Информационная система анализа научной деятельности (на примере теории управления)

Чхартишвили А.Г. (ИПУ РАН)
ВСПУ, 20 июня 2024 г.

ИСАНД:
основные вопросы

Зачем нужен ИСАНД?

ИСАНД – специализированный классификатор научных объектов, разработанный на основе онтологии научного знания, составленной при помощи экспертов.

Почему именно ИСАНД?

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.

Сегодня ИСАНД использует онтологию научного знания теории управления.

ИСАНД позволяет встроить в систему онтологию любой области знаний.

Целевая аудитория ИСАНД

- студенты
- аспиранты
- научные сотрудники
- преподаватели вузов
- независимые исследователи
- редакции журналов
- организаторы конференций
- руководители научных организаций и подразделений

Что спросить у ИСАНД?

Студентам, аспирантам, ученым	1. Какие ученые, организации, журналы, конференции являются наиболее важными для данного направления? 2. В какой журнал/конференцию лучше подать статью/доклад? 3. Какие публикации по научному направлению наиболее широко цитируются? 4. Как меняется спектр научных направлений со временем?
Редакциям журналов и организаторам конференций	1. Кто вносит основной вклад в показатели журнала (пишут и ссылаются)? 2. Как меняется спектр научных направлений журнала со временем? 3. Кому направить на рецензирование статью? 4. Соответствует ли список литературы содержанию статьи?

Что спросить у ИСАНД?

Руководителям научных организаций и коллективов	1. Какие научные направления являются наиболее популярными? 2. Какова динамика научных направлений во времени? 3. Кто в данном научном направлении? 4. Кого имеет смысл пригласить пригласить к сотрудничеству?
Организаторам науки	1. Кто ведет научную деятельность в данном научном направлении? 2. Какова «географическая структура» данного научного направления? 3. Как спрогнозировать перспективные направления исследований? 4. Как оценить эффективность научного направления?

Некоторые публикации

- Кузнецов О.П., Суховеров В.С. Онтологический подход к оценке тематики научного текста // Онтология проектирования. – 2016. – Т. 6, № 1. – С. 55–66.
- Губанов Д.А., Новиков Д.А., Чхартишвили А.Г. Социальные сети: модели информационного влияния, управления и противоборства. 3-е изд., перераб. и дополн. М.: МЦНМО, 2018. (Social Networks: Models of information influence, control and confrontation. Springer, 2019.)
- Теория управления: словарь системы основных понятий. - М.: ЛЕНАНД, 2024.
- Губанов Д.А., Кузнецов О.П., Курако Е.А., Лемтюжникова Д.В., Новиков Д.А., Чхартишвили А.Г. Информационная система анализа научной деятельности (ИСАНД) в области теории управления // Проблемы управления. 2024. № 3. С. 42-65.

Основные определения

Научный объект – публикация/ученый/журнал/конференция/ научная организация.

Профиль (тематический) – вектор, координаты которого описывают положение научного объекта в тематическом пространстве науки.

Основные подходы ИСАНД

Лингвистический (анализ текстов)

информационный поиск
извлечение информации
категоризация текстов
…

Сетевой (анализ связей)

семантические сети
структурные характеристики
меры влияния
…

ИСАНД: онтология научного знания

Уровни онтологии научного знания по теории управления

Нулевой уровень - метафакторы

- 3 фактора: **Математический аппарат**, **Предметная область**, **Сфера применения**

Первый уровень - факторы

- 52 фактора. Пример: **Исследование операций**

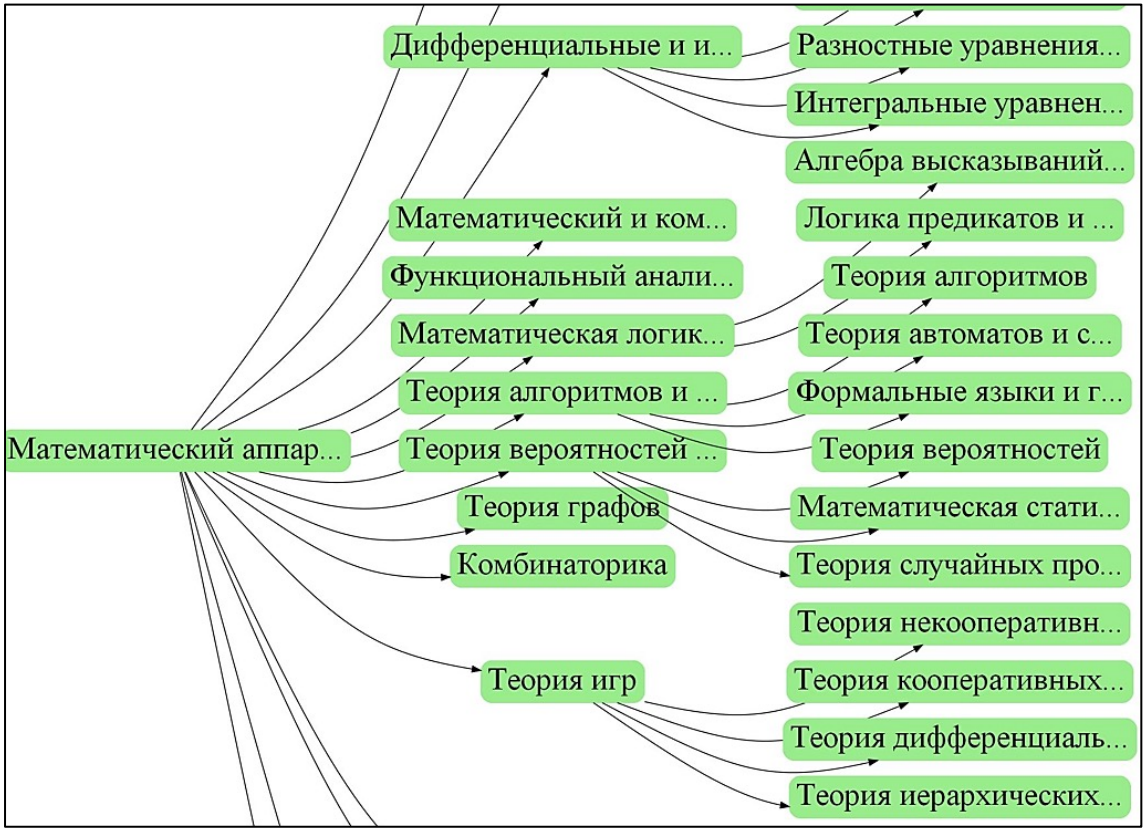
Второй уровень - подфакторы

- 161 фактор. Пример: **Управление запасами**

Третий уровень - базовые термины

- 3032 термина. Пример: **Оптимизация объемов сырья**

Онтология научного знания



ИСАНД: тематические профили

Как вычисляется тематический профиль

Вычисляются на основе терминов теории управления

Профиль публикации – вектор, размерность которого соответствует количеству тем или подтем

Вычисляются на основе профилей публикаций

Профиль ученого – нормированная сумма профилей его публикаций с учетом количества соавторов

Профиль организации – нормированная сумма профилей публикаций с учетом количества соавторов и их аффилиаций

Профиль журнала (конференции) – нормированная сумма профилей публикаций журнала (конференции)

Профиль публикации

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

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

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.

Примеры профилей 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

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 уп	0,032
Социальные системы	0,050
Робототехника	0,011
Образование	0,008
Технологические процессы	0,005
Авиация	0,000

Поляк Б.Т.

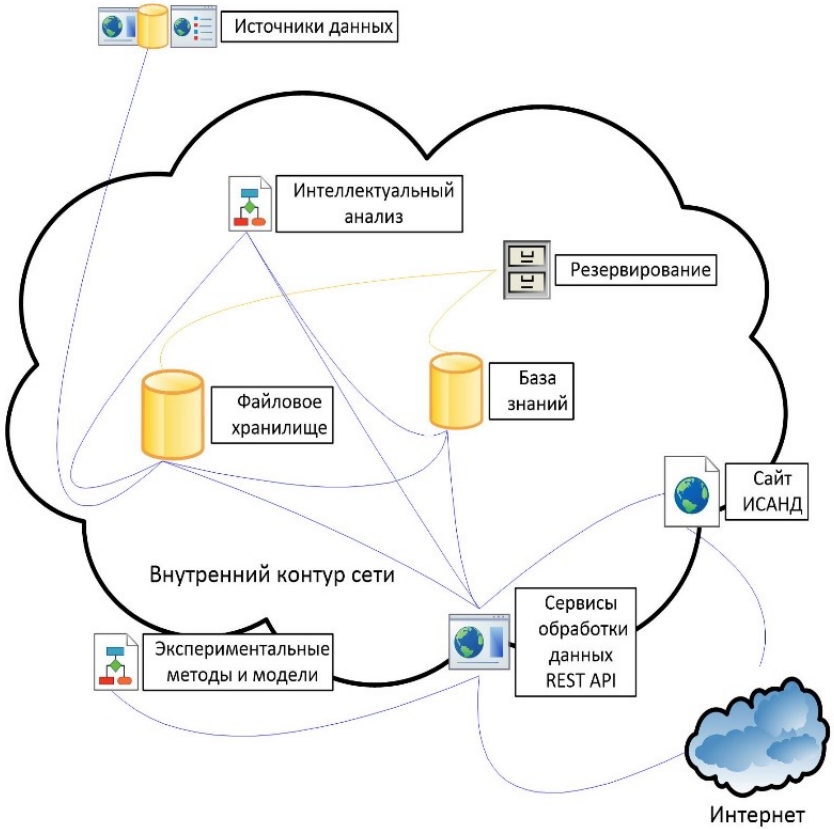
1-й уровень	Значение
Теория оптимизации	0,113
Информатика и теория инфо	0,079
Дифференциальные и интег	0,054
Алгебра и теория чисел	0,012
Математическая логика	0,007
Теория автоматического уп	0,491
Кибернетика и системный а	0,058
Навигация и управление дви	0,042
Мехатроника	0,020
Теория выбора и принятия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

Темы упорядочены по убыванию значения компоненты профиля; для каждого аспекта приведены пять наиболее весомых тем

18

ИСАНД: структура информационной системы

Архитектура информационной системы

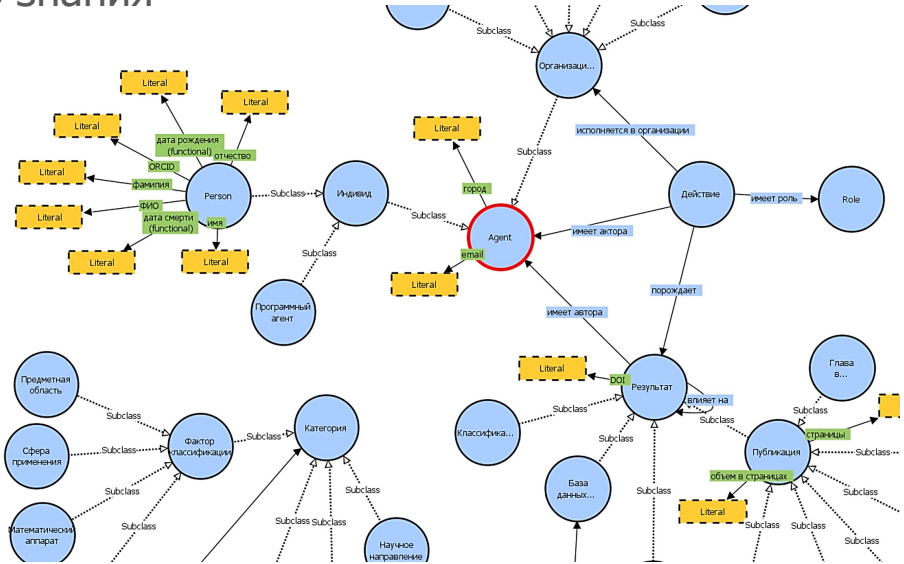


- файловое хранилище и подсистема загрузки данных
- база знаний
- подсистема интеллектуального анализа информации
- подсистема экспериментальных методов и моделей
- подсистема резервирования
- сервисы обработки данных
- сайт ИСАНД

Онтология научной деятельности

Включает онтологию научного знания

- **45** классов
  напр., «Публикация»
- **23** типа объектных
  отношений напр.,
  «*влияет на*»
- **37** типов простых
  свойств напр.,
  «*название*»
- 7 видов аннотаций,
  582 логические аксиомы



Онтология является основой базы знаний ИСАНД (*knowledge graph*).
Позволяет выполнить интеграцию с внешними открытыми источниками информации (используя технологии LinkedData)

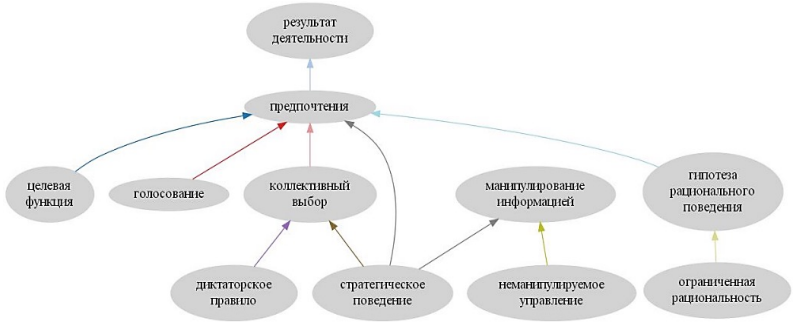
ИСАНД:
анализ связей

Анализ научных сетей

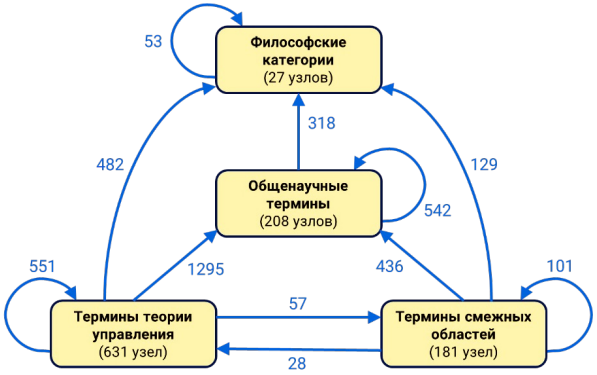
Основа построения «научных» сетей – онтология научной деятельности и научного знания (граф знаний)

- Терминологические сети
- Сети областей научного знания
- Сети сотрудничества
- Сети цитирования
- …

Терминологическая сеть структуры ТУ



Фрагмент графа связей между значимыми терминами



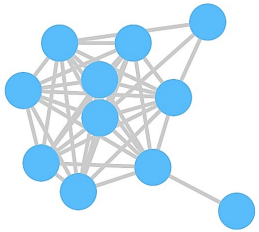
Сеть между группами терминами

Сети соавторства публикаций

Узлы – ученые, а ненаправленная дуга – наличие хотя бы одной совместной публикации
Сеть соавторства сотрудников лабораторий ИПУ РАН:

- Более 400 вершин
- Около 1 тыс. связей соавторства

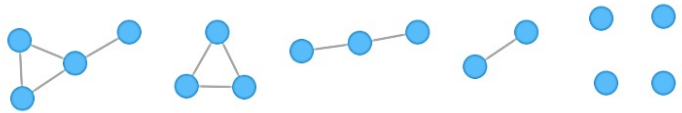
Пример лаборатории 1



Граф связный и плотный

Тесное взаимодействие сотрудников друг с другом при подготовке публикаций

Пример лаборатории 2



Несколько компонент связности

Отдельные группы в данной лаборатории работают автономно

ИСАНД:
анализ текстов

Выделение структурных элементов из публикации

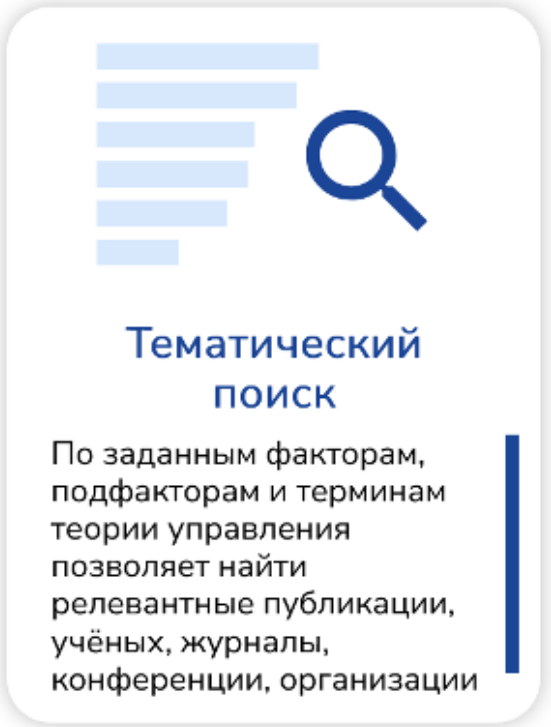
А также:

- список источников;
- дата публикации;
- название журнала;
- поддержка;

и другие структурные элементы.

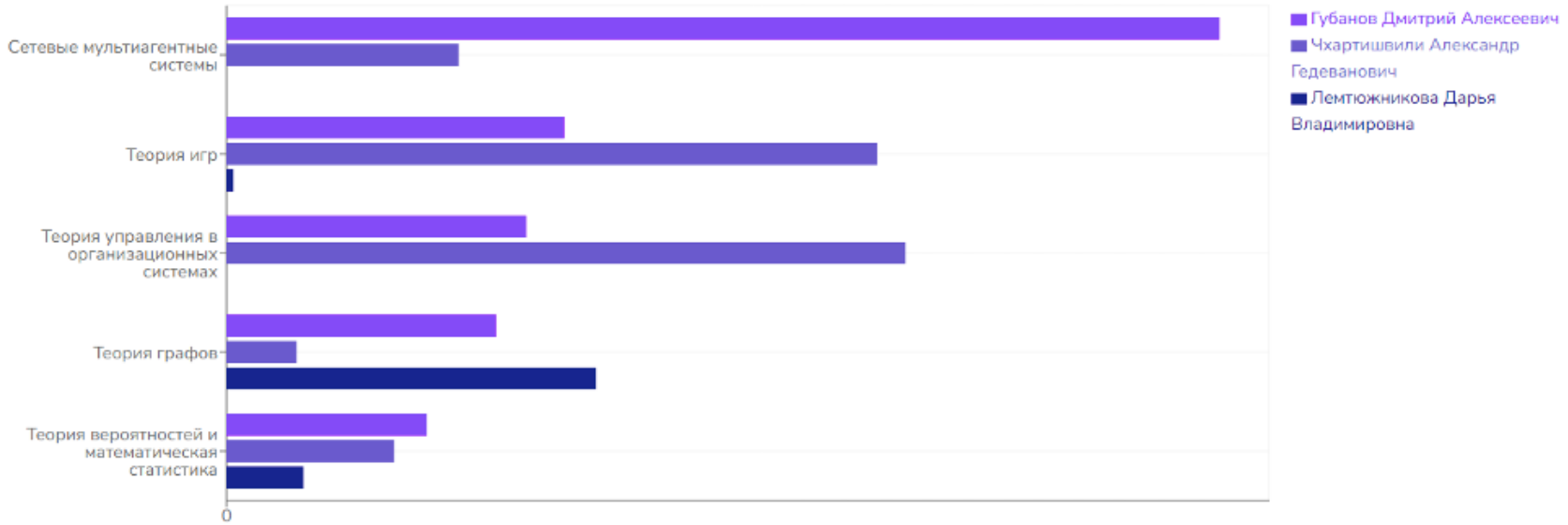
ИСАНД (isand.ipu.ru): пользовательский интерфейс

Тематический поиск



Позволяет отобрать релевантные научные объекты по заранее заданным факторам (1-й уровень тематической классификации), подфакторам (2-й уровень тематической классификации) и терминам теории управления

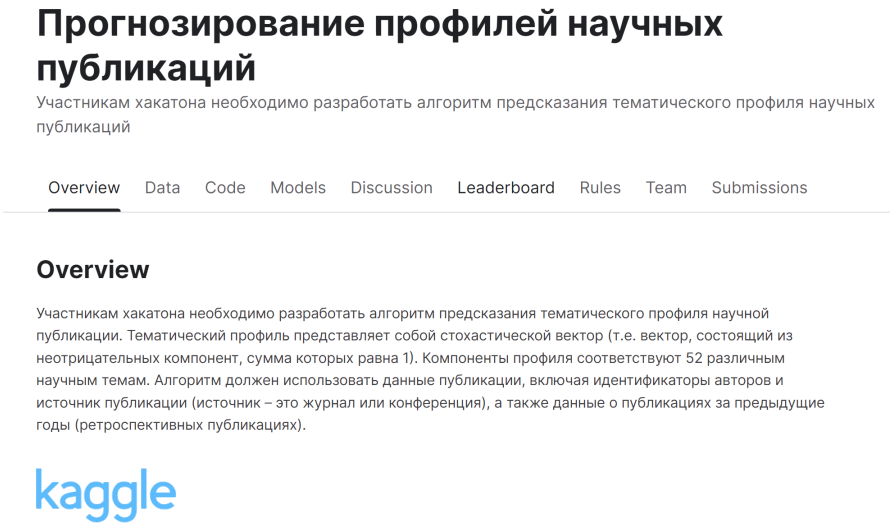
Профили ученых



ИСАНД: хакатон

Хакатон по прогнозированию профилей научных публикаций

Приглашаем к участию (с 17 июня по 26 июня):
https://www.kaggle.com/competitions/predicting-scientific-publication-profiles



Спасибо за внимание!

Чхартишвили А.Г. (ИПУ РАН)
sandro_ch@mail.ru