научно-практическая конференция

Science Analytics 2026

статистика, аналитика и оценка научных исследований

Мастерская знаний:

информационная среда будущего для коллективного поиска, понимания, систематизации, производства и передачи научных знаний

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН, руководитель лаборатории машинного обучения и семантического анализа

МГУ

Институт искусственного интеллекта МГУ им. М.В. Ломоносова

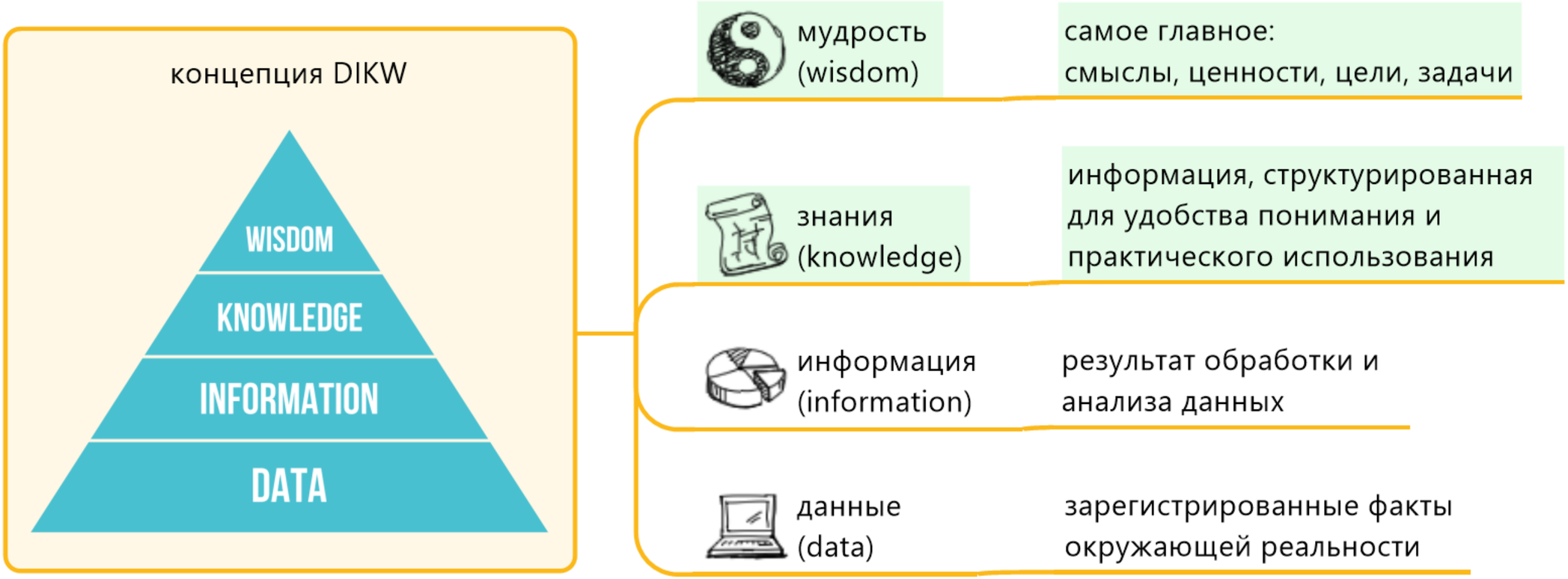
Концепция «Мастерской знаний»

«Огромное и все возрастающее богатство знаний разбросано сегодня по всему миру. Этих знаний, вероятно, было бы достаточно для решения всего громадного количества трудностей наших дней, но они рассеяны и неорганизованы. Нам необходима очистка мышления в своеобразной **мастерской ума**, **где можно получать, сортировать, суммировать, усваивать, разъяснять и сравнивать знания и идеи.**» – *Герберт Уэллс, 1940*

(An immense and ever-increasing wealth of knowledge is scattered about the world today; knowledge that would probably suffice to solve all the mighty difficulties of our age, but it is dispersed and unorganized. We need a sort of mental clearing house for the mind: **a depot** *(склад с мастерской)* **where knowledge and ideas are received, sorted, summarized, digested, clarified and compared** – *Herbert Wells, 1940*)

Теперь технологии IR/NLP/LLM позволяют ставить и решать такие задачи

Что такое «знания»



Технологии больших языковых моделей (Large Language Model, LLM) позволяют выделять знания и идеи из текста и систематизировать их

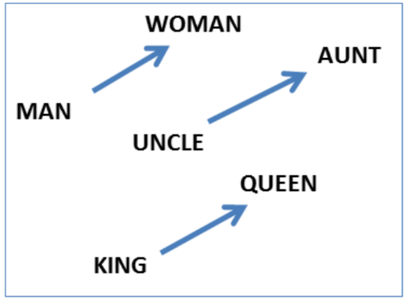
Эволюция подходов в обработке естественного языка

Декомпозиция задач по уровням «пирамиды NLP»

- морфологический анализ, лемматизация, опечатки,…
- синтаксический анализ, выделение терминов, NER,…
- семантический анализ, выделение фактов, тем,…

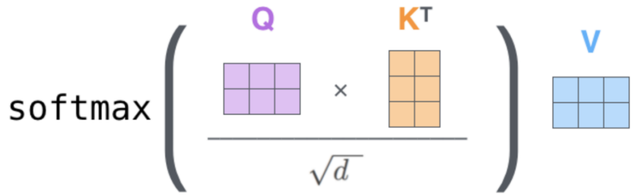
Векторные представления слов (embeddings)

- модели дистрибутивной семантики:
  word2vec [Mikolov, 2013], FastText [Bojanowski, 2016],…
- тематические модели LDA [Blei, 2003], ARTM [2014],…



Большие языковые модели (БЯМ, LLM)

- рекуррентные нейронные сети: LSTM, GRU,…
- «end-to-end» модели внимания, трансформеры:
  Google NMT [2016], BERT [2018], GPT-3 [2020], GPT-4 [2023],…



От поиска информации к «Мастерской знаний»

Недостатки обычного поиска:

- как искать новые знания?
- что делать с найденным?

Мастерская знаний – инструментарий для работы с текстовыми источниками на всём жизненном цикле научного проекта:

- ищу текстовые документы – чтобы сохранять их и накапливать
- накапливаю – чтобы их перечитывать, анализировать, понимать
- понимаю – чтобы получать, обрабатывать, систематизировать *знания*
- систематизирую – чтобы применять и передавать *знания и мудрость*

Теперь технологии IR/NLP/LLM позволяют ставить и решать такие задачи

Научный поиск на основе LLM и ИИ-агентов

perplexity

Consensus

NotebookLM

SEMANTIC SCHOLAR

Elicit

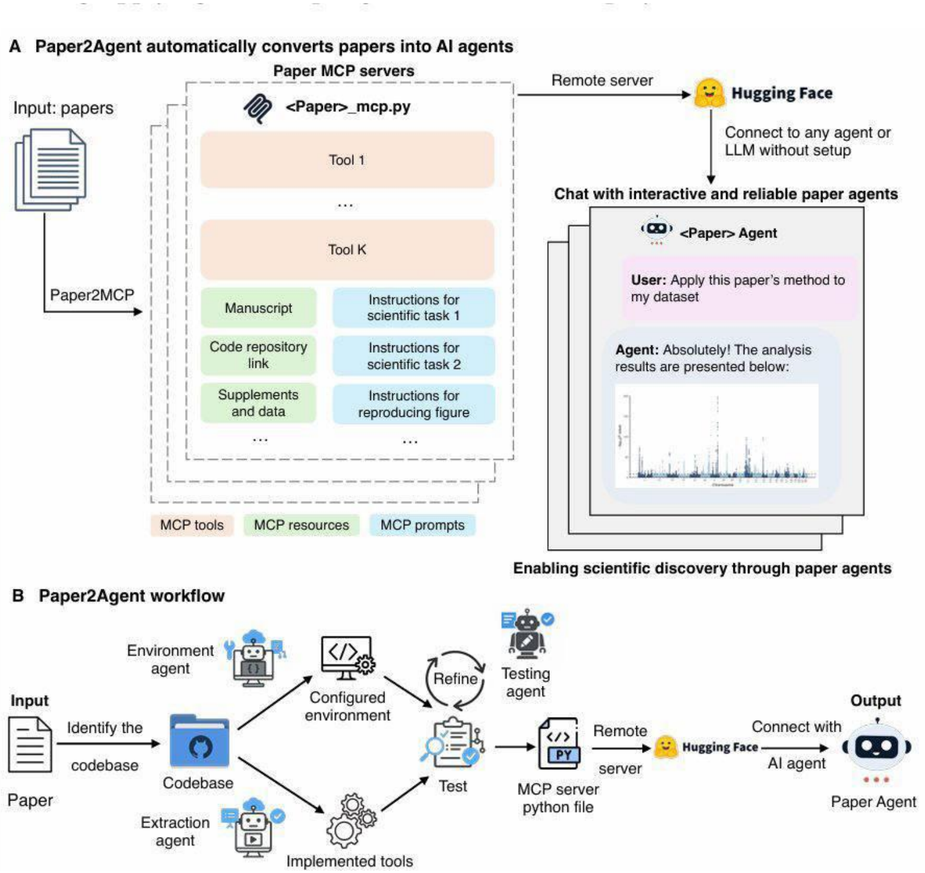
ChatPDF

ResearchRabbit

Открытые проблемы ИИ-систем:

- как зафиксировать долгосрочный тематический поисковый интерес?
- как выделять и как обновлять знания?
- как обеспечить ясность представления знаний — «посмотрел и всё понял»?
- как включить «коллективный разум»?

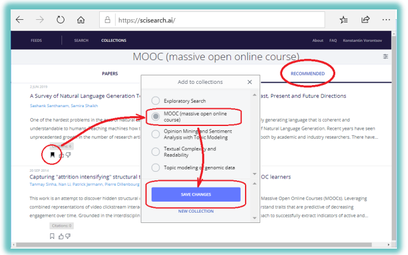
Paper2Agent — интерактивный ИИ-агент
https://github.com/jmiao24/Paper2Agent



Концепция сервисов «Мастерской знаний»

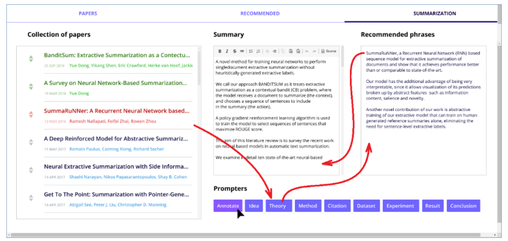
Подборка текстов фиксирует тематику поискового интереса пользователя или группы
Расширенная подборка: + результаты поиска семантически близких текстов

Поисково-рекомендательные сервисы:



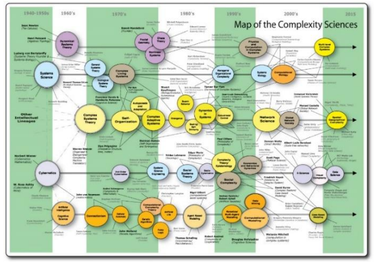
- поиск семантически близких документов по подборке
- контекстный поиск по фрагменту документа из подборки
- мониторинг новых документов по тематике подборки

Аналитические сервисы:



- полуавтоматическое реферирование подборки
- тематизация, картирование, онтологизация подборки
- хронологизация, выявление трендов по тематике подборки
- контент-анализ, сбор и анализ фактов из документов подборки

Коммуникативные сервисы:



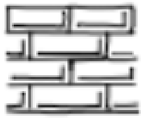
- совместное обсуждение, анализ, использование подборок
- создание нового контента в соавторстве на основе подборки

Декларация принципов «Мастерской знаний»

1. **Тематичность.** Рабочая среда пользователя образуется тематическими подборками
2. **Текстуальность.** Знания содержатся в текстах, написанных людьми для людей
3. **Коллегиальность.** Формы представления знаний служат взаимопониманию в группе
4. **Креативность.** Пользователи создают в среде свой контент, информационные продукты
5. **Доверенность.** Меньше генерации, больше экстракции, источников, ссылок
6. **Антропоцентричность.** Интенсификация творчества людей — цель, а не средство
7. **Когнитивность.** Представление знаний учитывает особенности восприятия и памяти
8. **Мультиязычность.** Автоматический перевод с языков источника на язык пользователя
9. **Расширяемость.** Платформа МЗ поддерживает возможность добавления сервисов
10. **Открытость.** Базовые функции общедоступны ради устранения цифрового неравенства
11. **Экономичность.** Чтобы сделать мир умнее, сначала сделать монетизируемый продукт
12. **Социоцентричность.** Проектируя систему, предвидеть социальные практики и эффекты

Миссия Мастерской Знаний

— устранять *барьеры* между человеком и знанием



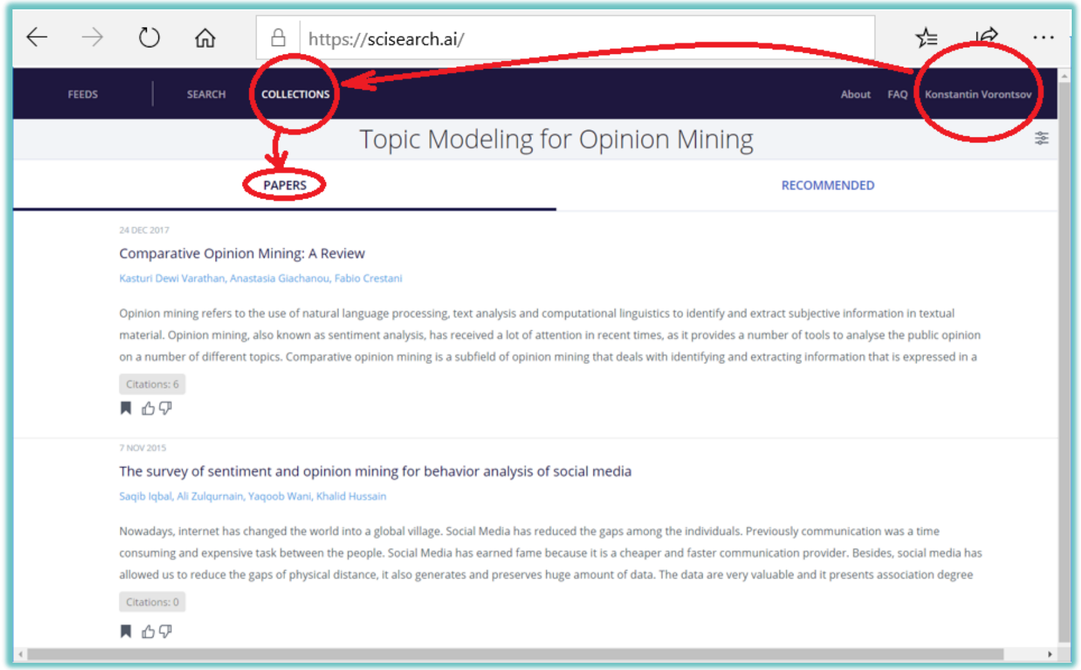
технологические — из-за избыточности, неструктурированности, ненадёжности информации

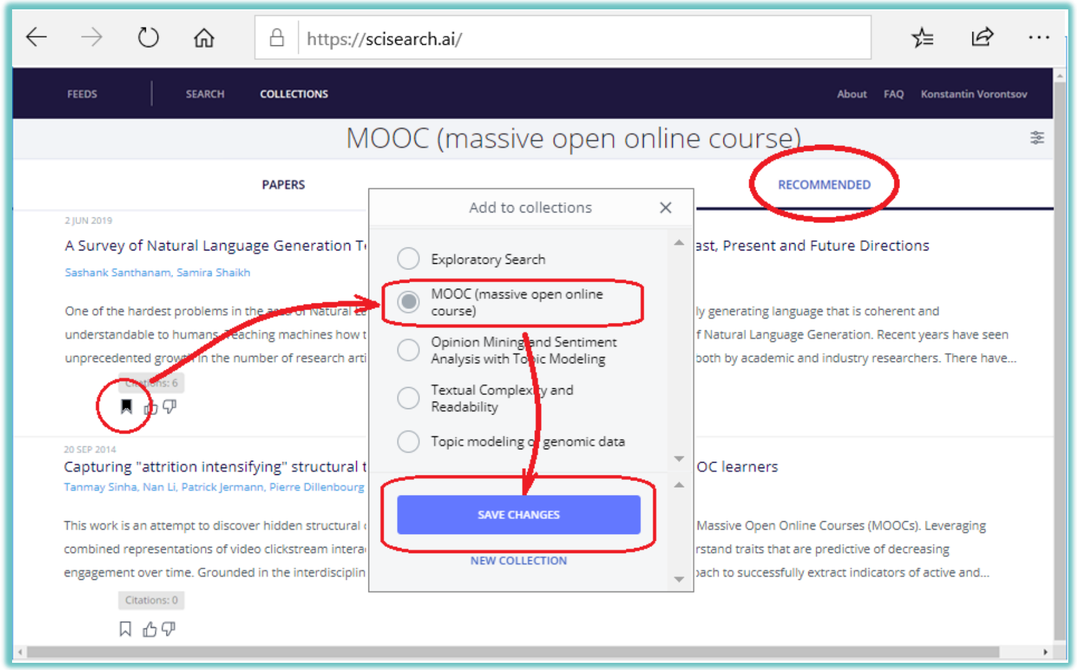
когнитивные — из-за ограниченности наших возможностей запоминания, понимания, анализа

коммуникативные — из-за различий в мотивациях, уровне компетенций, социальном и служебном положении

Сервис поиска и ранжирования рекомендаций

Цель: помочь пользователю быстро собрать тематическую подборку
по своей информационной потребности, бегло знакомясь с документами

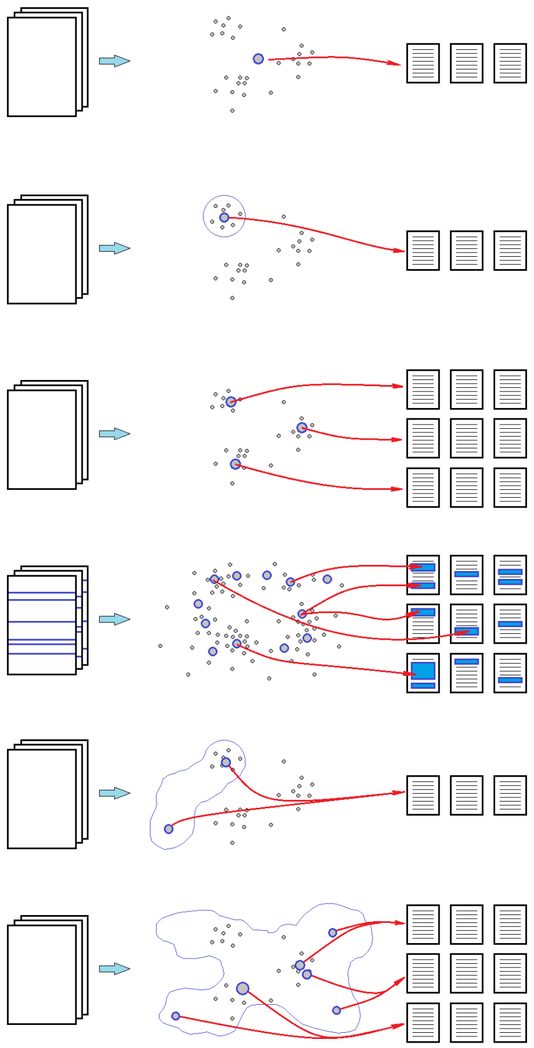




Герасименко Н.А., Ватолин А.С., Янина А.О., Воронцов К.В. SciRus: легкий и мощный мультиязычный энкодер для научных текстов // Доклады РАН, 2024, том 520

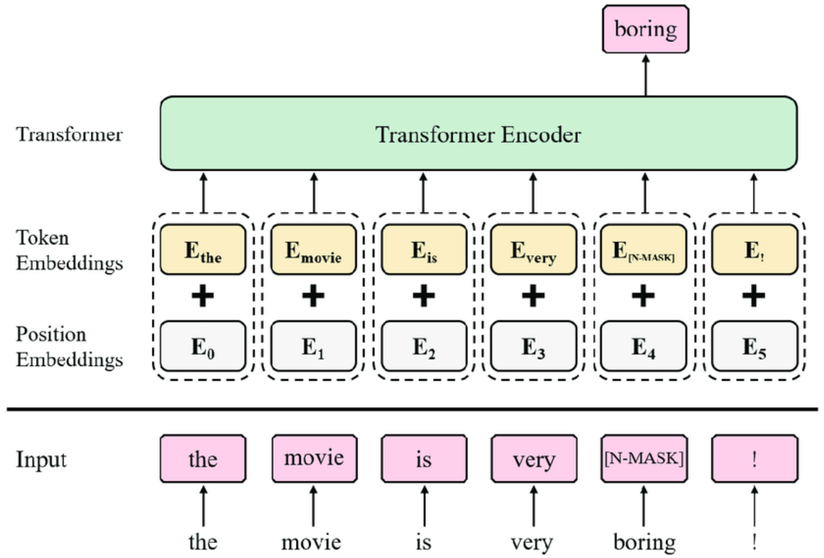
Ватолин А.С., Герасименко Н.А., Янина А.О., Воронцов К.В. RuSciBench: открытый бенчмарк для оценки семантических векторных представлений научных текстов на русском и английском языках // Доклады РАН, 2024, том 520

Стратегии векторного поиска документов



1. Поиск по среднему вектору подборки
   (самая простая, но не самая удачная стратегия)
2. **Поиск по документу из подборки или нескольким близким к нему документам**
3. Разбиение подборки на кластеры и
   поиск по центральным документам кластеров
4. Разбиение документов подборки на сегменты
   и поиск по сегментам документов
5. Поиск по документам смежной тематики
   для документа или части документов подборки
6. Поиск по тематике, смежной для всей подборки

LLM-кодировщики для научных текстов



- **SciBERT (2019)** *Beltagy et al.*
  SciBERT: A pretrained language model for scientific text
- **SPECTER (2020)** *Cohan et al.*
  SPECTER: Document-level representation learning using citation-informed transformers
- **LaBSE (2020)** *Feng et al.*
  Language agnostic BERT sentence embedding
- **MPNet (2020)** *Song et al.*
  MPNet: Masked and permuted pre-training for language understanding
- **SPECTER-2 (2022)** *Singh et al.*
  SciRepEval: A multi-format benchmark for scientific document representations
- **SciNCL (2022)** *Ostendorff et al.*
  Neighborhood contrastive learning for scientific document representations with citation embeddings
- **mE5 (2024)** *Wang et al.*
  Multilingual E5 text embeddings: A technical report. 2024.

Модель SciRus: мотивации исследования

МГУ

Модель должна быть применима в русскоязычных сервисах для экстрактивных задач: поиска, рекомендации, классификации, анализа научных текстовых документов — в различных сервисах и приложениях («Мастерская знаний», eLibrary.ru, научные электронные библиотеки)

Требования к модели:

- минимальный размер (23M параметров)
- при качестве, сопоставимом с лучшими (SOTA) моделями
- возможность вычисления эмбедингов без GPU
- мультиязычность: английский, русский, китайский и др.
- возможность дообучения модели по данным о цитировании
- оценивание качества — по стандартным + новым benchmark-ам

Герасименко Н.А., Ватолин А.С., Янина А.О., Воронцов К.В. SciRus: легкий и мощный мультиязычный энкодер для научных текстов. Доклады РАН, 2024

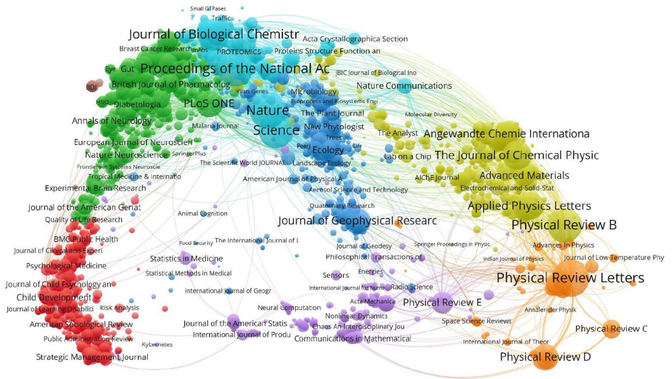
Данные для обучения модели научных текстов

Данные для обучения — title+abstract:

- **S2ORC — Semantic Scholar Open Research Corpus**
  30M (12B токенов), 85% en, 2% ru
- **eLibrary**:
  8.5M (2B токенов) ru
  5.2M (1.2B токенов) ru+en
- **ScienceChina** (title+abstract):
  5M аннотаций (0.85 токенов) en+zh

Данные для дообучения:

- **S2AG — Semantic Scholar Academic Graph**
  источники: Crossref, PubMed, Unpaywall и др.
  2.5B связей цитирования



Методики оценивания моделей (benchmarks)

SciDocs: 6 задач

- классификация статей по MeSH / по тематике
- предсказание цитирования / со-цитирования
- предсказание пользовательской активности, рекомендации статей

SciRepEval: 24 задачи, вкл. SciDocs (кроме рекомендаций):

- классификация, регрессия, сходство, поиск,
- подбор рецензента для статьи, разрешение неоднозначности авторов

RuSciBench: 14 задач

- 6 задач классификации OECD/ГРНТИ по аннотации ru / en / ru+en
- 4 задачи кросс-язычного поиска ru→en / en→ru / zh→en / en→zh
- 2 задачи предсказания цитирования / социтирования
- 2 задачи регрессии: предсказание года и цитируемости публикации

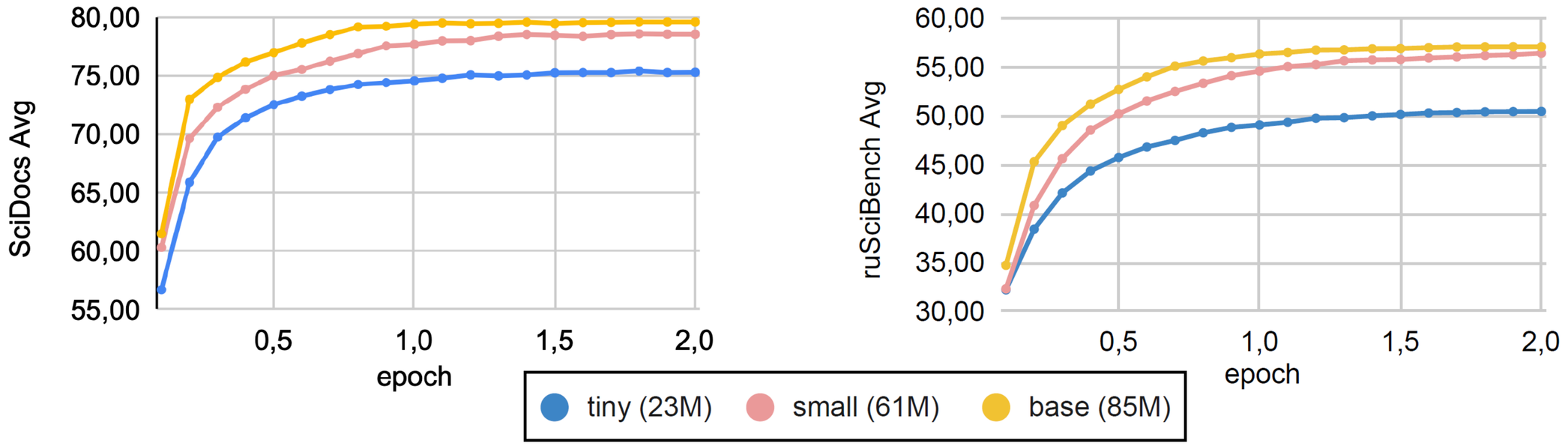
МГУ

N.Gerasimenko, A.Vatolin. RuSciBench benchmark. 2023. https://github.com/mlsa-iai-msu-lab/ru_sci_bench/tree/main

Этап 1: предобучение модели SciRus-tiny

Архитектура RoBERTa (Y.Liu et al., 2019), случайная инициализация:
tiny (sz=23M, dim=312), small (sz=61M, dim=768), base (sz=85M, dim=1024)

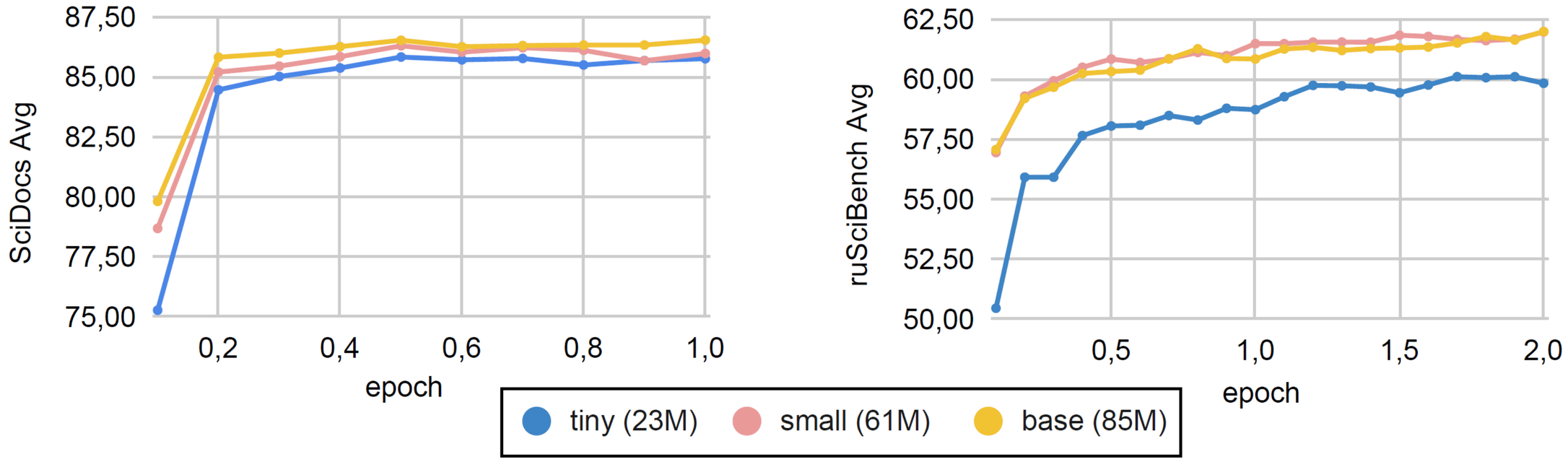
- критерий маскированного языкового моделирования MLM
- две эпохи обучения
- Avg — F1-мера, усреднённая по всем задачам бенчмарка



Этап 2: дообучение на парах title-abstract

Критерий: сближать эмбединги в контрастных парах название/аннотация, ru/en

- 30.6M пар из S2AG
- 17.8M пар из eLibrary



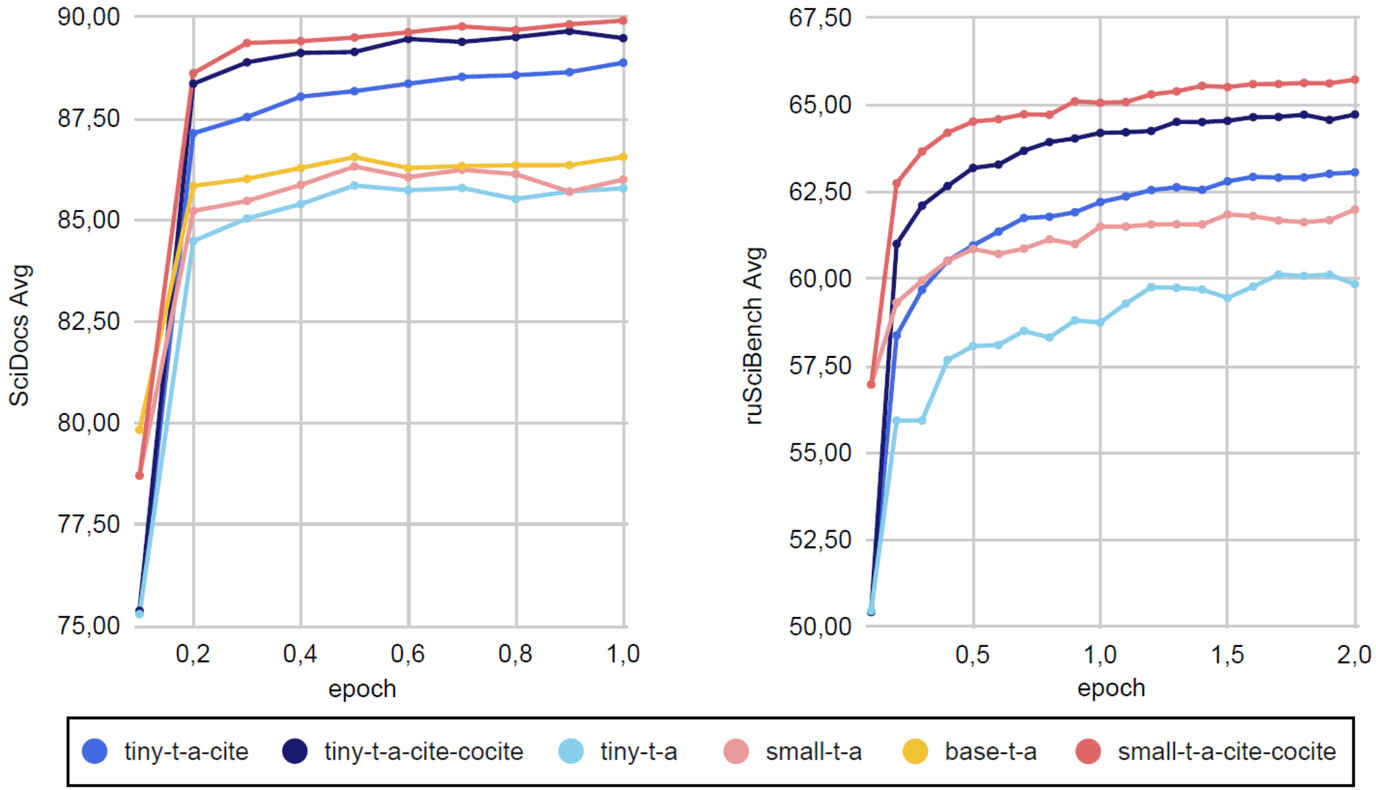
L.Wang et al. Text embeddings by weakly-supervised contrastive pretraining. 2022.

Этап 3: дообучение на парах cite-cocite

Критерий: сближать эмбединги пары документов (A,B) при цитировании:
«cite» — статья A цитирует статью B
«co-cite» — третья статья C цитирует статьи A и B

S2AG:
- 13.3M пар cite
- 62M пар co-cite

eLibrary:
- 40M пар cite
- 33.7M пар co-cite



Сравнение моделей по метрикам ruSciBench

	model_name	Model size	elibrary_oecd_full	translation_search	
			macro_f1	ru_en recall@1	en_ru recall@1
SOTA	e5-mistral-7b-instruct	7.11B	**67,28**	3,65	18,11
	scirus-tiny3.1	23M	65,40	97,40	98,80
	multilingual-e5-large	560M	63,70	**99,19**	**99,37**
	scirus-tiny2	23M	62,02	96,70	95,11
	multilingual-e5-base	278M	62,00	97,00	98,00
	LaBSE	471M	60,21	98,31	97,20
	LaBSE-en-ru	128M	60,05	98,26	96,93
	paraphrase-multilingual-mpnet-base-v2	118M	60,03	66,33	78,18
	FRED-T5-large	360M	59,80	22,25	0,79
	distiluse-base-multilingual-cased-v1	135M	58,69	92,04	90,83
	paraphrase-multilingual-MiniLM-L12-v2	118M	56,48	72,87	77,49
	mfaq	280M	54,84	86,75	90,11
	scirus-tiny	23M	54,83	88,00	88,00

- Сильнее модели, которая в ~20 раз больше
- Приблизились вплотную к SOTA, которую держит модель в ~300 раз больше

Сравнение моделей по метрикам SciRepEval

Model name	Model size	SciDocs	Out-of-Train	In-Train
all-mpnet-base-v2	110M	**91,03**	50,2	53,12
scincl	110M	90,84	**51,8**	55,6
scirus-tiny3.1	23M	90,1	50,08	**57,2**
SPECTER	110M	89,10	50,6	54,7
e5-large-v2	335M	88,70		
e5-base	109M	88,58		
e5-base-v2	109M	88,43		
multilingual-e5-large	560M	87,53	49,32	55,65
e5-small-v2	33.4M	86,99		
multilingual-e5-base	278M	86,91		
e5-mistral-7b-instruct 4byte	7.11B	86,03		
scirus-tiny2	23M	84,21		
sentence-transformers/LaBSE	471M	80,78		
e5_pretrain_longer_240000_similarity_step_5581	23M	80,51		
cointegrated/rubert-tiny2	29.4M	71,60		
allenai/scibert_scivocab_uncased	110M	69,04		
scirus-tiny	23M	67,92		
nreimers/MiniLM-L6-H384-uncased (e5-small-v2 pretrain)	33.4M	65,68		

SOTA (In-Train)

- Топ-3 в **SciDocs** и **Out-of-Train** (конкуренты в ~5 раз больше), SOTA в **In-Train**

Выводы по результатам сравнения моделей

1. **Размер и качество модели в сравнении с SciNCL**
   — меньше параметров: 23M против 110M
   — меньше размерность эмбедингов: 312 против 768
   — больше контекст: 1024 против 512
   — сопоставимое качество (SciDocs Avg): 90.10 против 91.03
2. **Контрастивное дообучение на парах title-abstract**
   — улучшает все метрики, особенно кросс-языковой поиск
3. **Контрастивное дообучение на парах cite / cocite**
   — компенсирует недостаточность кросс-языковых данных
4. **Open Source**
   — бенчмарк интегрирован в MTEB (Multilingual Text Embedding Benchmark)

Данные на Huggingface

Герасименко Н.А., Ватолин А.С., Янина А.О., Воронцов К.В. SciRus: легкий и мощный мультиязычный энкодер для научных текстов. Доклады РАН, 2024
Ватолин А.С., Герасименко Н.А., Янина А.О., Воронцов К.В. RuSciBench: открытый бенчмарк для оценки семантических векторных представлений научных текстов на русском и английском языках. Доклады РАН, 2024.
K.Enevoldsen,..., A. Vatolin et.al. MMTEB: Massive Multilingual Text Embedding Benchmark. 2025.

Первое внедрение (2024)

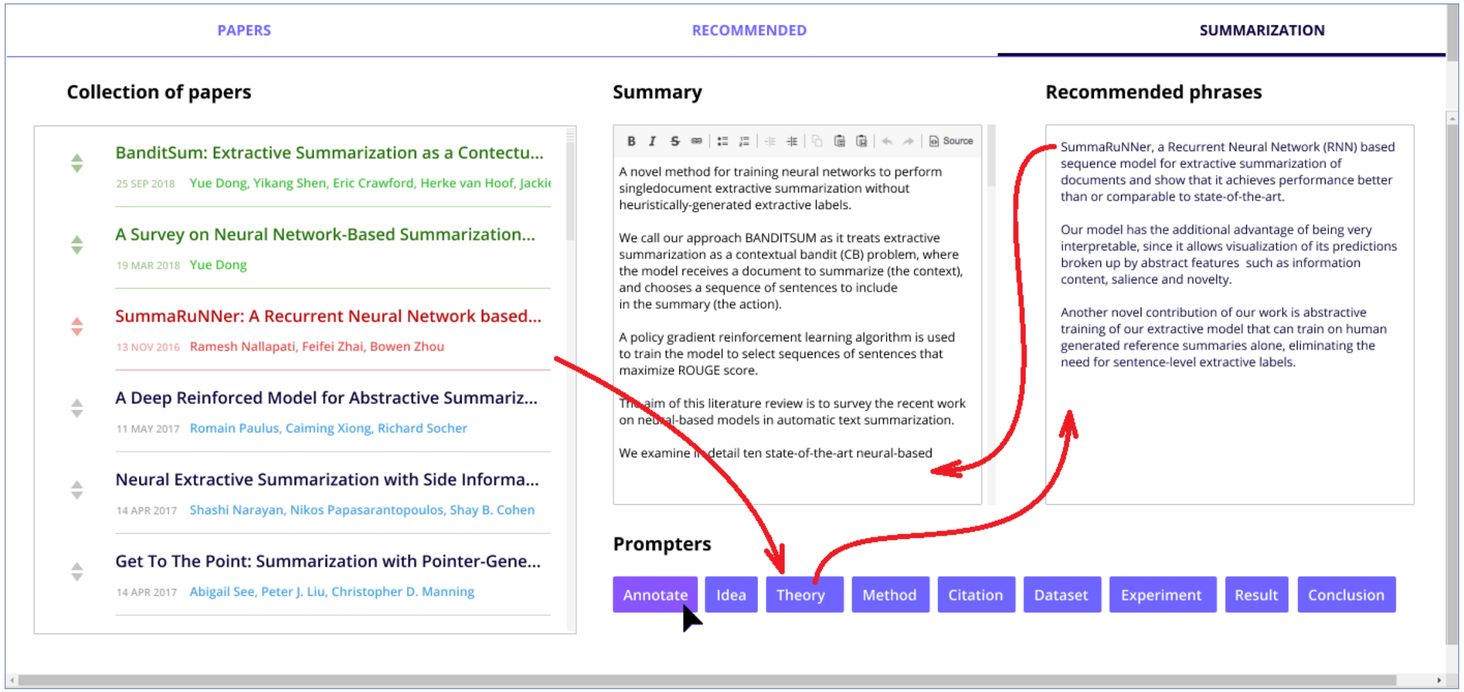
«Разработанная в рамках данного проекта модель уже широко используется в **Научной электронной библиотеке** для решения целого ряда задач, связанных с оценкой тематической близости научных документов. Уже протестирован специалистами полезный сервис для ученых, позволяющий *для заданной статьи или подборки статей найти тематически похожие документы*, как среди всего массива eLIBRARY.RU (более 55 млн. научных публикаций), так и только среди новых поступлений. Важной для нас особенностью данной модели является ее мультиязычность, поскольку **Научная электронная библиотека** содержит документы на различных языках.»

— Геннадий Еременко, генеральный директор НЭБ

Научная электронная библиотека, портал eLIBRARY.RU. Пресс-релиз 24-04-2024: «Открыт поиск близких по тематике публикаций с применением нейросети МГУ для анализа научных текстов.»
https://elibrary.ru/projects/news/search_similar_publ.asp

Сервис полуавтоматического реферирования

Цель: автоматизировать написание реферата по подборке, попутно помогая пользователю систематизировать свои знания в режиме «non-linear reading»



Суфлёры-экстракторы про:
- проблему, идею,
- теорию, метод, модель,
- эксперимент, датасет,
- результаты, выводы,
- достоинства, недостатки,
…

А. Власов. Методы полуавтоматической суммаризации подборок научных статей. МФТИ, 2020

С. Крыжановская. Технология полуавтоматической суммаризации подборок научных статей. МГУ, 2022

Сервис тематизации

Цель: «разложить по полочкам» подборку — выявить тематическую кластерную структуру, дать представление о каждой теме

Тематическая модель определяет

для каждой темы:

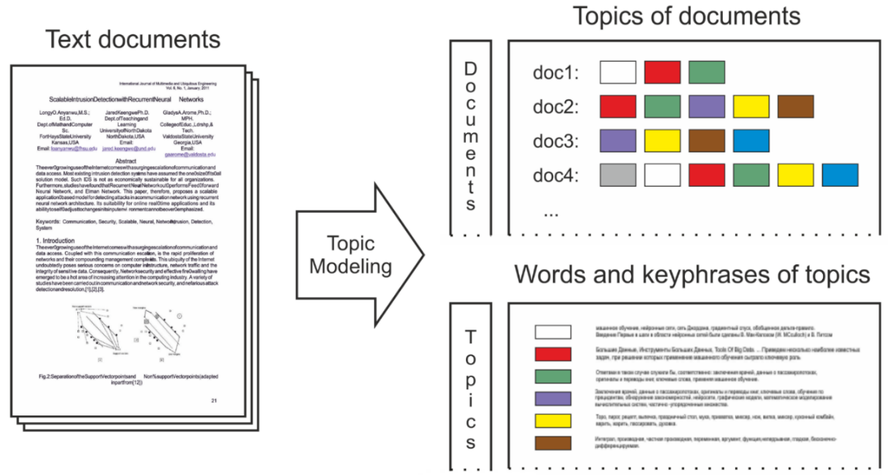
- слова, термины, фразы,
- название, суммаризацию

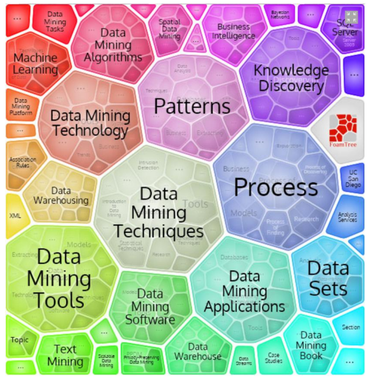
для каждого документа:

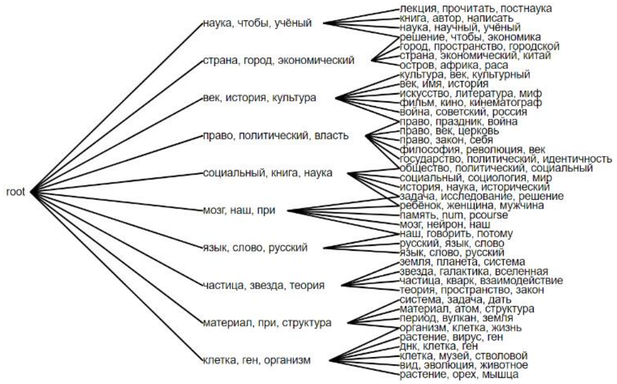
- состав тем и сегментацию по темам

для подборки в целом:

- визуализацию кластерной структуры
- тематический спектр, иерархию







Воронцов К.В. Вероятностное тематическое моделирование: Теория регуляризации ARTM и библиотека с открытым кодом BigARTM. — Москва: URSS. — 2025. — 224 с..
http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

Сервис хронологизации

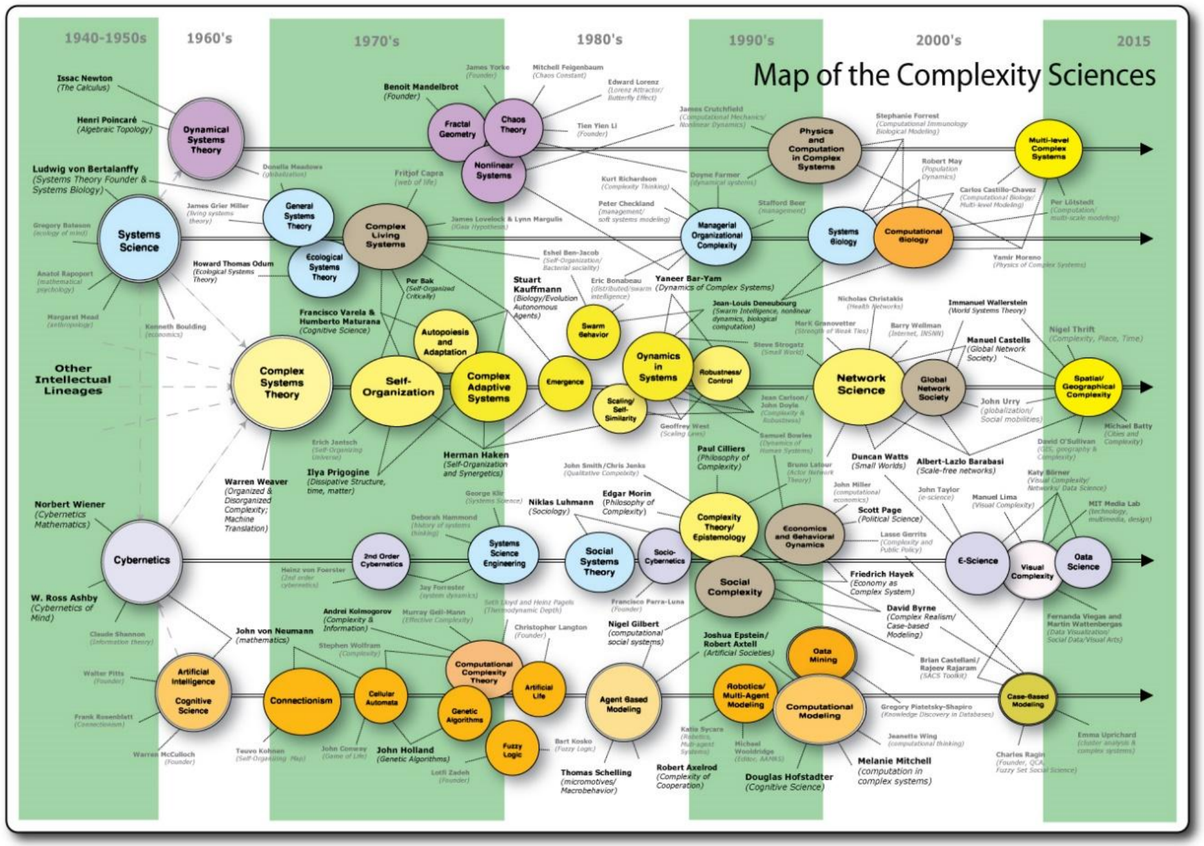
Цель: показать развитие во времени основных долгосрочных тем подборки, обозначив ключевые вехи, идеи и их авторов

Трёхуровневая тематическая иерархия:

- научные направления
- научные теории
- научные школы и учёные

Оси на карте:

- **время × спектр тем**
- читабельность,
- релевантность и др.



http://www.theoryculturesociety.org/brian-castellani-on-the-complexity-sciences

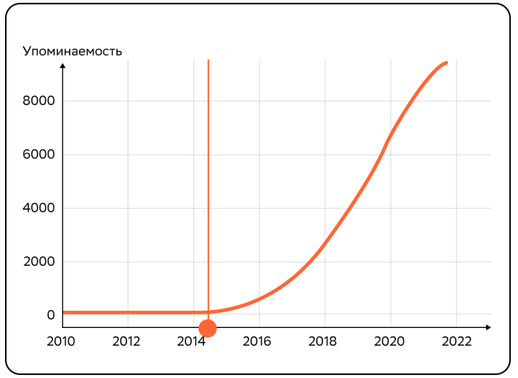
Сервис поиска и анализа трендов

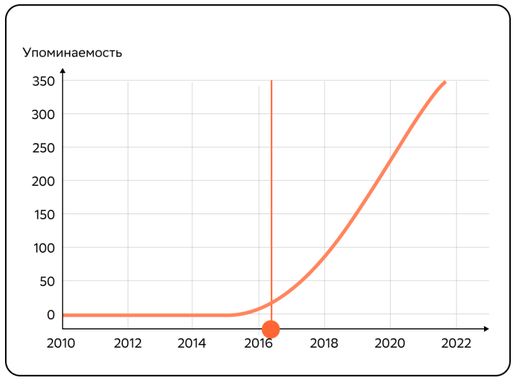
Цель: автоматизировать выявление в подборке новых научных тем, момента их появления, терминологии, интервала роста

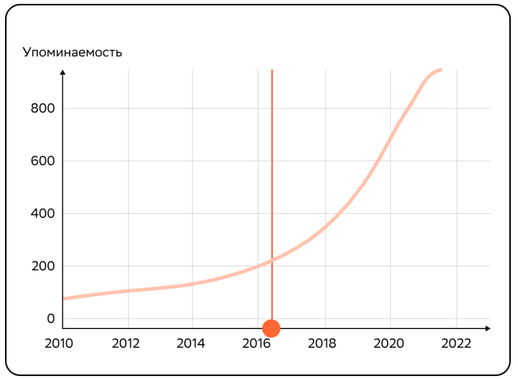
Темпоральная тематическая модель
дообучается без учителя (без размеченных данных) последовательно на месячных интервалах

Для валидации модели экспертами отобраны 87 трендовых тем из области Data Science

Результат: >60% тем детектируются в течение года после появления темы



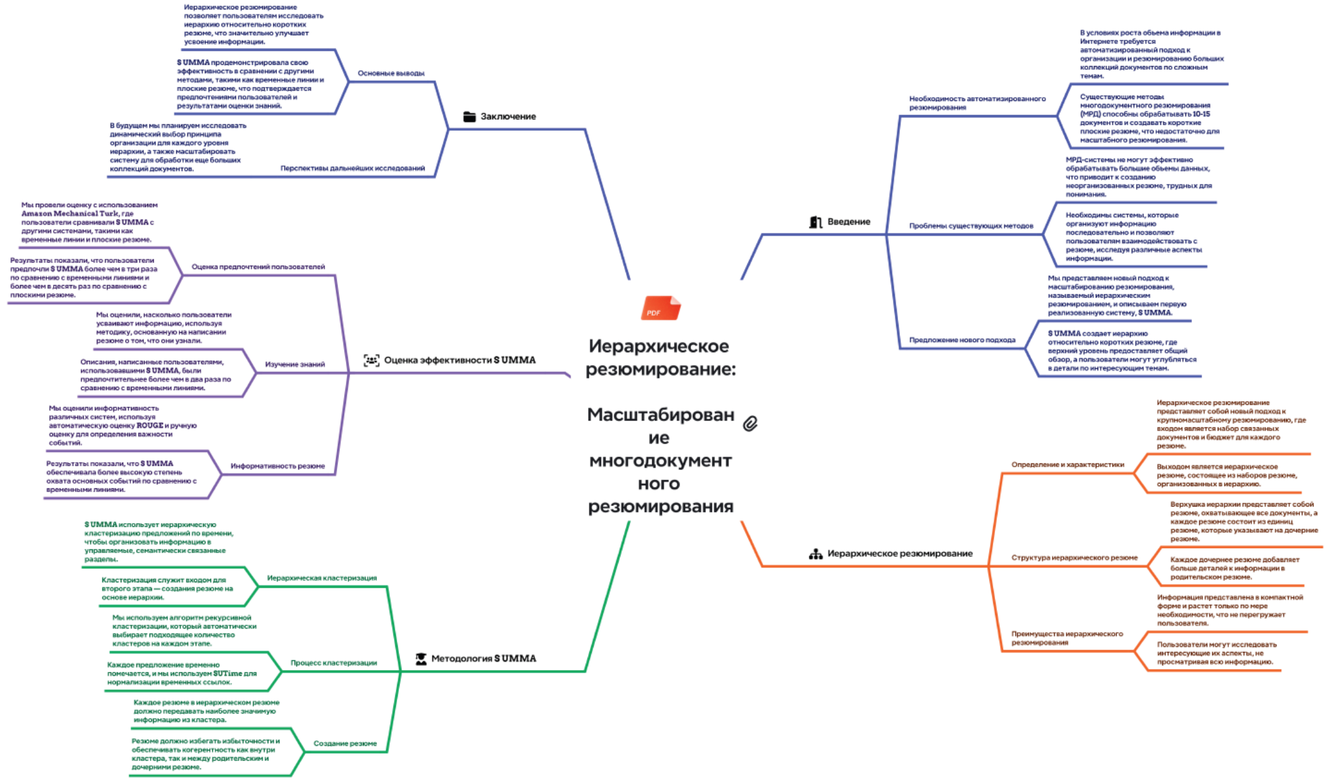




Герасименко Н. А., Чернявский А. С., Никифорова М. А., Никитин М. Д., Воронцов К. В. Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях // Доклады РАН. 2022.

Сервис картирования

Цель: автоматизировать структурирование знаний в виде mind-map, путём выделения главных идей и разделения их на главные подыдеи иерархически



https://mapify.so/ AI Mind Map Summarizer — один из лучших сервисов майндмэпинга на основе LLM

От интеллект-карт (mind-maps) к картам знаний

Интеллект-карты (mind maps)

текстографическое отображение
того, как темы (мысли, идеи)
разбиваются на подтемы иерархически

максимально близкое к тому, как
мы храним знания у себя в головах

предложены в 70-е годы
британским **психологом**
Тони Бьюзеном

От интеллект-карт (mind-maps) к картам знаний

нацелены на повышение эффективности

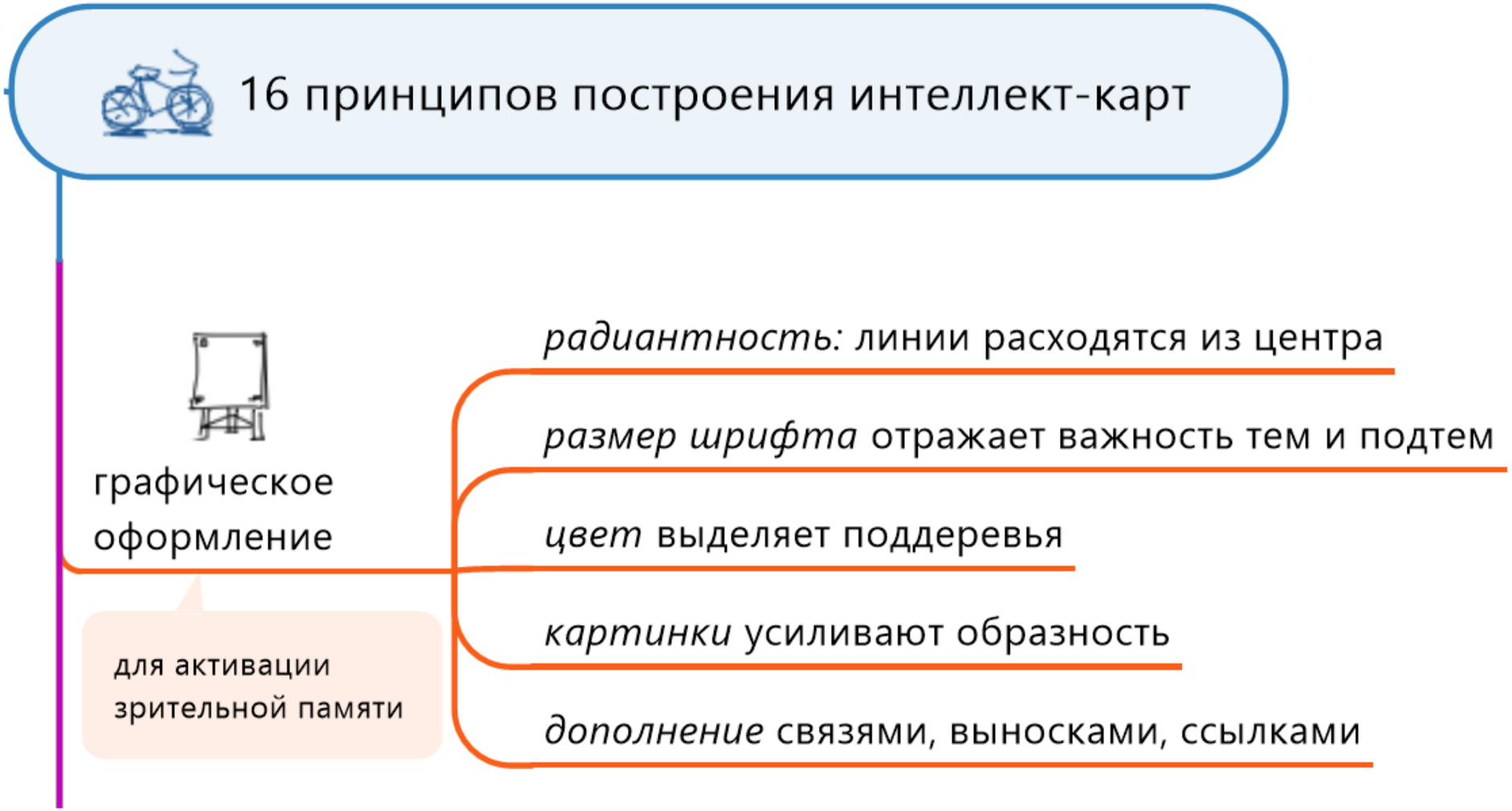
- конспектирования
- понимания
- запоминания
- систематизации
- поиска консенсуса

благодаря активизации обоих полушарий мозга, учёта особенностей восприятия, мышления, памяти

техника запоминания

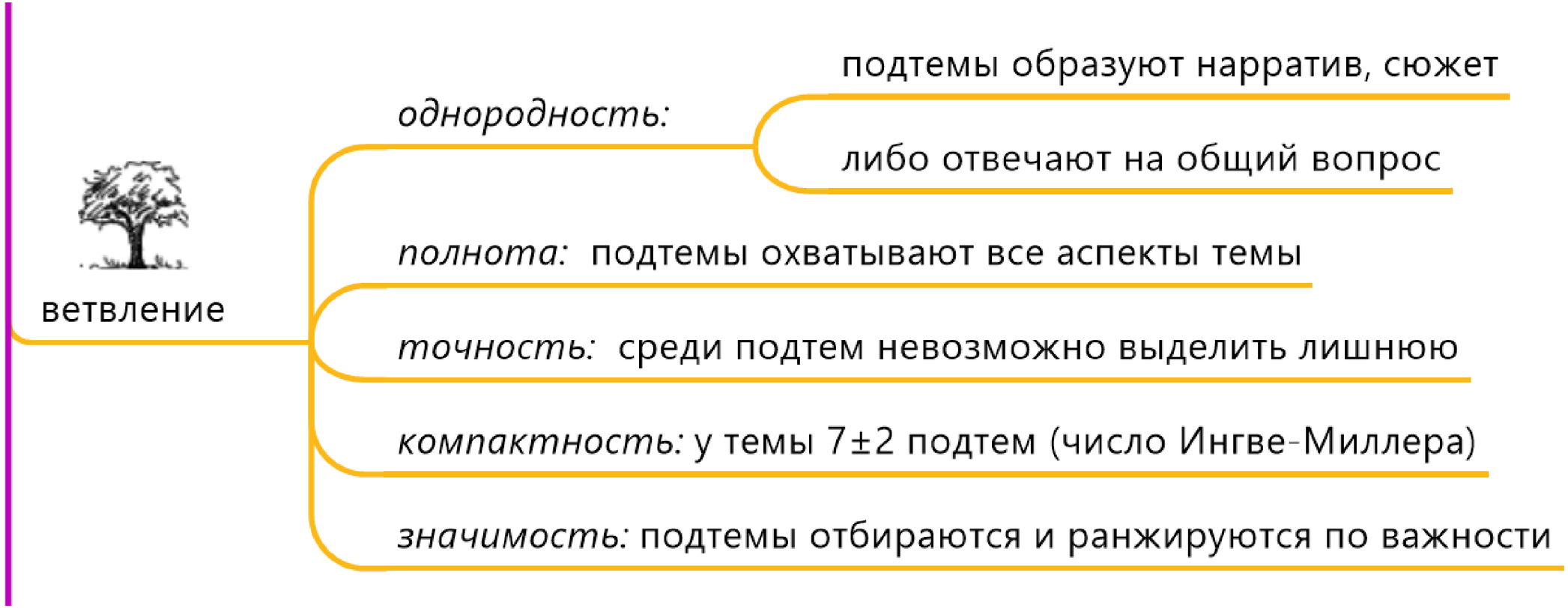
- посмотреть, понять, обсудить, договориться, принять
- самостоятельно воспроизвести через 10 минут → сутки → неделю → месяц

От интеллект-карт (mind-maps) к картам знаний



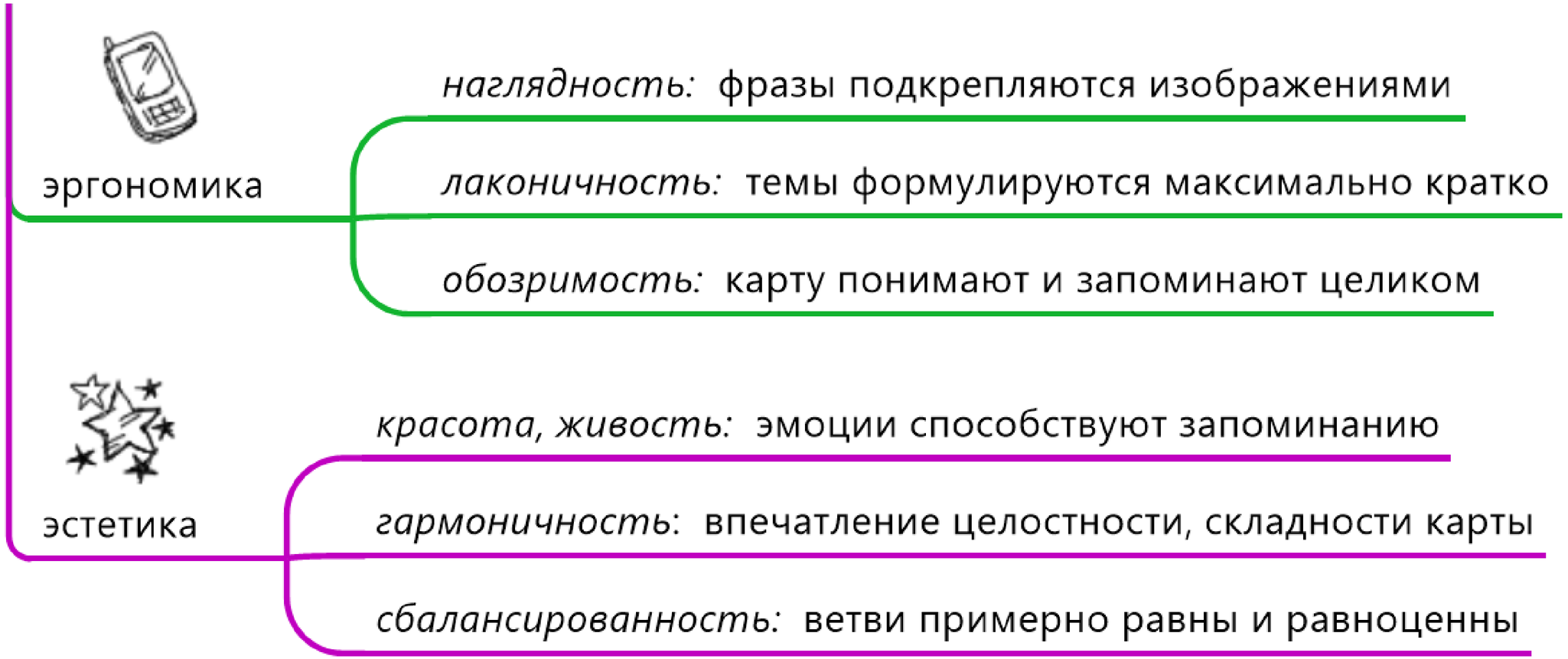
От интеллект-карт (mind-maps) к картам знаний

(16 принципов построения интеллект-карт)



От интеллект-карт (mind-maps) к картам знаний

(16 принципов построения интеллект-карт)



От интеллект-карт (mind-maps) к картам знаний

6 принципов, усиливающих интеллект-карты до **карт знаний**

(1)
читабельность

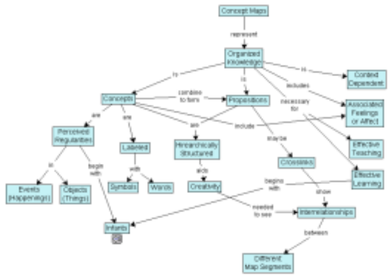
компромисс с
лаконичностью
и обозримостью

любой фрагмент карты
читается как нарратив

легко и однозначно

даже автоматически

в отличие
от других
способов
представления
знаний

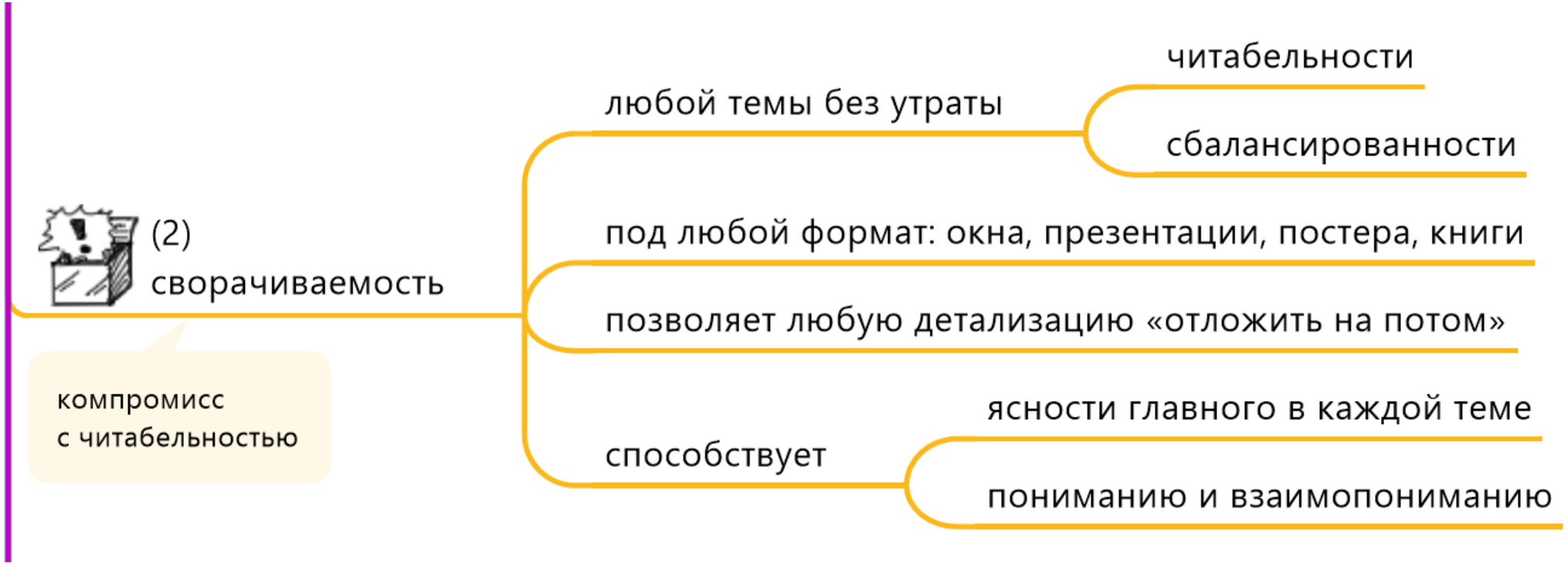


онтологий

фреймов и др.

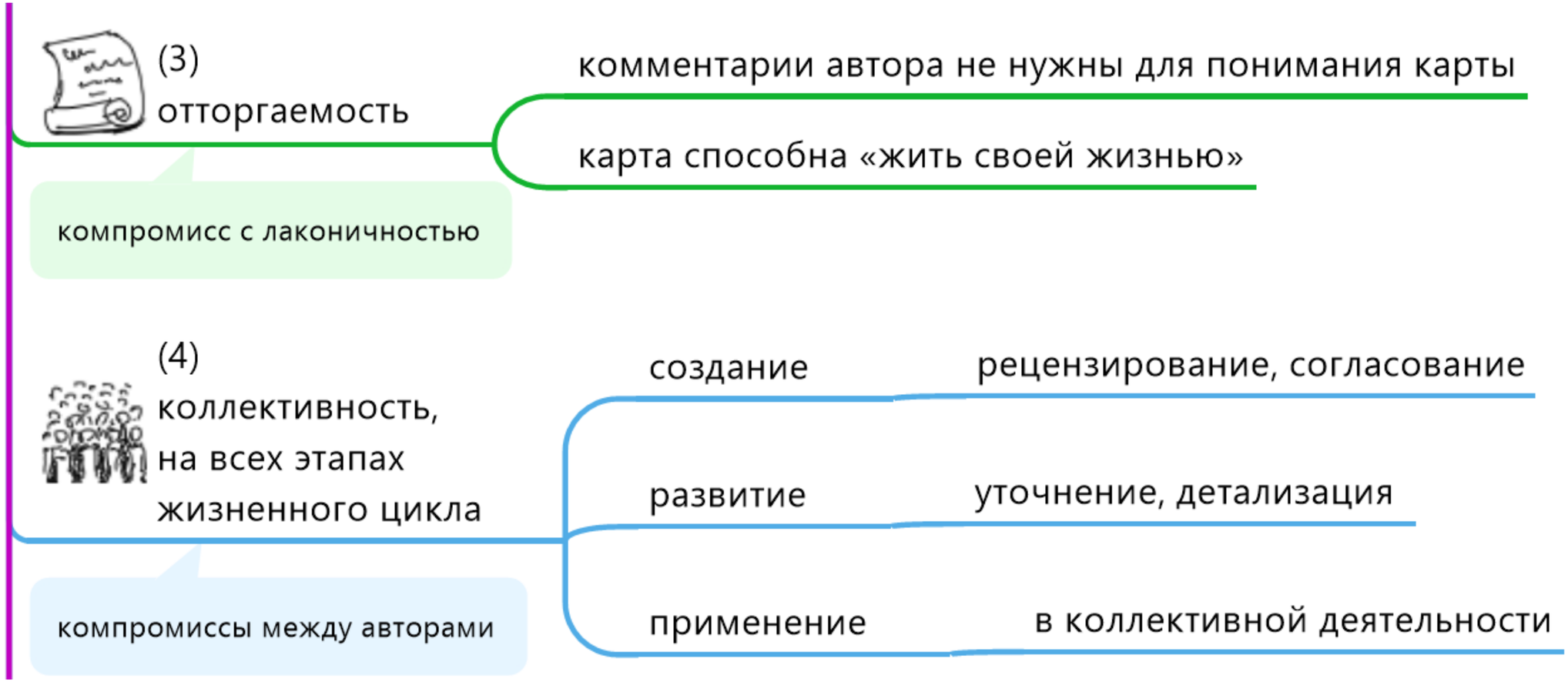
От интеллект-карт (mind-maps) к картам знаний

(6 принципов, усиливающих интеллект-карты до **карт знаний**)



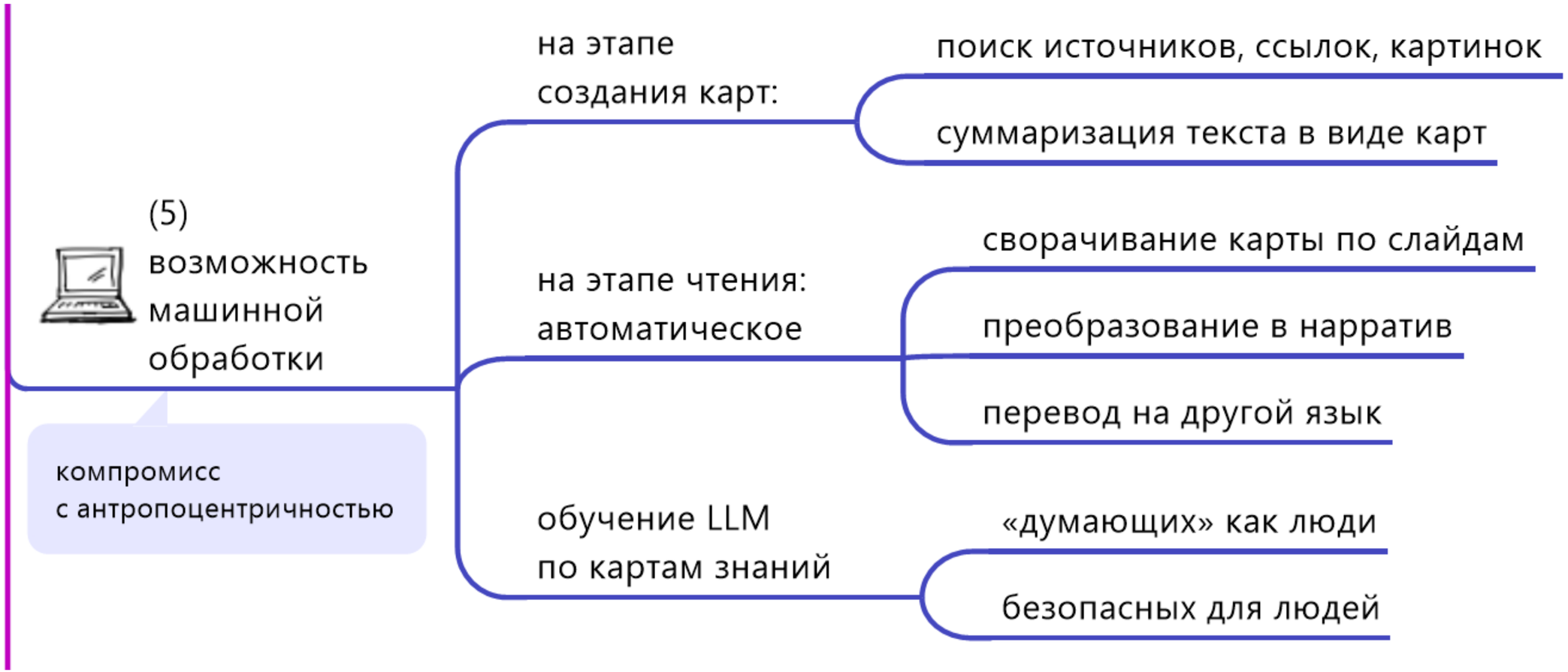
От интеллект-карт (mind-maps) к картам знаний

(6 принципов, усиливающих интеллект-карты до **карт знаний**)



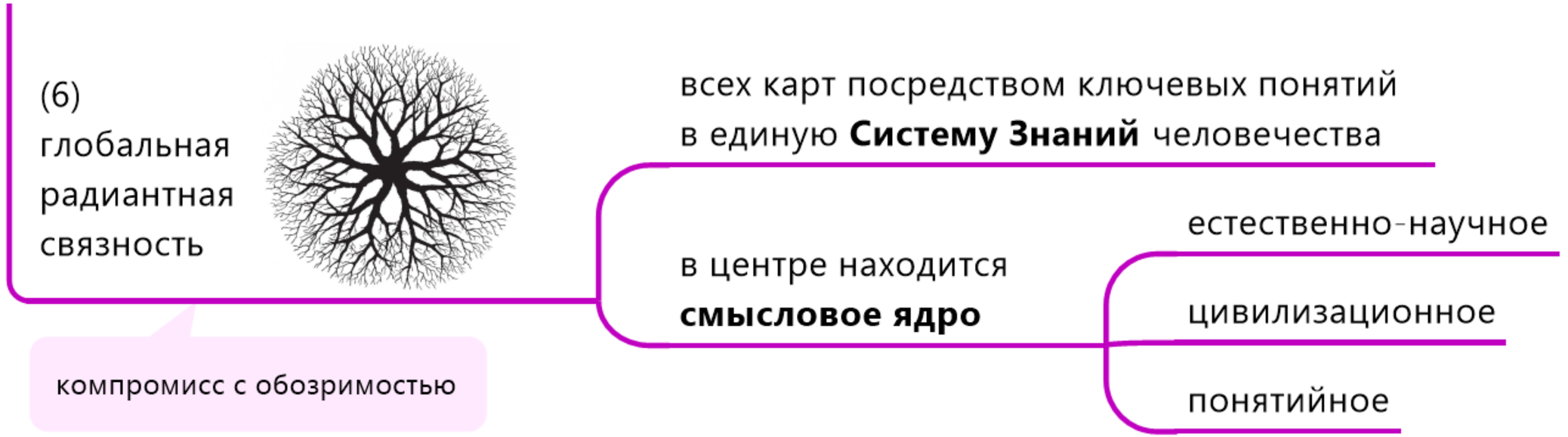
От интеллект-карт (mind-maps) к картам знаний

(6 принципов, усиливающих интеллект-карты до **карт знаний**)



От интеллект-карт (mind-maps) к картам знаний

(6 принципов, усиливающих интеллект-карты до **карт знаний**)



Воронцов К.В., Курилов В.А. Карты знаний: на пути к доверенным языковым моделям и системам представления знаний. BIS Journal, №3 (54), 2024.

Карты знаний. Выводы

- **Представление знаний**, универсальное для человека и машины
  — 16+6 принципов реализуются через промпты для LLM
- **Перспективный инструмент** «коллективного разума»,
  развивающий навыки работы с научной информацией:
  — во всём выделять главное (7±2),
  — делать это быстро, формулировать лаконично,
  — достигая в команде единства понимания целей, идей, смыслов
- **Обучение ИИ** по тексто-графическим представлениям потребует:
  — освоить картирование знаний (индивидуально и в коллективе)
  — накапливать обучающие выборки и бенчмарки

Sobolevsky F.A., Vorontsov K.V. Text Tree Edit Distance: A Language Model-Based Metric for Text Hierarchies. 2025

Как активировать визуальное аналитическое мышление (эволюционно обусловленное, намного более мощное)

1. порядка сотни карт: просмотреть, обсудить, поспорить, принять
2. десятки карт: построить самому, следуя 16+6 принципам
3. испытать «моменты ясности», инсайты, когда карта
   - «красиво сложилась»
   - привела к согласию
   - легко и ярко запомнилась,
   - легла в основу деятельности

индивидуальная практика и опыт

4. сделать построение карт регулярной профессиональной практикой
   - индивидуальной
   - коллективной

Мастерская знаний. Выводы

Мастерская знаний — не проект, а адаптивная концепция:
реализаций может быть много и разных

Миссия: устранять барьеры между человеком и знанием,
активировать «коллективный разум» на всём жизненном цикле научного проекта

Реализация: в основном экстрактивные методы
(LLM-энкодеры, векторный поиск, ранжирование, тематические модели),
генеративные LLM — по принципу «минимальной достаточности»

Конец технократии: придётся проектировать информационные системы
не просто как инженерно-технические (сделал потому, что мог),
а как социально-технические (во благо человеческой цивилизации)

Антропоцентричное определение ИИ

Искусственный интеллект —

вычислительные технологии, создаваемые для повышения производительности созидательного интеллектуального труда людей

не замена человека

не «загадочный новый тип разума»

не повод уподобиться Богу, чтобы «творить по образу и подобию Своему»

Спасибо за внимание!

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

зав. кафедрой ММП ВМК МГУ,
зав. лаб. МОСА Института ИИ МГУ,
зав. кафедрой ИС и кафедрой МОЦГ МФТИ,
г.н.с. ФИЦ «Информатика и управление» РАН

k.vorontsov@iai.msu.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov

Научный семинар ИПУ РАН
«Проблемы управления знаниями»