# Информационная система анализа научной деятельности (ИСАНД)

Чхартишвили Александр Гедеванович (лаб. 57)

27 ноября 2023 г.

# ОСНОВНЫЕ УЧАСТНИКИ

- Губанов Д.А. (лаб. 11)
- Кузнецов О.П. (лаб. 11)
- Курако Е.А. (лаб. 79)
- Лемтюжникова Д.В. (лаб. 90)
- Чхартишвили А.Г. (лаб. 57)

и др.

# ЭКСПЕРТЫ

| | | | |
|---|---|---|---|
| Агаев Р.П. | Алчинов А.И. | Антипин С.И . | Арутюнов А.В. |
| Барабанов И.Н. | Бахтадзе Н.Н. | Бурков В.Н. | Васильев С.Н. |
| Вишневский В.М. | Вытовтов К.А. | Галяев А.А. | Дранко О.И. |
| Калашников А.О. | Калянов Г.Н. | Каравай М.Ф. | Каршаков Е.В. |
| Краснова С.А. | Кузнецов О.П. | Кульба В.В. | Лазарев А.А. |
| Лебедев В.Г. | Лычагин В.Г. | Макаренко А.В. | Мешков Д.О. |
| Мещеряков Р.В. | Михальский А.И. | Назин А.В. | Нижегородцев Р.М. |
| Новиков Д.А. | Рапопорт Л.Б. | Рощин А.А. | Рубинович Е.Я. |
| Толок А.В. | Уткин В.А. | Фархадов М.П. | Хлебников М.В. |
| Хрипунов С.П. | Чхартишвили А.Г. | Щепкин А.В. | Ядыкин И.Б. |

# ПРОБЛЕМАТИКА

Автоматизированный анализ научных публикаций, поддержки принятия решений в научных организациях и сопровождения деятельности ученых, например:

- Кому направить на рецензирование статью или заявку на грант?
- В какую рубрику журнала поместить отрецензированную статью?
- Какие есть статьи по заданной тематике?

Потенциальные пользователи: исследователи, студенты, аспиранты, редакции журналов, руководители организаций и подразделений, …

Необходимо прежде всего определение тематики научного текста и/или сферы компетентности ученого (позиционирования объектов научной деятельности в тематическом пространстве науки)

# ПРОБЛЕМАТИКА

- Тематические классификаторы научных текстов (УДК, OECD, классификатор РНФ, ГРНТИ и пр.) не вполне удовлетворяют требованиям к структуре и содержанию тематического пространства в т.ч. в силу своей универсальности
- ИСАНД – специализированный классификатор, разрабатывается на основе онтологии научного знания теории управления, составленной при помощи экспертов

# ВНУТРЕННЯЯ МОТИВАЦИЯ

## Онлайновая социальная сеть

- пользователи создают (публикуют) контент и связаны различными отношениями (дружба, комментирование и т.п.)

## Сеть научных публикаций

- авторы создают (публикуют) работы и связаны различными отношениями (соавторство, цитирование и т.п.)

# ВНУТРЕННЯЯ МОТИВАЦИЯ

Подходы к интеллектуальному анализу научных исследований

- Лингвистический (анализ текстов)
- Сетевой (анализ связей)

# ОНТОЛОГИЯ НАУЧНОГО ЗНАНИЯ ТЕОРИИ УПРАВЛЕНИЯ

- **Нулевой уровень –** математический аппарат / предметная область / сфера применения
- **Первый уровень - темы**

  52 фактора. Пример: Исследование операций
- **Второй уровень - подтемы**

  161 фактор. Пример: Управление запасами
- **Третий уровень - базовые термины**

  3032 термина. Пример: Оптимизация объемов сырья

# ПРОФИЛИ (ПЕРВОГО И ВТОРОГО УРОВНЯ)

*Профиль* – вектор, координаты научного объекта (например, научной статьи или ученого) в тематическом пространстве науки

- **Профиль публикации** – стохастический вектор, размерность которого соответствует количеству факторов или подфакторов
- **Профиль ученого** – нормированная сумма профилей его публикаций с учетом количества соавторов (каждая публикация входит с весом, обратно пропорциональным количеству авторов). Наряду с профилем важными характеристиками ученого являются общее количество его публикаций и общее количество «авторств», т.е. публикаций с учетом количества соавторов
- **Профиль организации** – нормированная сумма профилей его публикаций с учетом количества соавторов и их аффилиаций. Наряду с тематическими профилями важными характеристиками организации являются общее количество связанных с ней публикаций и скорректированное количество публикаций (с учетом количества соавторов и аффилиаций)
- **Профиль журнала (конференции)** рассчитывается аналогично

# ПРОФИЛЬ ПУБЛИКАЦИИ

- $V = \{v_1, \ldots, v_n\}$ – множество вершин 1-го уровня (*темы*)
- $V_i = \{v_{i1}, \ldots, v_{in_i}\}$ – множество вершин 2-го уровня (*подтем*) для *i*-й вершины 1-го уровня, $m = \sum_{i \in N} n_i$ - общее число подтем
- $Q_{ij} = \{1, \ldots, q_{ij}\}$ – множество вершин-терминов, характеризующих *ij*-ю подтему
- $L$ – множество публикаций
- $\Delta_{lij}$ – сумма числа вхождений в *l*-ю публикацию базовых терминов из $ij$-й подтемы
- Профиль второго уровня публикации *l*:

$$x_l = (x_{l1}, \ldots, x_{lij}, \ldots, x_{lnm}), \qquad \text{где } x_{lij} = \frac{\Delta_{lij}}{\sum_{i \in N} \sum_{j \in N_i} \Delta_{lij}},$$

$$l \in L, \qquad j \in N_i, \qquad i \in N.$$

- Профиль первого уровня публикации *l*:

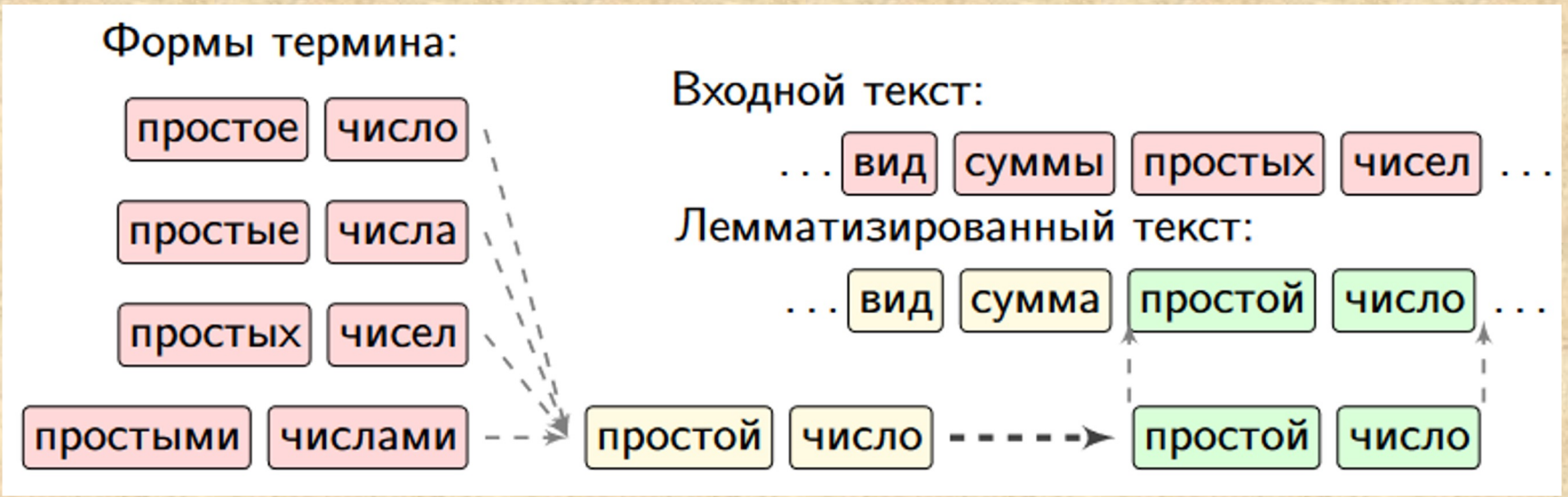
$$X_l = (X_{l1}, \ldots, X_{li}, \ldots, X_{ln}), \text{ где } X_{li} = \sum_{j \in N_i} x_{lij}, \; l \in L, \; i \in N.$$

# ПОИСК ТЕРМИНОВ

Основа алгоритма составления профиля – поиск терминов

Чем больше терминов в работе соответствует заданному направлению исследований, тем больше соответствующая оценка в профиле

Поиск терминов осуществляется при помощи синтаксического анализа



# ПРОФИЛЬ УЧЕНОГО

Обозначим

- $K$ – множество ученых
- $r(l)$ – количество авторов $l$-ой публикации
- $\omega(k,\ l) = \begin{cases} 1, \text{ если } k\text{–й ученый является автором } l\text{–ой публикации;} \\ 0, \text{ в противном случае;} \end{cases}$

Профили второго и первого уровня $k$-го ученого считаются на основе его публикаций

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

$$Y_{i}^{k} = \sum_{j \in N_i} y_{ij}^{k}, \qquad k \in K, \qquad i \in N$$

# РАССТОЯНИЕ МЕЖДУ ПРОФИЛЯМИ

Предлагается применять следующее расстояние между двумя профилями, задаваемыми стохастическими векторами $p = (p_1, \dots, p_n)$ и $q = (q_1, \dots, q_n)$:

$$d(p,q) = 1 - \sum_{j=1}^{n} \min(p_j, q_j) = \frac{1}{2} \sum_{j=1}^{n} |p_j - q_j|$$

Замечание. В данной метрике расстояние между профилями первого уровня всегда не превосходит расстояние между профилями второго уровня тех же объектов

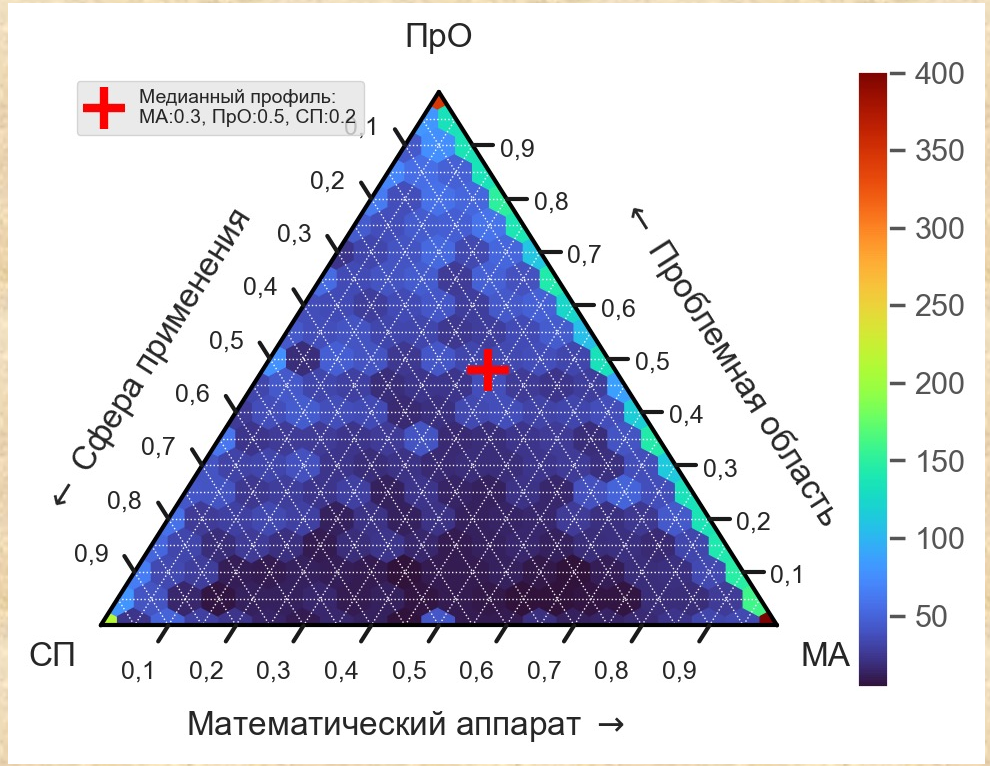
# АНАЛИЗ ПРОФИЛЕЙ ПУБЛИКАЦИЙ И УЧЕНЫХ

- Профили ученых рассчитываются на основе профилей публикаций. Профили публикаций – на основе терминов теории управления
- Корпус ИПУ РАН - 19,5 тыс. русскоязычных публикаций сотрудников Института проблем управления им. В. А. Трапезникова РАН (БД ИПУ РАН) за 1921–2022 годы (https://www.ipu.ru)
- Общая схема построения профилей

1. Извлечение и предварительная обработка текстов (токенизация, лемматизация, выявление n-грамм и преобразование текстов в векторное представление)
2. Расчет и анализ профилей публикаций
3. Расчет и анализ профилей авторов публикаций (ученых)

# ПРОФИЛИ ПУБЛИКАЦИЙ ИПУ РАН В ТУ

- Расчет сначала профилей 1-го уровня для этих публикаций
- Расчет профилей 0-го уровня
(Математический аппарат, Проблемная область, Сфера применения)




Тепловая карта профилей 0-го уровня публикаций БД ИПУ РАН

Медианный профиль на основе текстов

**(0,3; 0,5; 0,2)**

**В ТУ медианный профиль публикации «затрагивает» на 30 % математический аппарат, 50 % предметную область и 20 % сферу применения**

# ПРОФИЛИ УЧЕНЫХ ИПУ РАН В ТУ

- Расчет профилей 1-го уровня для ученых – авторов публикаций (6,4 тыс.)
- Визуализация преобразованных профилей в двумерном пространстве
  - Расчет расстояние между профилями ученых
  - Нелинейное снижение размерности пространства профилей при помощи алгоритма машинного обучения UMAP

Профили ученых в двумерном пространстве


Крестиками профили авторов данного доклада и некоторых других ученых ИПУ РАН:

- Губанов Д.А.
- Кузнецов О.П.
- Новиков Д.А.
- Поляк Б.Т.
- Суховеров В.С.
- Хлебников М.В.
- Чхартишвили А.Г.

# ПРИМЕРЫ ПРОФИЛЕЙ УЧЕНЫХ 1-ГО УРОВНЯ

## Характерные профили 1-го уровня*

### Пример профиля 1-го уровня: Новиков Д.А.

| 1-й уровень | Значение |
|---|---|
| Теория игр | 0,032 |
| Теория графов | 0,016 |
| Теория вероятностей и мате | 0,014 |
| Теория множеств и отношен | 0,008 |
| Комбинаторика | 0,005 |
| Теория управления в органи | 0,408 |
| Теория выбора и принятия р | 0,111 |
| Кибернетика и системный а | 0,064 |
| Исследование операций | 0,055 |
| Информационная безопасно | 0,035 |
| Образование | 0,070 |
| Социальные системы | 0,027 |
| Социально-экономические с | 0,011 |
| Робототехника | 0,011 |
| Военное дело | 0,008 |

### Пример профиля 1-го уровня: Кузнецов О.П.

| 1-й уровень | Значение |
|---|---|
| Теория графов | 0,085 |
| Математическая логика | 0,038 |
| Дифференциальные и интег | 0,017 |
| Функциональный анализ | 0,012 |
| Теория алгоритмов и форма | 0,012 |
| Искусственный интеллект и | 0,319 |
| Теория управления в органи | 0,133 |
| Информационная безопасно | 0,104 |
| Теория выбора и принятия р | 0,087 |
| Теория автоматического упр | 0,032 |
| Социальные системы | 0,050 |
| Робототехника | 0,011 |
| Образование | 0,008 |
| Технологические процессы | 0,005 |
| Авиация | 0,000 |

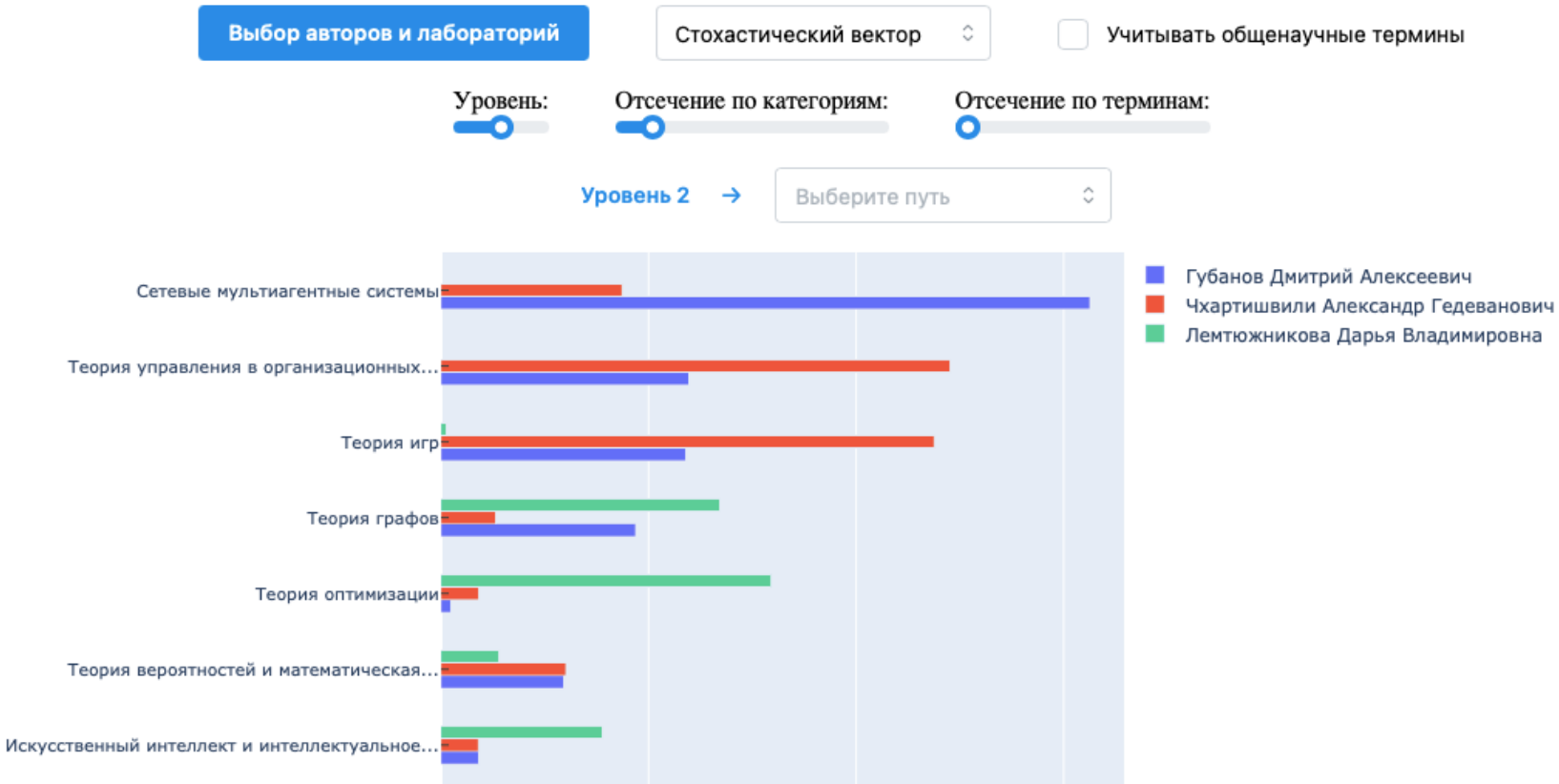
### Пример профиля 1-го уровня: Поляк Б.Т.

| 1-й уровень | Значение |
|---|---|
| Теория оптимизации | 0,113 |
| Информатика и теория инфо | 0,079 |
| Дифференциальные и интег | 0,054 |
| Алгебра и теория чисел | 0,012 |
| Математическая логика | 0,007 |
| Теория автоматического упр | 0,491 |
| Кибернетика и системный а | 0,058 |
| Навигация и управление дви | 0,042 |
| Мехатроника | 0,020 |
| Теория выбора и принятия р | 0,020 |
| Образование | 0,012 |
| Робототехника | 0,004 |
| Космос | 0,004 |
| Социальные системы | 0,002 |
| Военное дело | 0,000 |

**В усеченном виде. Темы упорядочены по убыванию «внимания» (значения компоненты профиля); для каждого аспекта приведены только пять наиболее весомых тем.*

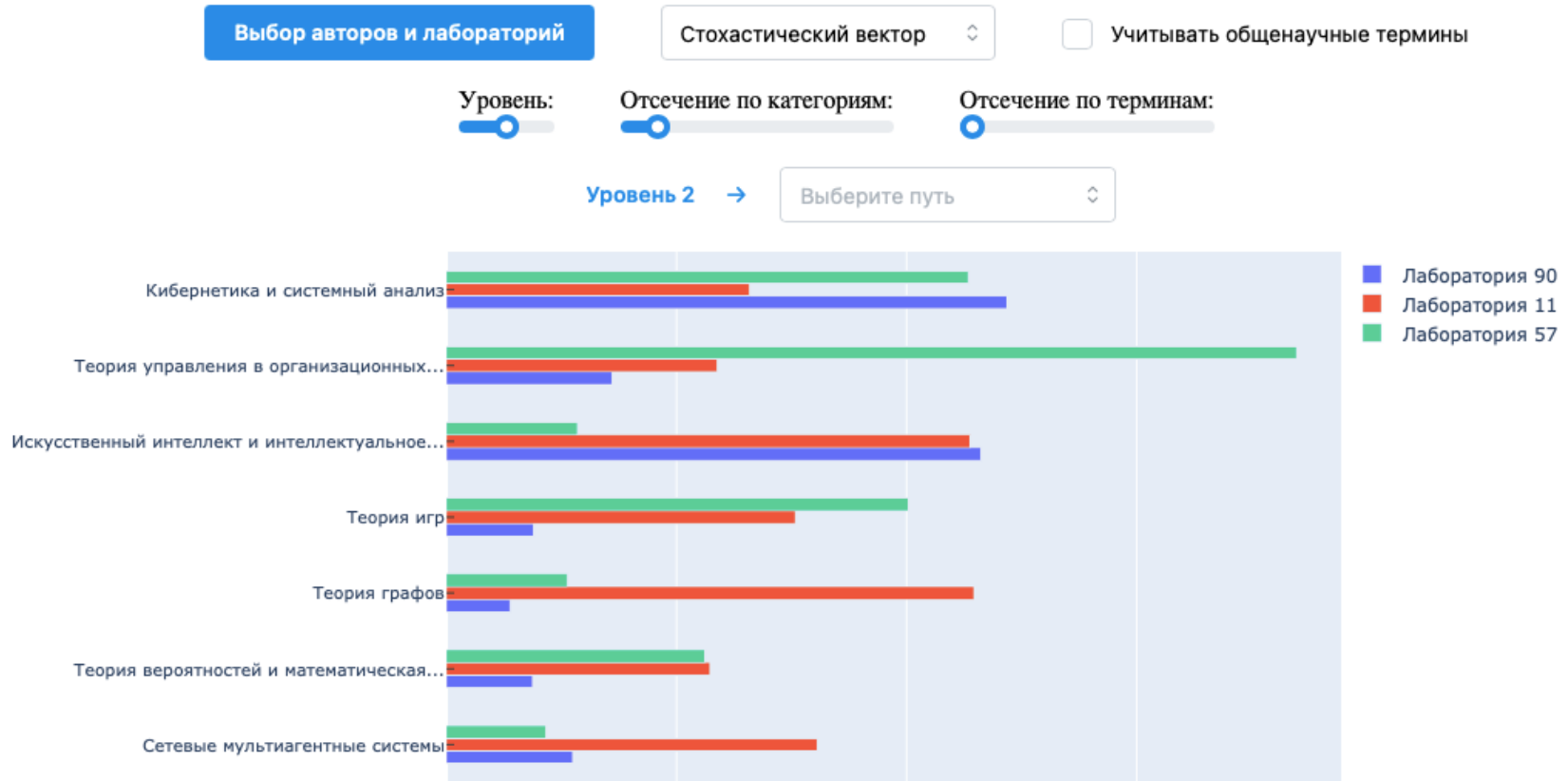
# ИНТЕРФЕЙС: ПРОФИЛИ УЧЕНЫХ

Главная › Профили ученых



# ИНТЕРФЕЙС: ПРОФИЛИ ЛАБОРАТОРИЙ

Главная › Профили ученых



# СЦЕНАРИИ ИСПОЛЬЗОВАНИЯ ИСАНД

На данный момент реализованы следующие методы использования ИСАНД конечным пользователем

1. **Какие термины чаще всего используются в публикациях заданного агента?**
2. **Каким ученым отправить приглашение на данную научную конференцию?**
3. **Как можно определить, что агент начал публиковаться по нехарактерному для него научному направлению?**
4. **Как меняется спектр научных направлений агента со временем?**
5. **Какие агенты являются наиболее активными в данном научном направлении?**
6. **Кто мог бы выступить рецензентом данной статьи (доклада)?**

**Можно создавать свои сценарии и реализующие их методы, используя REST-сервисы ИСАНД (в том числе SPARQL)**

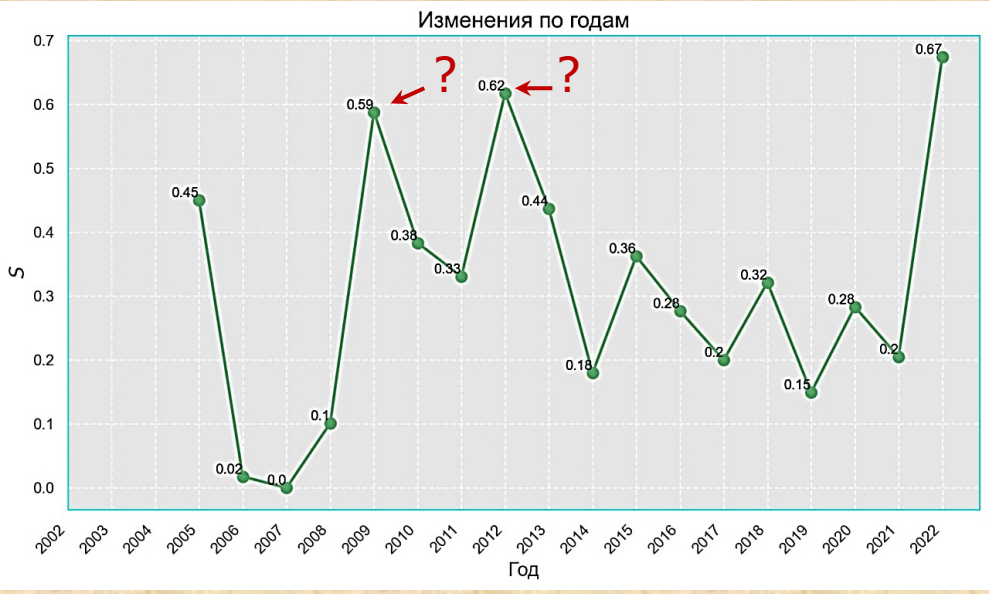
# ПРИМЕР РЕАЛИЗАЦИИ СЦЕНАРИЯ

## Как меняется спектр научных направлений агента со временем?

Задано:

- Персона “Чхартишвили А.Г.”
- Время изменения: Δ = 1

Результат расчета:




| год | публикаций | с учетом соавторства | *s* | Δ компонент |
|---|---|---|---|---|
| 2002 | 1 | 0,50 | | Теория управления в орган |
| 2003 | 0 | 0,00 | | |
| 2004 | 2 | 1,50 | | Теория управления в орган |
| 2005 | 8 | 5,50 | 0,45 | Социально-экономические |
| 2006 | 1 | 0,50 | 0,02 | Социальные системы (-0.0 |
| 2007 | 2 | 2,00 | 0,00 | Теория игр (0.0), Теория у |
| 2008 | 5 | 4,00 | 0,10 | Теория управления в орган |
| 2009 | 10 | 4,17 | 0,59 | Сетевые мультиагентные с |
| 2010 | 8 | 2,92 | 0,38 | Теория игр (-0.26), Теория |
| 2011 | 10 | 4,92 | 0,33 | Сетевые мультиагентные с |
| 2012 | 4 | 2,50 | 0,62 | Теория управления в орган |
| 2013 | 6 | 2,92 | 0,44 | Теория управления в орган |
| 2014 | 10 | 6,00 | 0,18 | Теория управления в орган |
| 2015 | 4 | 2,50 | 0,36 | Теория игр (0.3), Теория у |
| 2016 | 7 | 3,50 | 0,28 | Сетевые мультиагентные с |
| 2017 | 5 | 2,50 | 0,20 | Алгебра и теория чисел (0 |
| 2018 | 12 | 6,50 | 0,32 | Алгебра и теория чисел (-0 |
| 2019 | 10 | 4,11 | 0,15 | Функциональный анализ (0 |
| 2020 | 10 | 5,03 | 0,28 | Функциональный анализ (- |
| 2021 | 13 | 6,75 | 0,20 | Алгебра и теория чисел (-0 |
| 2022 | 6 | 2,50 | 0,67 | Сетевые мультиагентные с |

# ЗАКЛЮЧЕНИЕ

- Разрабатываемая система ИСАНД в целом адекватно оценивает компетенции ученых
- Анализ публикаций ожидаемо показывает, что основные исследования ученых ИПУ РАН ведутся в области прикладных теорий
- В дальнейшем планируется развитие ИСАНД с точки зрения
  - данных (расширение базы за счет российских журналов, конференций, организаций по теории управления)
  - функционала (новые методы анализа и сценарии использования)
  - интерфейса (новые методы визуализации результатов)

# ЗАКЛЮЧЕНИЕ

- Разрабатываемая система ИСАНД в целом адекватно оценивает компетенции ученых
- Анализ публикаций ожидаемо показывает, что основные исследования ученых ИПУ РАН ведутся в области прикладных теорий
- В дальнейшем планируется развитие ИСАНД с точки зрения
  - данных (расширение базы за счет российских журналов, конференций, организаций по теории управления)
  - функционала (новые методы анализа и сценарии использования)
  - интерфейса (новые методы визуализации результатов)

**Желаете присоединиться к проекту?**