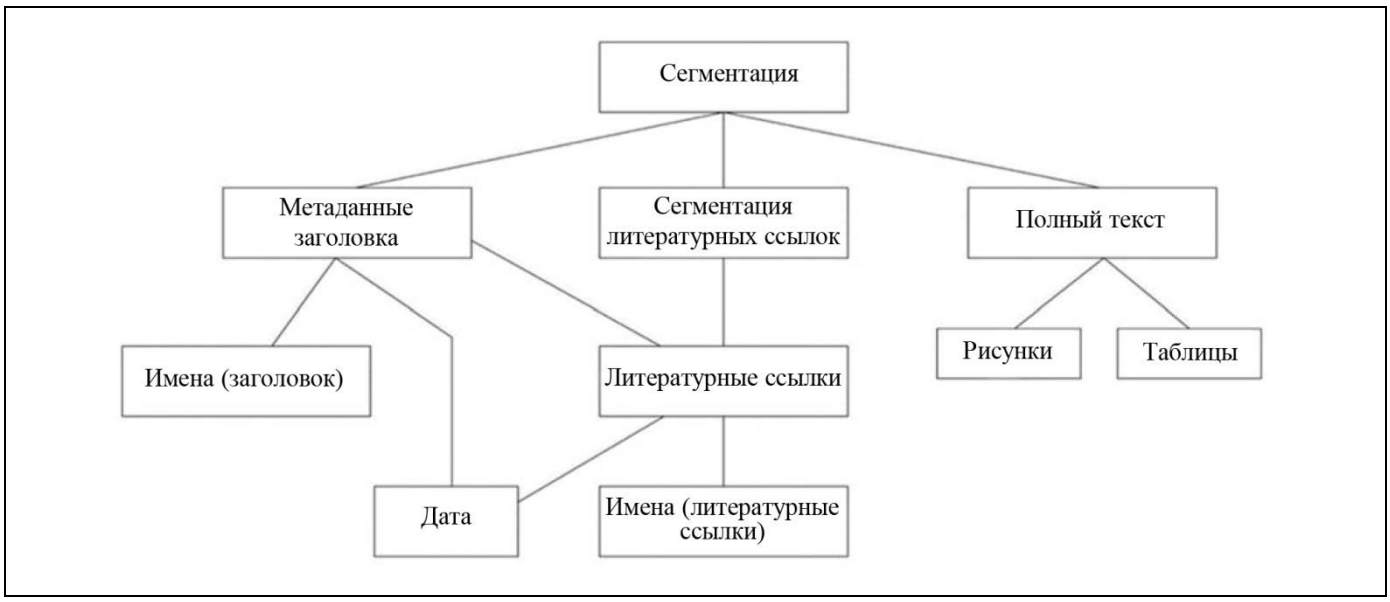


Рис. 4. Каскад структурных элементов в GROBID

Модель *Сегментация* используется для определения основных структурных элементов документа, таких как титульная страница, заголовок, основная часть, сноски, библиографические разделы и т. д. Обнаруженные моделью сегментации области заголовков передаются в модель заголовков. Модель заголовков обучается распознавать такую информацию, как название, авторы, принадлежность, аннотация и т. д. Некоторые модели маркировки могут использоваться в нескольких местах документа. Например, модель даты, используемая для сегментации исходной даты на годы, месяцы и т. д. и обеспечение нормализации даты в соответствии со стандартом ISO, вызывается, когда даты идентифицируются не только в области заголовка, но и при разборе зоны ссылок. Аналогично модели рисунков или таблиц применяются для структурирования всех рисунков и таблиц документа. Структурирование одного и того же типа сущности зависит также и от позиции этой сущности. Например, в шапке статьи обычно указываются полные имена авторов и они связаны с маркерами аффилиации, а имена авторов в строке ссылок обычно намного короче и никогда не смешиваются с информацией об аффилиации.

Метод обучения библиотеки GROBID, примененный при создании ИСАНД, основывается на предположении о стабильности структуры статей журналов в течение времени. В соответствии с этим предположением произведен анализ статей из разных источников с последующей группировкой и выявлением наиболее часто встречающихся шаблонов. Начальная фаза обучения включает 400 таких шаблонов, применяемых для обучения моделей, в том числе следующие шаблоны: «Метаданные заголовка», «Сегментация», «Адрес организации», «Имена (заголовок)» и «Сегментация литературных ссылок». В ходе экспериментов использовались PDF-файлы статей для создания обучающих данных при помощи инструмента GROBID. Внесение корректировок вызывало трудности из-за ограничений в методике исправления обучающих файлов. Если в обучающем файле отсутствовала часть текста из статьи, ее добавление было невозможно, что делало такой файл непригодным для обучения. В результате некоторые шаблоны были исключены из тренировочного набора в связи с вышеуказанными ограничениями. Дообучение GROBID происходит на основе предварительно проаннотированных обучающих данных. Каждая статья для дообучения поставляется с PDF-файлом, набором предварительно аннотированных XML-файлов и набором файлов без расширения XML, содержащих список токенов с характеристиками для обучения. После завершения процесса обучения были получены значения метрики *f*1 (показанные в таблице), которые отражают эффективность моделей.

Метрика *f*1 для модели GROBID

Модель GROBID	*f*1 (микро-усреднение)	*f*1 (макро-усреднение)
Метаданные заголовка	80,95	73,25
Сегментация	83,95	70,33
Адрес организации	82,71	78,29
Имена (заголовок)	93,97	87,71
Сегментация литературных ссылок	91,18	81,01

Для понимания значений метрики *f*1 рассмотрим несколько примеров.

Пример. 1. Успешное извлечение.

- **Исходная строка:** «Влияние квантовых эффектов на проводимость наноструктур. Иванов А.А., Петров Б.Б., Сидорова В.В. - Московский физико-технический институт, Долгопрудный, Россия.»
- **Распознанная строка:**
  - **название:** «Влияние квантовых эффектов на проводимость наноструктур»;
  - **авторы:** «Иванов А.А.», «Петров Б.Б.», «Сидорова В.В.»;
  - **аффилиация:** «Московский физико-технический институт, Долгопрудный, Россия».
- **Токены:**
  - **слова:** «Влияние», «квантовых», «эффектов», «на», «проводимость», «наноструктур», «Иванов», «А.А.», «Петров», «Б.Б.», «Сидорова», «В.В.», «Московский», «физико-технический», «институт», «Долгопрудный», «Россия» (17 токенов);
  - **знаки препинания:** «.», «,», «,», «-», «,», «,», «.» (7 токенов);
  - **итого:** 17 (слова) + 7 (знаки препинания) = 24 токена.
- **Анализ:**
  - **TP** (истинные положительные): все токены, составляющие название статьи, имена авторов, название института, город и страну, правильно отнесены к соответствующим категориям.
  - **FN** (ложные отрицательные): отсутствуют.
  - **FP** (ложные положительные): отсутствуют.
  - **TN** (истинные отрицательные): все знаки препинания, кроме двух запятых в аффилиации, классифицированы как «не относящиеся к метаданным».

○ **Precision** (точность): TP / (TP + FP) = 100 %.

○ **Recall** (полнота): TP / (TP + FN) = 100 %.

○ ***F*1 score:** 2 * (1 * 1) / (1 + 1) = 100 %. ♦

Первый пример демонстрирует идеальный случай, где все элементы, включая название статьи, имена авторов, название института, город и страну, были извлечены и классифицированы верно. Это соответствует *f*1 score 100 %, указывая на высокую точность модели.

Пример 2. Частичная ошибка.

• **Исходная строка:** «Новые подходы к машинному обучению. Автор: Смирнова Е.Д., Институт системного программирования РАН, Москва, Россия».

• **Распознанная строка:**

○ **название:** «Новые подходы к машинному обучению»;

○ **авторы:** «Смирнова Е.Д.»;

○ **аффилиация:** «Институт системного программирования РАН» (город и страна пропущены).

• **Токены**:

○ **слова:** «Новые», «подходы», «к», «машинному», «обучению», «Автор», «Смирнова», «Е.Д.», «Институт», «системного», «программирования», «РАН», «Москва», «Россия» (14 токенов);

○ **знаки препинания:** «.», «:», «,», «,», «,» (5 токенов);

○ **итого:** 14 + 5 = 19 токенов

• **Анализ:**

○ **TP:** Токены в названии, имени автора и названии института классифицированы верно.

○ **FN:** Токены «Москва» и «Россия» ошибочно отнесены к категории «не относящиеся к метаданным».

○ **FP:** Отсутствуют.

○ **TN:** Остальные токены (знаки препинания, слово «Автор:») классифицированы как «не относящиеся к метаданным» – верно.

○ В исходной строке 19 токенов относятся к метаданным, из них 11 распознаны верно (TP), 2 пропущены (FN).

○ **Precision:** TP / (TP + FP) = 11 / (11 + 0) = 100 %.

○ **Recall:** TP / (TP + FN) = 11 / (11 + 2) ≈ 84,62 %.

○ ***F*1 score:** 2 * (1 * 0.8462) / (1 + 0.8462) ≈ 91,67 %. ♦

Второй пример иллюстрирует частичную ошибку, где город и страна в аффилиации были пропущены. Несмотря на эту ошибку, *f*1 score остается достаточно высоким (около 91,67%), так как бóльшая часть информации извлечена корректно.

Пример. 3. Ошибка в определении элемента.

• **Исходная строка:** «Keywords: Deep learning, Natural language processing, Text analysis».

• **Распознанная строка:**

○ **название:** «Keywords: Deep learning, Natural language processing, Text analysis»;

○ **авторы:** (не определены);

○ **аффилиация:** (не определена).

• **Токены:**

○ **слова:** «Keywords», «Deep», «learning», «Natural», «language», «processing», «Text», «analysis» (8 токенов);

○ **знаки препинания:** «:», «,», «,» (3 токена);

○ **итого:** 8 + 3 = 11 токенов.

• **Анализ:**

○ **TP:** Отсутствуют.

○ **FN:** Все токены, включая «Keywords:», «Deep», «learning» и т. д., неправильно классифицированы. Они должны были быть отнесены к категории «ключевые слова» или «не относящиеся к метаданным».

○ **FP:** Все токены ошибочно отнесены к категории «название».

○ **TN:** Отсутствуют, так как все токены отнесены к какой-либо категории.

○ В исходной строке 11 токенов. TP = 0, все токены отнесены к неправильной категории (FN = 11, FP = 11).

○ **Precision:** TP / (TP + FP) = 0 / (0 + 11) = 0 %.

○ **Recall:** TP / (TP + FN) = 0 / (0 + 11) = 0 %.

○ ***F*1 score:** 2 * (0 * 0) / (0 + 0) = 0 %. ♦

Третий пример показывает критическую ошибку: строка с ключевыми словами была ошибочно распознана как название статьи, а авторы и аффилиация не были определены вовсе. В этом случае *f*1 score равен 0 %, что указывает на неспособность модели справиться с задачей.

Пример 4. Пропуск информации.

• **Исходная строка:** «Применение методов анализа данных в медицине. А. Петров, Б. Иванов. – Научно-исследовательский институт имени Н.И. Пирогова».

• **Распознанная строка:**

○ **название:** «Применение методов анализа данных в медицине»;

○ **авторы:** «А. Петров», «Б. Иванов»;

○ **аффилиация:** (не определена, название института ошибочно не отнесено к аффилиации);

• **Токены:**

○ **слова:** «Применение», «методов», «анализа», «данных», «в», «медицине», «А.», «Петров», «Б.», «Иванов», «Научно-исследовательский», «институт», «имени», «Н.И.», «Пирогова» (15 токенов);

○ **знаки препинания:** «.», «,», «.», «-» (4 токена);

○ **итого:** 15 + 4 = 19 токенов.

• **Анализ:**

○ **TP:** название статьи, имена авторов.

○ **FN:** Все токены названия института.

○ **FP:** Отсутствуют.

○ **TN:** Остальные токены (знаки препинания, точка) классифицированы как «не относящиеся к метаданным» – верно.

○ Допустим, в исходной строке 19 токенов относятся к метаданным, из них 10 распознаны верно (TP), 5 пропущены (FN).

○ **Precision:** TP / (TP + FP) = 10 / (10 + 0) = 100 %.

○ **Recall:** TP / (TP + FN) = 10 / (10 + 5) = 66,67 %.

○ **F1 score:** 2 * (1 * 0,6667) / (1 + 0,6667) = 80 %. ♦

Четвертый пример демонстрирует случай пропуска информации, где название института ошибочно не отнесено к аффилиации. Это приводит к снижению *f*1 score до 80 %, подчеркивая важность корректной классификации элементов метаданных.

Доступ к заголовкам и аннотациям в базе статей позволил провести анализ их сходства с применением метрик, таких как расстояние Джаро – Винклера, Левенштейна и косинусное расстояние. Анализ результатов подтвердил высокую точность соответствия заголовков и аннотаций в рамках выбранных метрик. В частности, модели, такие как «Метаданные заголовка» и «Сегментация», достигли значительных успехов. Тем не менее, другие модели, в частности «Адрес организации», «Имена (заголовок)» и «Сегментация литературных ссылок», выявили потребность в дополнительном улучшении для достижения желаемых результатов.

4.2. Выделение именных групп и кореференция

Предобработка текстового слоя для вычисления профилей состоит из операций преобразования слов в нижний регистр, лемматизации (приведения слов к нормальной форме), удаления стоп-слов (слов, знаков, символов, которые самостоятельно не несут никакой смысловой нагрузки), подготовки общего словаря для всех документов, преобразования слов в векторы (с помощью фреймворка pytorch), с которыми умеет работать нейронная сеть. Особенный интерес вызывает задача кореференции, которую необходимо решать для обеспечения полноты профиля, т. е. для того, чтобы учитывались не только все упоминания термина, но также и его косвенные упоминания, когда вместо него в тексте используется местоимение или синоним.

Стандартом решения многих задач обработки текстов, связанных с классификацией слов в тексте, является использование языковых моделей, проводящих токенизацию входного текста по словам. Это следует из того, что интуитивно проще классифицировать слово, когда оно представлено только одним токеном. Ввиду большого числа слов в используемом словаре такие языковые модели требуют значительных затрат памяти и вычислительных ресурсов. Для языков с богатой морфологией модели должны хранить информацию о каждой возможной словоформе каждого слова, что увеличивает размер словаря в среднем в два десятка раз. В качестве альтернативного подхода используется токенизация текста по наборам подряд стоящих символов, называемых подсловами (*subword* или *word pieces*). Это позволяет модели оперировать словарем ограниченных размеров [22]. Однако при использовании такой стратегии токенизации необходимы дополнительные механизмы объединения векторных представлений нескольких токенов, соответствующих одному слову [23].

Задача кореференции – это задача обработки естественных языков. В заданном тексте устанавливаются группы именных групп (слов или словосочетаний), обозначающих один и тот же объект [24]. Предполагается, что задача может быть решена более точно при использовании токенизации по подсловам. Задача осложняется необходимостью классифицировать не слова, а именные группы – группы подряд стоящих слов [25]. В ИСАНД применяется решение задачи кореференции при использовании токенизации по подсловам путем вычисления для каждой пары токенов двух оценок. Первая оценка выражает вероятность того, что два токена входят в одну именную группу. Вторая оценка выражает вероятность того, что два токена входят в две разные кореферентные именные группы. Совмещение двух оценок позволяет получить модель решения задачи кореференции, которая наследует все преимущества моделей с токенизацией по подсловам: меньший размер модели, более точная работа с языками с богатой морфологией.

Модель кореференции основана на том, что текст на естественном языке представляет собой описание действий или состояний различных объектов. Именная группа – это словосочетание, ссылающееся на объект внеязыковой действительности, называемый референтом. Именные группы обычно выражаются в тексте последовательностью из существительного и синтаксически подчиненных ему слов. Если две именные группы ссылают-

ся на один и тот же референт, то они называются кореферентными. Задача кореференции заключается в поиске всех пар кореферентных именных групп.

Первое полученное решение кореференции было основано на предположении, что для большей части пар именных групп в тексте можно однозначно определить наличие или отсутствие кореферентной связи при помощи системы правил. Полученная система правил отсеивала 71 % пар именных групп, однозначно определяя для них наличие или отсутствие кореференции. Для остальных пар использовалась типичная для данной задачи стратегия сравнения векторов признаков двух именных групп. В этот вектор кодировалась информация о положении именной группы в тексте, грамматических и синтаксических признаках их главных слов, а также некоторая другая. Нейронная сеть, состоящая из набора полносвязных слоев, определяла итоговую оценку вероятности наличия кореферентной связи. Данный поход имеет следующие недостатки: он опирается на сторонние решения синтаксического и морфологического анализа; набор заданных на этапе создания модели признаков именных групп, по которым определяется кореференция, может быть неполон.

В данный момент исследуется подход, основанный на выставлении каждой паре токенов текста оценки вероятности того, что оба токена находятся или в одной именной группе, или в двух кореферентных именных группах. Оценка основана на модификации механизма внимания self-attention и использует только векторные представления токенов для принятия решения. Такой подход позволяет решать одновременно и задачу определения именных групп в тексте, и задачу кореференции между ними. В настоящее время подход имеет следующие недостатки: возможность работы только в ограниченном окне токенов, необходимость обучения на больших корпусах и нестрогое покрытие именных групп: только часть токенов кореферентных групп получает высокую оценку. Построенная модель обладает высокой точностью, но не является достаточно полной. Это указывает на то, что модель находит только малую часть правильных пар токенов, но почти не ошибается.

5. ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ СЕТЕЙ

Научная деятельность порождает немалое количество объектов (публикаций, авторов, организаций, журналов и т. д.), связанных между собой различными связями (см. § 1) и тем самым образующих сеть. Узлы этой сети могут быть связаны цитированием (одна публикация цитирует другую), авторством (автор связан со своей публикацией), соавторством (авторы одной и той же публикации) и т. д.

Одной из наиболее простых и наглядных является сеть соавторства. В ней узлами являются ученые, а ненаправленная дуга между двумя узлами означает наличие хотя бы одной совместной публикации. Сеть соавторства позволяет, например, наглядно изобразить структуру публикационного сотрудничества в рамках научных подразделений. Это может оказаться полезным в ряде ситуаций (например, для нового сотрудника или руководителя подразделения). Рассмотрим в качестве примера следующий граф, на котором узлами являются сотрудники одного из реальных научных подразделений (рис. 5).

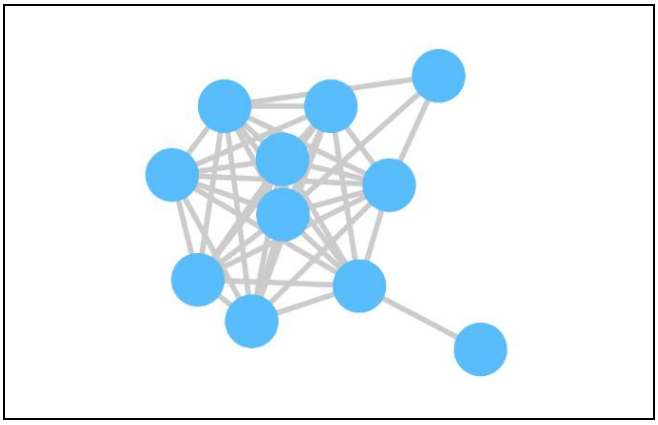


Рис. 5. Связная сеть соавторства сотрудников научного подразделения

Видно, что граф является связным, а связи в нем – достаточно плотными. Это означает, что в данной лаборатории сотрудники достаточно тесно взаимодействуют друг с другом при подготовке публикаций.

Пример другой, в некотором смысле противоположной ситуации, представлен на рис. 6.

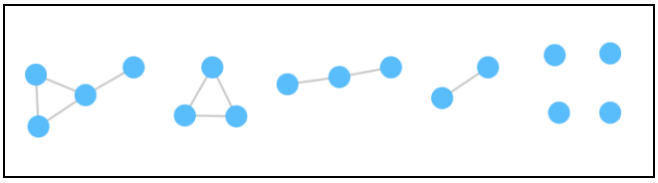


Рис. 6. Несвязная сеть соавторства сотрудников научного подразделения

Здесь имеются несколько компонент связности, в том числе включающих в себя изолированные узлы. Это может означать, что отдельные группы в данной лаборатории работают автономно.

Анализировать сети соавторства можно и с точки зрения выявления общих закономерностей. Рассмотрим вопрос о том, как наличие совместных публикаций соотносится с расстоянием между тематическими профилями авторов.

Например, среднее расстояние между профилями сотрудников ИПУ РАН является довольно большим и составляет 0,85 (в смысле метрики, описанной в § 3). Это означает, что публикации сотрудников относятся, вообще говоря, к различным областям теории управления.

Определим критерий наличия *сильной связи* между двумя авторами: существует третий автор, который имеет хотя бы одну публикацию с первым автором без второго и хотя бы одну публикацию со вторым без первого (т. е. связан отдельно с каждым из двух авторов). Оказывается, что профили авторов, соединенных сильной связью, в среднем ближе друг к другу (среднее расстояние 0,59), чем профили авторов, связь которых не является сильной (среднее расстояние 0,64), и это различие является статистически значимым. Это наблюдение отражает наличие взаимосвязи между двумя понятиями близости авторов – в смысле расстояния между профилями в тематическом пространстве теории управления и в смысле расстояния между узлами в графе соавторства.

6. ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ ИСАНД

На сегодняшний день ИСАНД предоставляет следующие возможности: построение тематического профиля ученого или подразделения, тематическое ранжирование ученого или подразделения, создание профиля связности тем, наложение профиля на граф глоссария по теории управления, а также исследование проекций профилей ученых на двумерном пространстве.

Напомним (см. п. 1.1), что система использует онтологию по теории управления, состоящую из четырех блоков: общенаучные термины, математический аппарат, предметная область и сфера применения. Общенаучные термины включают термины, которые встречаются в различных научных темах. Остальные три блока включают факторы, которые, в свою очередь, делятся на подфакторы. Каждый подфактор определяется терминами, которые были выбраны экспертами соответствующих научных областей. Таким образом, с помощью ИСАНД можно построить профиль ученого, который покажет, насколько часто данный специалист использует термины из соответствующих факторов и подфакторов.

Функциональность сайта реализована в шести разделах (рис. 7).

Рис. 7. Рабочие окна ИСАНД

6.1. Тематический поиск

Раздел «Тематический поиск» позволяет отобрать релевантные публикации, ученых, журналы, конференции, организации и города по заранее заданным факторам (первый уровень тематической классификации), подфакторам (второй уровень тематической классификации) и терминам теории управления. Он предполагает первоначальный выбор одного метафактора из четырех:

- общенаучная проблематика,
- предметная область,
- математический аппарат,
- сфера применения (рис. 8).

Следующим этапом предполагается выбор темы из предложенного списка с возможностью поиска темы, а также сортировки тем по алфавиту и по популярности (рис. 9).

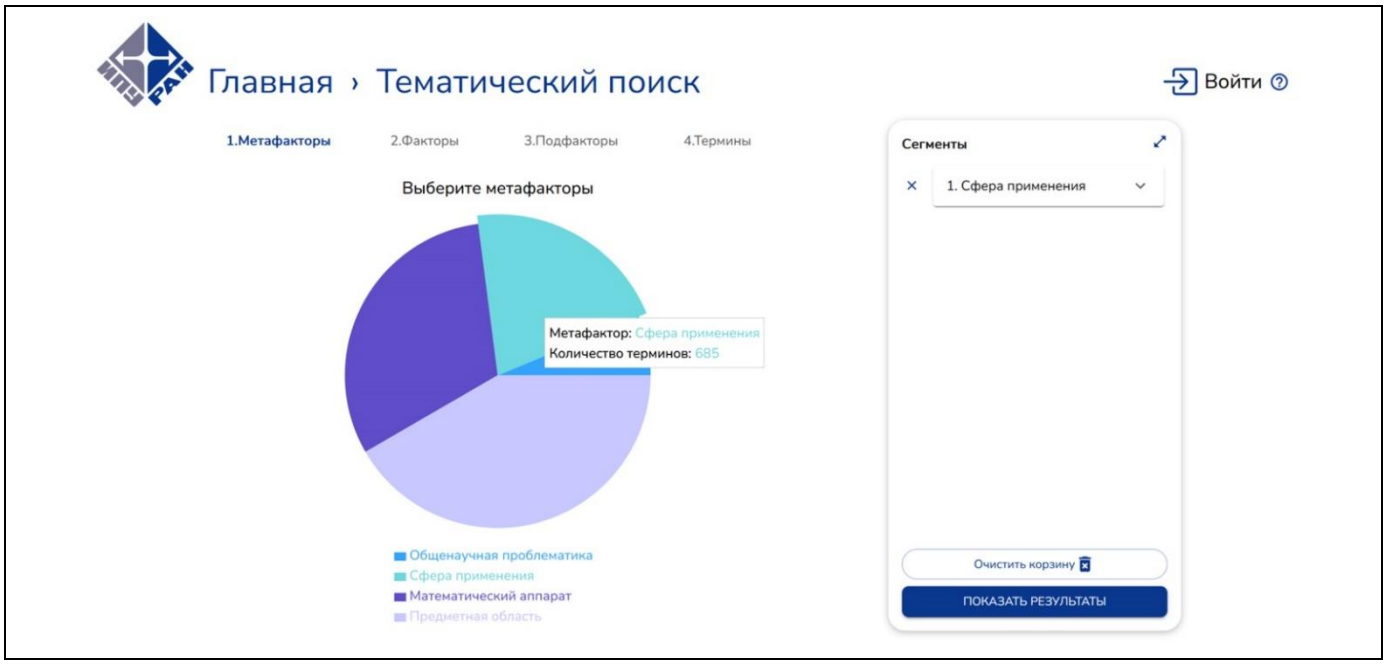


Рис. 8. Выбор метафакторов для тематического поиска

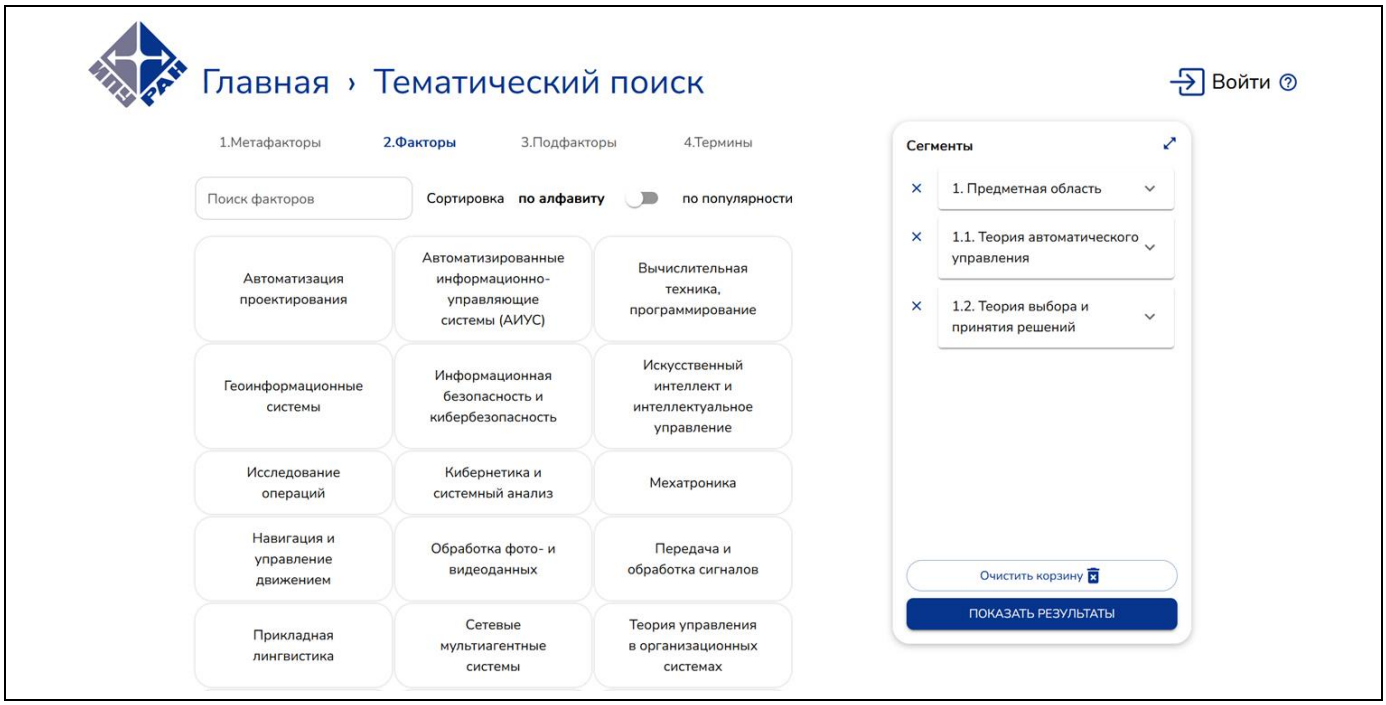


Рис. 9. Выбор факторов для тематического поиска

Для уточнения выборки могут быть заданы также подфакторы и термины на аналогичных интерфейсах.

Выдача результатов группируется по публикациям, авторам, городам, журналам, организациям и конференциям (рис. 10). В выдаче результатов показывается количество терминов, найденных в материалах с учетом примененных фильтров и групп. Таким образом, ученый может получить ответы на вопросы: какие публикации являются наиболее релевантными выбранной тематике, какие ученые преимущественно занимаются выбранной темой, на каких конференциях чаще всего можно услышать доклады по данному направлению, в каких журналах стоит публиковать статью по заявленным темам и, наконец, в каких организациях и городах работают ученые, связанные с заданными направлениями.

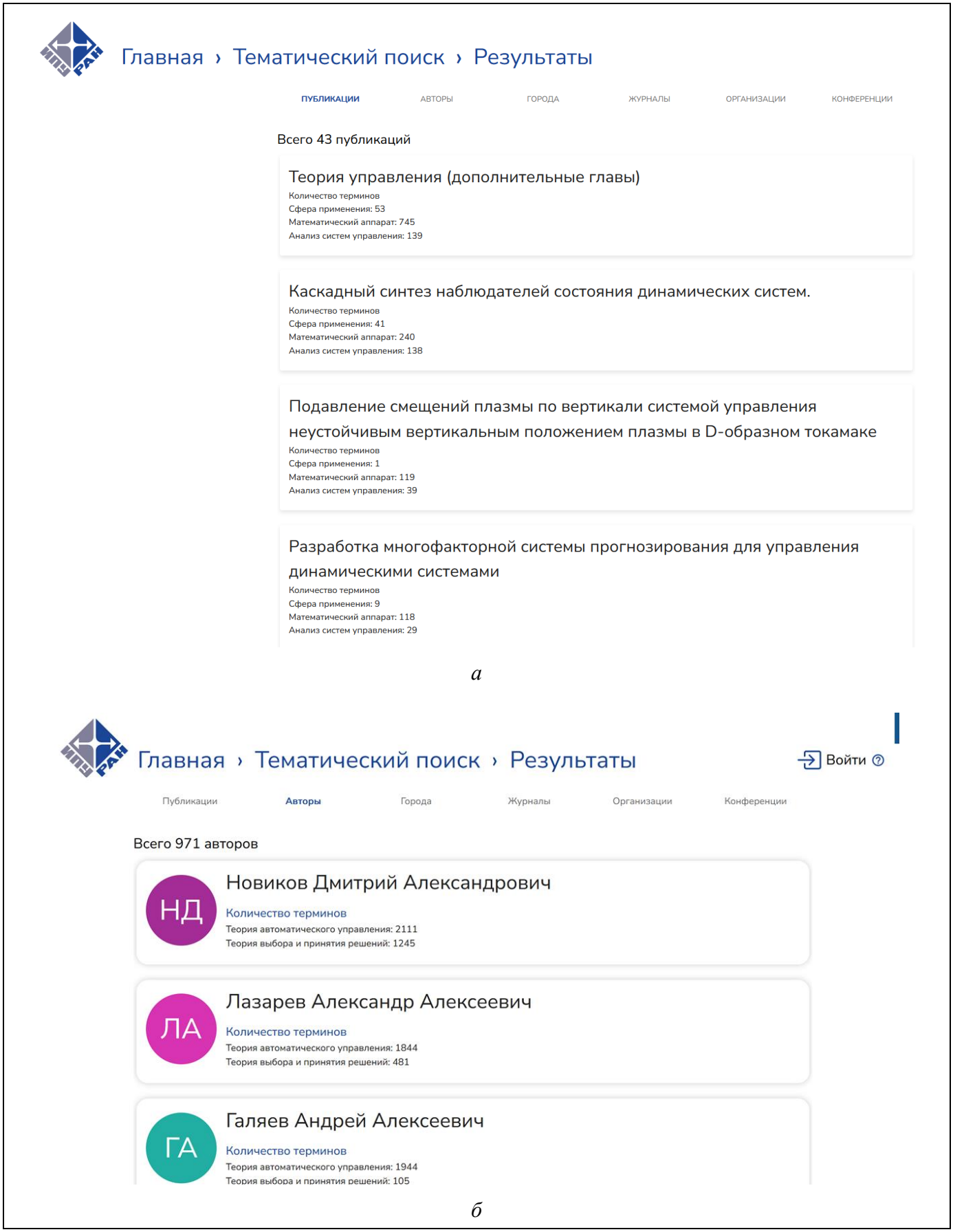


Рис. 10. Отсортированный список результатов тематического поиска: *а* – по публикациям, *б* – по авторам

В профиле публикации выводятся название, аннотация и авторы (рис. 11).

Построение тематического профиля дает пользователю возможность увидеть профиль с четырех ракурсов: предметная область, математический аппарат, сфера применения, общенаучная проблематика в диаграммах по факторам, подфакторам и терминам (рис. 12, *а* – *в* соответственно).

Тематический профиль является эффективным инструментом для решения многих задач анализа научной деятельности – например, для тематического анализа научных групп.

6.2. Профили ученых

В ходе научной деятельности регулярно возникают ситуации, когда несколько научных коллективов занимаются исследованием одной задачи. Это может происходить как в рамках общего проекта, так и в процессе реструктуризации научных подразделений. Для планирования научной деятельности необходимо понимать, какими компетенциями обладают сотрудники.

Предлагается определение направлений исследований научных коллективов на основании тем их публикаций. Более детальный анализ сравнения научных групп проводится с помощью некоторых критериев, приведенных в этой работе.

Данный раздел предоставляет возможность построить и сравнить тематические профили выбранных ученых. Для построения сравнительных диаграмм первым шагом необходимо выбрать авторов для сравнения. При выборе авторов есть возможность выбрать отдельные работы либо все работы автора.

Слева на вертикальной оси диаграммы выводятся термины, столбец диаграммы показывает количество вхождений в публикации. Значение столбца диаграммы может быть приведено к одной из пяти схем отображения.

• «Абсолютный вектор» отражает суммарное количество вхождений терминов.

• «Стохастический вектор» – абсолютный вектор с нормализованными столбцами.

• «Булев вектор» – компоненты этого вектора могут принимать два значения: единица – если количество терминов больше значения «отсечение по терминам» и ноль – в остальных случаях.

• «По количеству использованных терминов» – вариант абсолютного вектора, когда «отсечение по терминам» убирает те столбцы, где «уникальных» терминов меньше, чем значение «отсечение по терминам». В случае абсолютного вектора «отсечение по терминам» работает с общим числом вхождений терминов, в данном случае – с уникальным.

• «Термины» – просмотр терминов.

Выбор уровня означает выбор уровня графа глоссария – дерева терминов, в соответствии с которым проводится сравнение публикаций. Большее значение уровня обеспечивает более детальный анализ. Отсечение по категориям и по терминам убирает минимальные приведенные значения, чтобы сделать график более выразительным. Чекбокс «Учитывать общенаучные термины» добавляет в выборку слова, относящиеся к общенаучной проблематике.

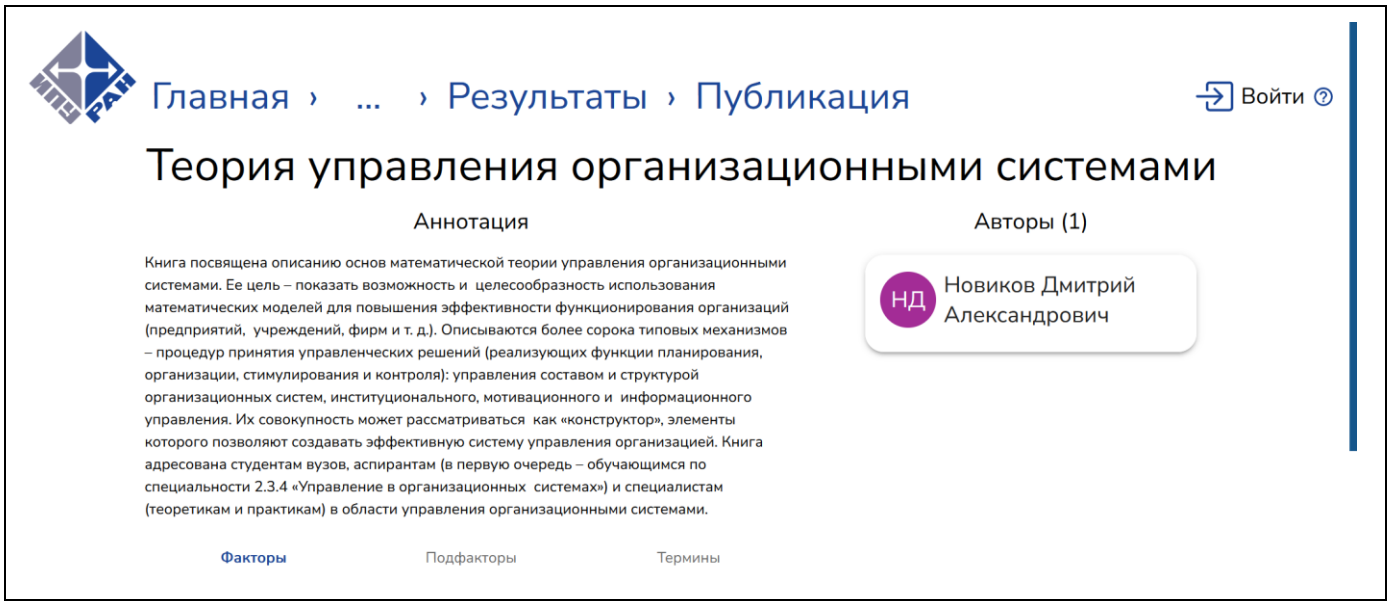


Рис. 11. Карточка публикации

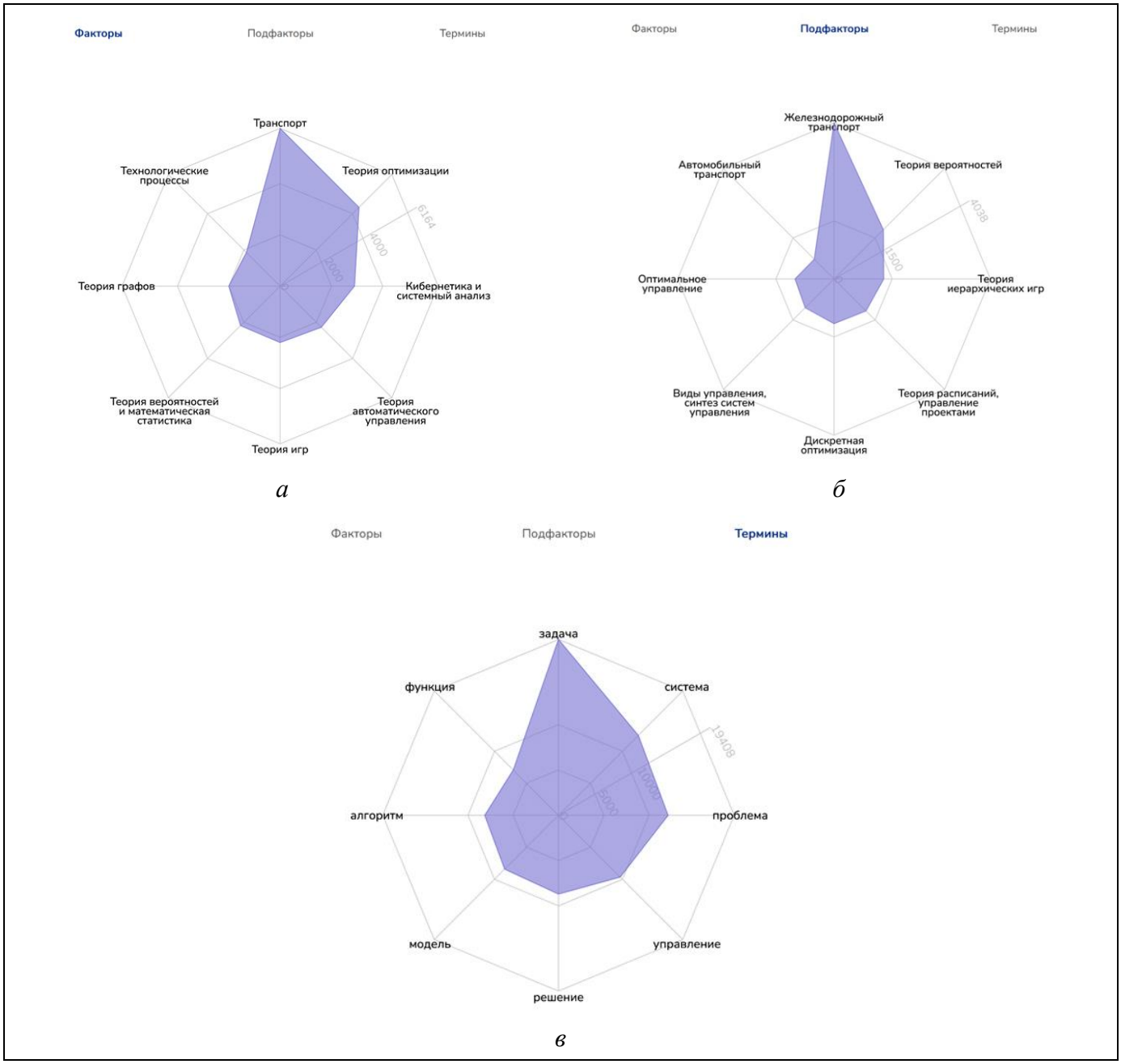


Рис. 12. Тематический профиль научного объекта: *а* – факторы, *б* – подфакторы, *в* – термины

Опция «Выберите путь» позволяет уточнить тему – второй уровень графа глоссария. Временна́я шкала позволяет выбрать промежуток времени для сравнения публикаций. Имеется возможность углубленного сравнения профилей по подфакторам второго уровня (рис. 13).

6.3. Тематическое ранжирование

В данном разделе пользователь получает список релевантных ученых, отсортированный по количеству использованных терминов выбранных факторов (первый уровень тематической классификации) или подфакторов (второй уровень тематической классификации). Уровень задает глубину анализа в соответствии с используемым графом глоссария. На нулевом уровне тематической классификации выбираются метафакторы, на первом – факторы, на втором – подфакторы (рис. 14).

6.4. Граф классификатора

Данный раздел позволяет узнать частоту использования терминов в публикациях выбранного автора. Запрашиваемая информация отображается в виде графа: его вершина соответствует термину,

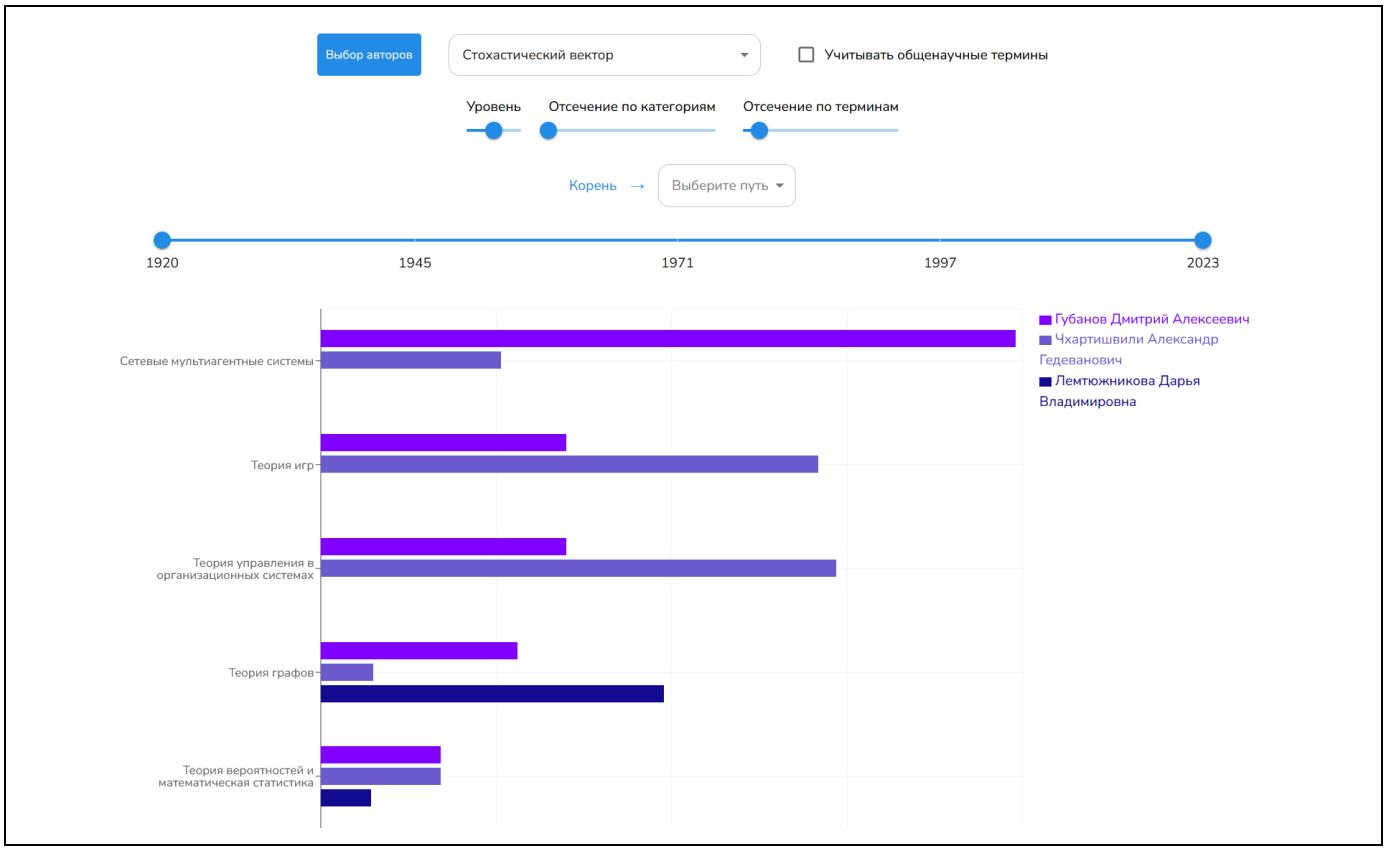


Рис. 13. Сравнение профилей ученых по подфакторам

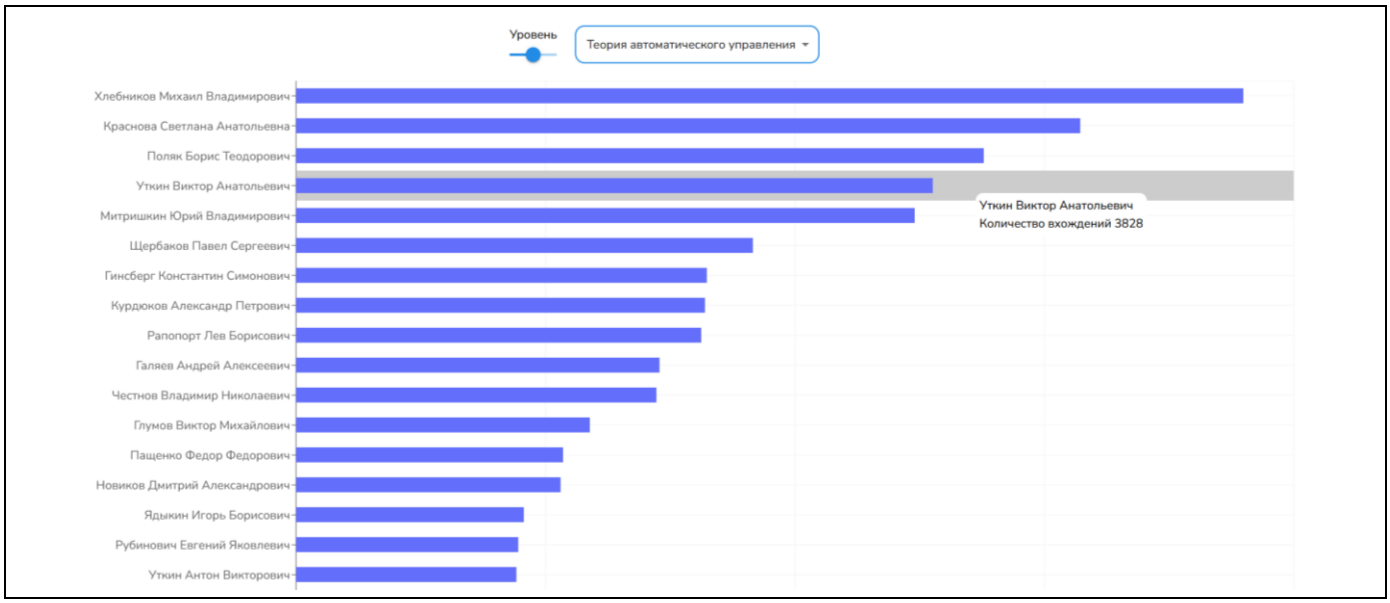


Рис. 14. Тематическое ранжирование

а ребро – совместному употреблению терминов в одной публикации.

Построение графов связности терминов производится по автору, выбор которого расположен в верхнем поле интерфейса. Применение бегунка «Отсечение по частоте» позволяет выводить только высокочастотные термины при увеличении значения и отсекать низкочастотные. Уровень терминов означает используемый уровень графа глоссария, по которому производится анализ.

На графе классификатора выводятся термины, откалиброванные по частоте использования, к ним

может быть применено два вида значимой расцветки: по количеству вхождений и по количеству связей. Шкала значений расцветки размещается справа от области графа.

Отображение ребер и названий влияет на внешний вид графа и добавляет названия и ребра к вершинам терминов.

Включение общенаучных терминов добавляет точки общенаучных терминов на граф (рис. 15).

6.5. Граф глоссария

Данный раздел является наглядным инструментом отображения используемых терминов теории управления для выбранного ученого. Он предоставляет возможность исследовать окрестность терминов разных порядков. Достигается это путем построения ориентированного графа взаимосвязи терминов на основании глоссария. При этом вершина графа a соответствует термину a, ориентированное ребро (a, b) – использованию термина a в определении термина b.

Граф глоссария – это граф, содержащий термины и связи типа «Определяется через». Если в определении термина b встречается термин a, то термин a соединен с термином b направленным ребром.

Первым шагом выбирается автор в верхнем поле, и на графе подсвечиваются все термины, которые автор использовал в работах, а также связи между ними.

Поле «Наследовать от» позволяет выбрать термин, через который будут определены другие термины. Глубина проработки наследования задается в соседнем поле с числовой шкалой.

Режим расцветки реализован в соответствии с определяющей силой термина – отношением количества терминов, которые определены через данный термин, к количеству терминов, через которые данный термин определен.

Если выбрать функцию «Подсветить термин», то будет подсвечен выбранный термин с исходящими стрелками к терминам, которые определены

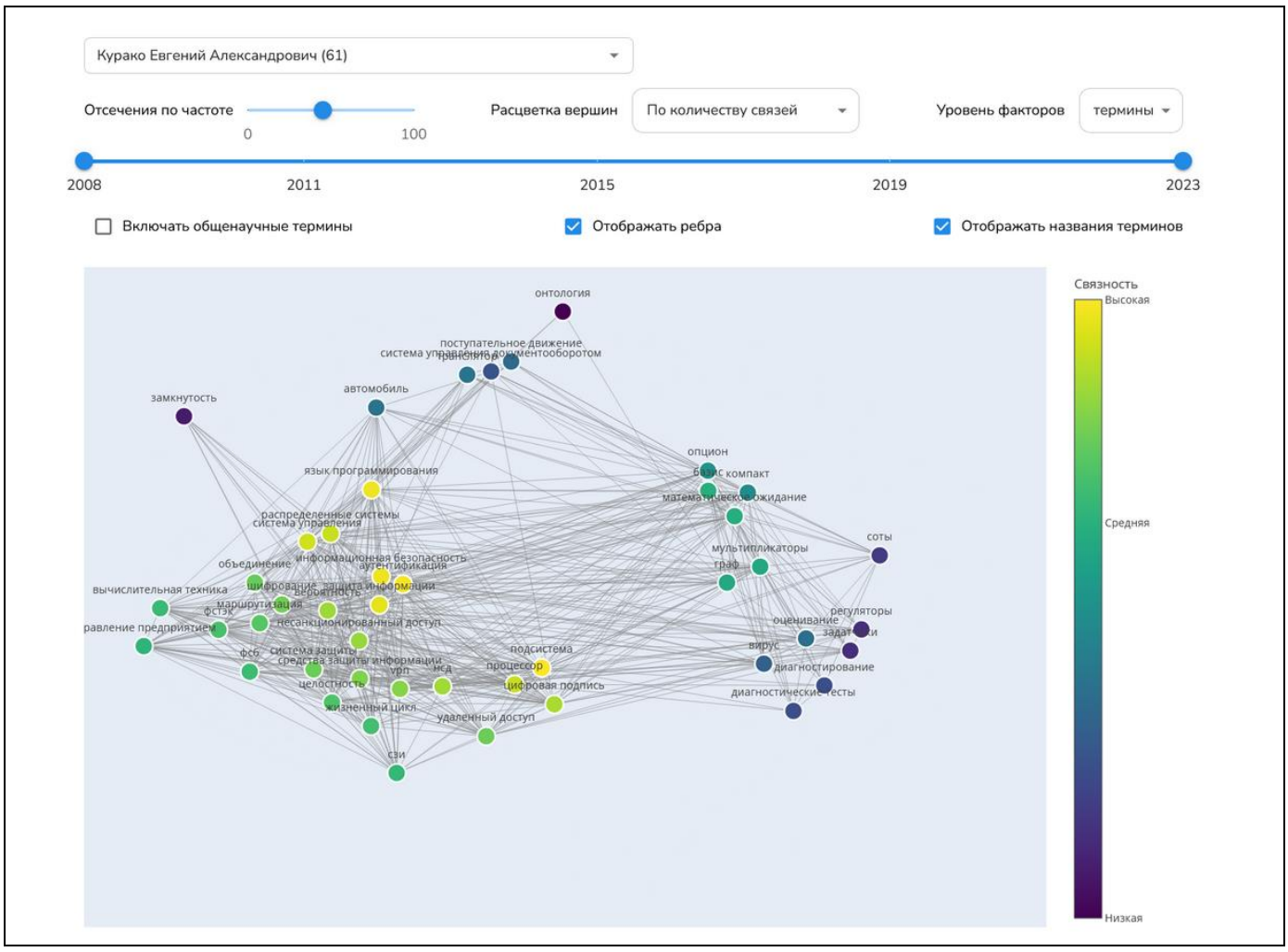


Рис. 15. Построение графа классификатора

через него, и входящими стрелками от тех терминов, через которые определен выбранный термин.

Режим «Всегда отображать названия терминов» позволяет видеть названия терминов на графе (рис. 16).

6.6. Глоссарий

Глоссарий представляет собой сборник терминов теории управления и их описаний (рис. 17, см. также https://www.ipu.ru/education/glossary).

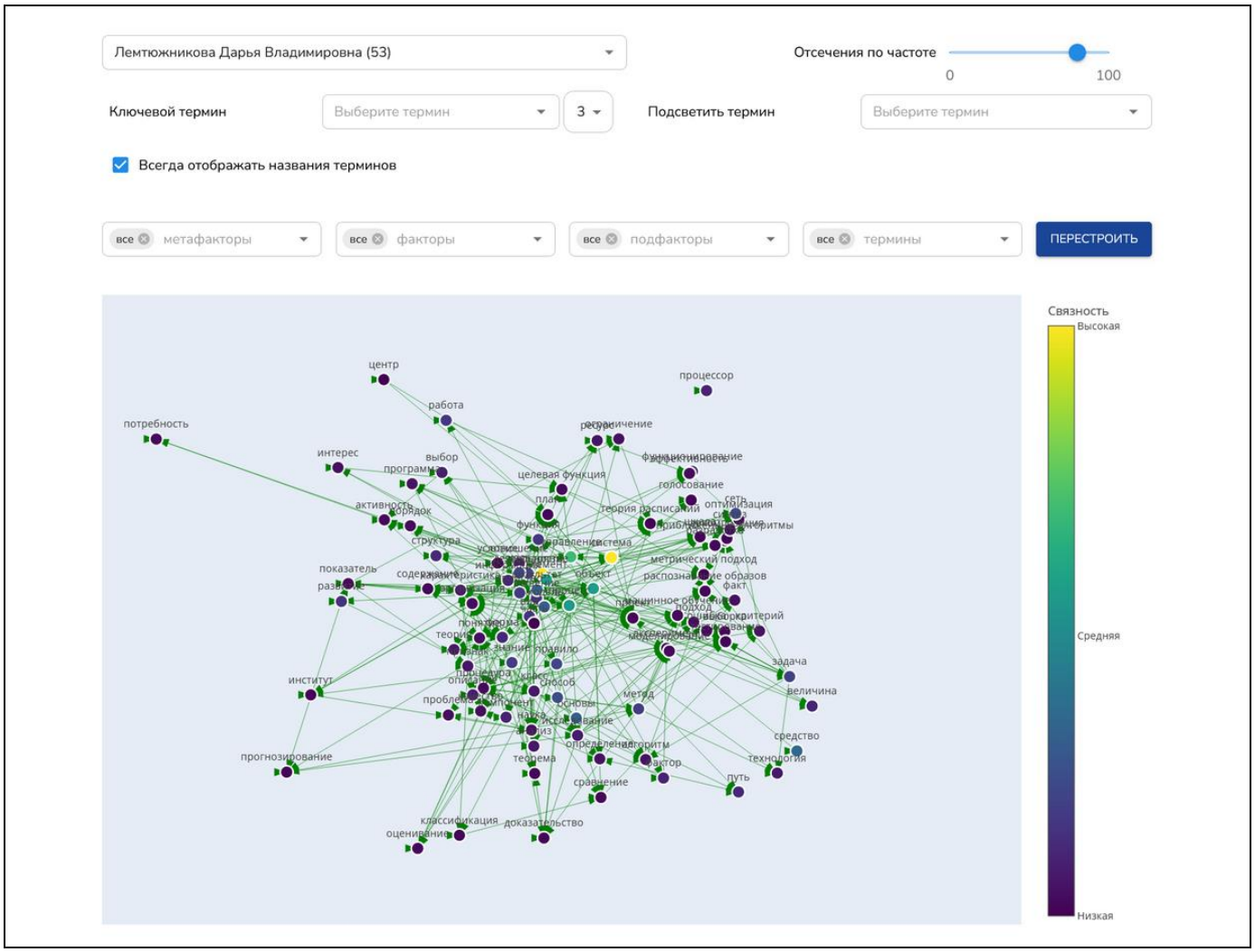
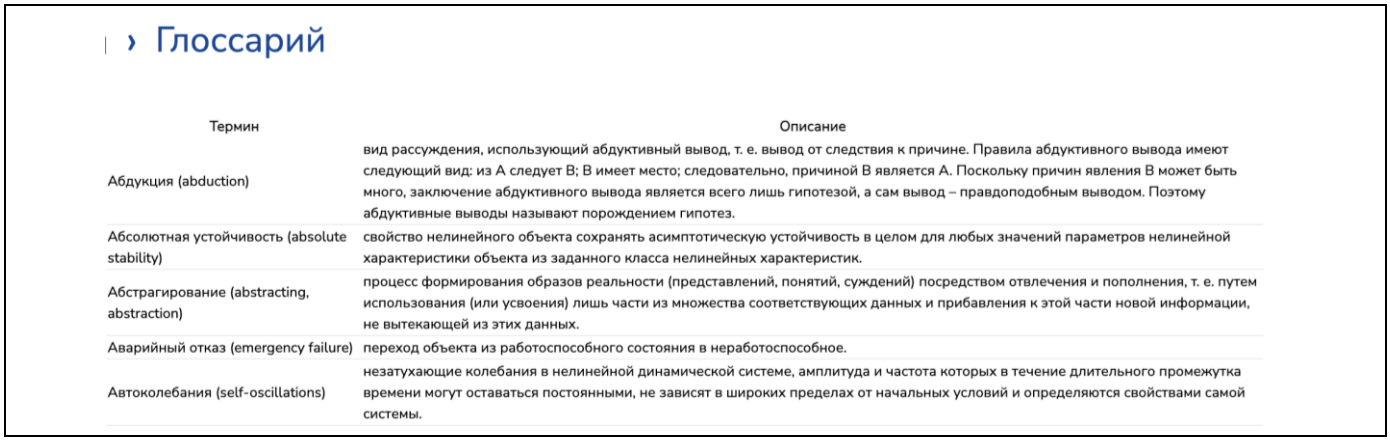


Рис. 16. Термины автора в графе глоссария



› Глоссарий

Термин	Описание
Абдукция (abduction)	вид рассуждения, использующий абдуктивный вывод, т. е. вывод от следствия к причине. Правила абдуктивного вывода имеют следующий вид: из A следует B; B имеет место; следовательно, причиной B является A. Поскольку причин явления B может быть много, заключение абдуктивного вывода является всего лишь гипотезой, а сам вывод – правдоподобным выводом. Поэтому абдуктивные выводы называют порождением гипотез.
Абсолютная устойчивость (absolute stability)	свойство нелинейного объекта сохранять асимптотическую устойчивость в целом для любых значений параметров нелинейной характеристики объекта из заданного класса нелинейных характеристик.
Абстрагирование (abstracting, abstraction)	процесс формирования образов реальности (представлений, понятий, суждений) посредством отвлечения и пополнения, т. е. путем использования (или усвоения) лишь части из множества соответствующих данных и прибавления к этой части новой информации, не вытекающей из этих данных.
Аварийный отказ (emergency failure)	переход объекта из работоспособного состояния в неработоспособное.
Автоколебания (self-oscillations)	незатухающие колебания в нелинейной динамической системе, амплитуда и частота которых в течение длительного промежутка времени могут оставаться постоянными, не зависят в широких пределах от начальных условий и определяются свойствами самой системы.

Рис. 17. Глоссарий

ЗАКЛЮЧЕНИЕ

Средства системы ИСАНД позволяют впервые автоматизировать решение научно-организационных задач в области теории управления, связанных с подбором экспертов и рецензентов с определенными компетенциями, поиском публикаций по определенной тематике, анализом тематики научного коллектива и ее эволюции и т. д. Кроме того, в перспективе ИСАНД предоставляет широкие возможности для наукометрических исследований, проводимых с целью установления близостью публикаций, авторов и коллективов, построения сети соавторства и цитирования, тематических трендов. По мере расширения базы публикаций и пополнения словаря терминов (нижнего уровня онтологии научного знания) эти возможности будут совершенствоваться. Разумеется, эти возможности относятся только к теории управления, однако в структурном плане система ИСАНД могла бы стать прототипом для аналогичных систем в других областях науки.

Авторы статьи выражают благодарность коллегам, которые приняли активное участие в проекте: Авдеевой З.К., Агаеву Р.П., Алчинову А.И., Антипину С.И., Арутюнову А.В., Барабанову И.Н., Батову А.В., Бахтадзе Н.Н., Богачевой Д.Н., Буркову В.Н., Васильеву С.Н., Вишневскому В.М., Владимировой С.С., Вытовтову К.А., Гаврилову М.С., Галяеву А.А., Голеву А.В., Гребенкову Д.И., Дранко О.И., Дранову Е.М., Жиляковой Л.Ю., Калашникову А.О., Калугину К.А., Каляному Г.Н., Караваю М.Ф., Карпухиной Д.Р., Каршакову Е.В., Кирьянову П.А., Козловой А.А., Красновой С.А., Красоткину С.А., Кудинову И.Д., Кульбе В.В., Лазареву А.А., Латипову А.Р., Лебедеву В.Г., Лычагину В.Г., Макаренко А.В., Мельничуку В.С., Мешкову Д.О., Мещерякову Р.В., Михальскому А.И., Назину А.В., Нижегородцеву Р.М., Осколкову Н.С., Рапопорту Л.Б., Рощину А.А., Рубиновичу Е.Я., Сальникову А.М., Сергееву В.А., Сокольскому К.А., Стрыгину Д.Д., Суховерову В.С., Сычу В.В., Толоку А.В., Уткину В.А., Фархадову М.П., Федянину Д.Н., Хлебникову М.В., Хрипунову С.П., Шарафиеву А.Ф., Шекунову М.А., Щепкину А.В., Ядыкину И.Б.

ЛИТЕРАТУРА

1. *ГОСТ* 7.90–2007. Система стандартов по информации, библиотечному и издательскому делу. Универсальная десятичная классификация. Структура, правила введения и индексирования. – М: Стандартинформ, 2010. – 26 с. [GOST 7.90 2007. Sistema standartov po informacii, bib-liotechnomu i izdatel'skomu delu. Universal'naja desjatich-naja klassifikacija. Struktura, pravila vvedenija i indek-sirovanija. – Moscow: Standartinform, 2010. – 26 p.(In Russian)]
2. Revised Field of Science and Technology (FOS) Classification in the Frascati Manual. – Paris: Organisation for Economic Co-operation and Development, 2007. – 12 p.
3. *Классификатор* Российского научного фонда (РНФ). – URL: https://rscf.ru/contests/classification (дата обращения: 24.04.2024). [Klassifikator Rossijskogo nauchnogo fonda (RNF). – URL: https://rscf.ru/contests/classification (accessed April 24, 2024). (In Russian)]
4. *Государственный* рубрикатор научно-технической информации. – URL: https://grnti.ru (дата обращения: 24.04.2024). – [Gosudarstvennyj rubrikator nauchno-tehnicheskoj informacii. – URL: https://grnti.ru (accessed April 24, 2024). (In Russian)]
5. *Кузнецов О.П., Суховеров В.С.* Онтологический подход к оценке тематики научного текста // Онтология проектирования. – 2016. – Т .6, № 1. – С. 55–66. [*Kuznetsov, O.P., Sukhoverov, V.S.* Ontologicheskii podkhod k otsenke tematiki nauchnogo teksta // Ontologiya proektirovaniya. – 2016. – Vol. 6, no. 1. – P. 55–66. (In Russian)]
6. *Gruber, T.R.* A translation Approach to Portable Ontology Specifications // Knowledge Acquisition. – 1993. – Vol. 5, no. 2. – P. 199–220.
7. *Borst, W.N.* Construction of Engineering Ontologies for Knowledge Sharing and Reuse: PhD Thesis. – Enschede: Centre for Telematics and Information Technology (CTIT), 1997.
8. *OWL* 2 Web Ontology Language: Primer (Second Edition). Ed. by P. Hitzler, M. Krötzsch, B. Parsia, P.F. Patel-Schneider, S. Rudolph. – W3C, 2012. – URL: http://www.w3.org/TR/2012/REC-owl2-primer-20121211/.
9. *Sengupta, K., Hitzler, P.* Web Ontology Language (OWL) // Encyclopedia of Social Network Analysis and Mining. – 2014. – P. 2374–2378.
10. *Губанов Д.А., Кузнецов О.П., Суховеров В.С., Чхартишвили А.Г.* О построении профилей в тематическом пространстве теории управления // Материалы 9-й Международной конференции «Знания-Онтологии-Теории» (ЗОНТ-2023). – Новосибирск, 2023. – С. 89–94. [*Gubanov, D.A., Kuznetsov, O.P., Sukhoverov, V.S., Chkhartishvili, A.G.* O postroenii profilei v tematicheskom pro-stranstve teorii upravleniya // Materialy 9-i Mezhdunarodnoi konferentsii «Znaniya-Ontologii-Teorii» (ZONT-2023). – Novosibirsk, 2023. – P. 89–94. (In Russian)]
11. *Теория* управления: словарь системы основных понятий. – М.: ЛЕНАНД, 2024. – 128 с. [*Teoriya* upravleniya: slovar' sistemy osnovnykh ponyatii. – M.: LENAND, 2024. – 128 p. (In Russian)]
12. *Губанов Д.А., Новиков Д.А.* Анализ терминологической структуры теории управления // Управление большими системами. – 2024. (в печати) [*Gubanov, D.A., Novikov, D.A.* Analiz terminologicheskoi struktury teorii upravleniya // Large-Scale Systems Control. – 2024. (In Russian, in print)]
13. *Теория* управления. Терминология. Вып. 107. – М.: Наука, 1988. – 56 с. [*Teoriya* upravleniya. Terminologiya. – Vol. 107. – M.: Nauka, 1988. – 56 p. (In Russian)]
14. *Karba, R., Kocijan, J., Bajd, T.,* et al. Terminological Dictionary of Automatic Control, Systems and Robotics. – Heidelberg: Springer, 2024. – 249 p.
15. *Glossary* of Control Engineering Terms. – URL: www.actc-control.com/glossary.
16. *Новиков Д.А., Новиков А.М.* Методология научного исследования. – М.: Либроком, 2010. – 280 с. [*Novikov, D.A., Novikov, A.M.* Metodologiya nauchnogo issledovaniya. – M.: Librokom, 2010. – 280 p. (In Russian)]
17. *Gomes Junior, A. de A., Schramm, V.B.* Problem Structuring Methods: A Review of Advances over the Last Decade // Syst. Pract. Action. Res. – 2022. – Vol. 35. – P. 55–88.

18. *Tkaczyk, D., Szostek, P., Fedoryszak, M.,* et al. CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature // International Journal on Document Analysis and Recognition (IJDAR). – 2015. – Vol. 18, no. 4. – P. 317–335.
19. *Krause, J., Shapiro, I., Saier, T., Farbe, M.* Bootstrapping Multilingual Metadata Extraction: A Showcase in Cyrillic // Proceedings of the Second Workshop on Scholarly Document Processing. – Mexico, 2021. – P. 66–72.
20. *Lopez, P.* GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications // Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL. – Corfu, 2009. – P. 473–474.
21. *Wen, Y., Fan, C., Chen, G.,* et al. A Survey on Named Entity Recognition // Communications, Signal Processing, and Systems. – Singapore: Springer, 2020. – P. 1803–1810.
22. *Mielke, S.J., Alyafeai, Z., Salesky, E., et al.* Between Words and Characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP // arXiv:2112.10508. – 2021. – DOI: https://doi.org/10.48550/arXiv.2112.10508.
23. *Acs, J., Kadar, A., Kornai, A.* Subword pooling makes a difference // arXiv:2102.10864. – 2021. – DOI: https://doi.org/10.48550/arXiv.2102.10864.
24. *Liu, R., Mao, R., Luu, A.T., Cambria, E.* A Brief Survey on Recent Advances in Coreference Resolution // Artificial Intelligence Review. – 2023. – Vol. 56. – P. 14439–14481.
25. *Joshi, M., Levy, O., Weld, D.S., Zettlemoyer, L.* BERT for coreference resolution: Baselines and analysis // arXiv:1908.09091. – 2019. – DOI: https://doi.org/10.48550/arXiv.1908.09091.

Статья представлена к публикации членом редколлегии В.Г. Лебедевым.

Поступила в редакцию 14.05.2024,
после доработки 10.06.2024.
Принята к публикации 10.06.2024.

Губанов Дмитрий Алексеевич – д-р техн. наук, Институт проблем управления им. В. А. Трапезникова РАН, г. Москва, ✉ dmitry.a.g@gmail.com,
ORCID iD: https://orcid.org/0000-0002-0099-3386

Кузнецов Олег Петрович – д-р техн. наук, Институт проблем управления им. В. А. Трапезникова РАН, г. Москва, ✉ olpkuz@yandex.ru,
ORCID iD: https://orcid.org/0000-0002-5061-3855

Курако Евгений Александрович – канд. техн. наук, Институт проблем управления им. В. А. Трапезникова РАН, г. Москва, ✉ kea@ipu.ru,
ORCID iD: https://orcid.org/0009-0008-4746-1943

Лемтюжникова Дарья Владимировна – канд. физ.-мат. наук, Институт проблем управления им. В. А. Трапезникова РАН; МАИ (национальный исследовательский университет), г. Москва, ✉ darabbt@gmail.com,
ORCID iD: https://orcid.org/0000-0002-5311-5552

Новиков Дмитрий Александрович – д-р техн. наук, академик РАН, Институт проблем управления им. В. А. Трапезникова РАН, г. Москва, ✉ novikov@ipu.ru,
ORCID iD: https://orcid.org/0000-0002-9314-3304

Чхартишвили Александр Гедеванович – д-р физ.-мат. наук, Институт проблем управления им. В. А. Трапезникова РАН, г. Москва,✉ sandro_ch@mail.ru,
ORCID iD: https://orcid.org/0000-0002-2970-1244

ISAND: AN INFORMATION SYSTEM FOR SCIENTIFIC ACTIVITY ANALYSIS (IN THE FIELD OF CONTROL THEORY AND ITS APPLICATIONS)

D. A. Gubanov*, O. P. Kuznetsov**, E. A. Kurako***, D. V. Lemtyuzhnikova****, D. A. Novikov*****, and A. G. Chkhartishvili******

*–******Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
****Moscow Aviation Institute (National Research University), Moscow, Russia

*✉ dmitry.a.g@gmail.com, **✉ olpkuz@yandex.ru, ***✉ kea@ipu.ru, ****✉ darabbt@gmail.com, *****✉ novikov@ipu.ru, ******✉ sandro_ch@mail.ru

Abstract. This paper describes the approaches underlying ISAND, an information system for scientific activity analysis in the field of control theory and its applications. ISAND is being developed at the Trapeznikov Institute of Control Sciences, the Russian Academy of Sciences. The ISAND ontology is oriented toward the representation and collection of knowledge in the field of control theory and its applications, namely, scientific knowledge (the ontology of control theory) and knowledge related to the scientific activity of agents (organizations, journals, conferences, and individual researchers) in this field. Based on this ontology, the ISAND architecture is a complex program system to collect, store, and analyze publications and their metadata from external sources. The ISAND algorithm for building the thematic profiles of scientific objects (publications, researchers, organizations, journals, and conferences), as well as ISAND text processing and network analysis capabilities, are presented. Finally, the main possibilities of using ISAND are considered.

Keywords: scientific activity analysis, control theory and its applications, information system, classification, ontology, thematic profile, thematic space, term, text processing, network analysis.

Acknowledgments. The authors are grateful to the following colleagues for their active participation in the project: Z.K. Avdeeva, R.P. Agaev, A.I. Alchinov, S.I. Antipin, A.V. Arutyunov, I.N. Barabanov, A.V. Batov, N.N. Bakhtadze, D.N, Bogacheva, V.N. Burkov, S.N. Vassilyev, V.M. Vishnevsky, S.S. Vladimirova, K.A. Vytovtov, M.S. Gavrilov, A.A. Galyaev, A.V. Golev, D.I. Grebenkov, O.I. Dranko, E.M. Dranov, L.Yu. Zhilyakova, A.O. Kalashnikov, K.A. Kalugin, G.N. Kalyanov, M.F. Karavay, D.R. Karpukhina, E.V. Karshakov, P.A. Kiryanov, A.A. Kozlova, S.A. Krasnova, S.A. Krasotkin, I.D. Kudinov, V.V. Kul'ba, A.A. Lazarev, A.R. Latipov, V.G. Lebedev, V.G. Lychagin, A.V. Makarenko, V.S. Melnichuk, D.O. Meshkov, R.V. Meshcheryakov, A.I. Mikhalsky, A.V. Nazin, R.M. Nizhegorodtsev, N.S. Oskolkov, L.B. Rapoport, A.A. Roshchin, E.Ya. Rubinovich, A.M. Salnikov, V.A. Sergeev, K.A. Sokolsky, D.D. Strygin, V.S. Sukhoverov, V.V. Sych, A.V. Tolok, V.A. Utkin, M.P. Farkhadov, D.N. Fedyanin, M.V. Khlebnikov, S.P. Khripunov, A.F. Sharafiev, M.A. Shekunov, A.V. Shchepkin, and I.B. Yadykin.