



Институт перспективных исследований проблем
искусственного интеллекта и интеллектуальных систем
МГУ им. М.В. Ломоносова

Проект «Мастерская знаний»

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН, зав. лабораторией МОСА ИИИ МГУ

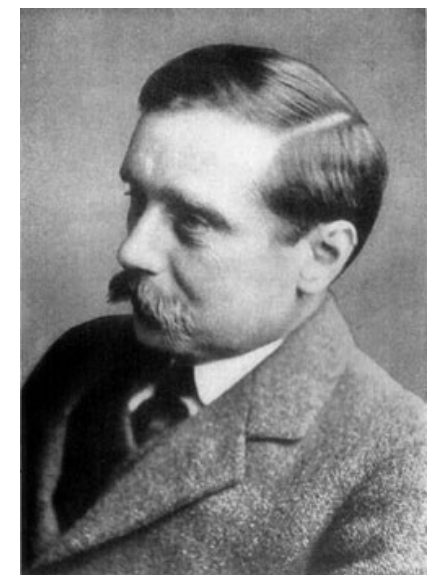
Научный семинар по анализу данных
Московского городского педагогического университета

10 апреля 2024

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**, 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 where knowledge and ideas are received, sorted, summarized, digested, clarified and compared** – *Herbert Wells, 1940*)



Сегодня технологии IR/ML/NLP/NLU позволяют решать такие задачи

Эволюция подходов в обработке естественного языка

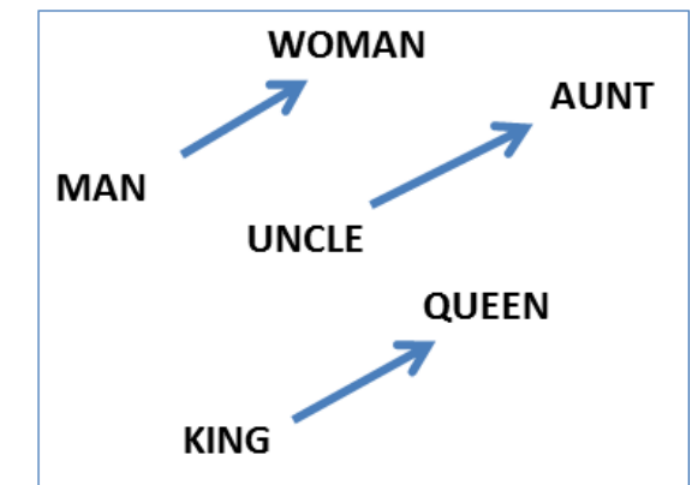
Декомпозиция задач по уровням «пирамиды NLP»

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



Модели векторных представлений слов (эмбедингов)

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



Нейросетевые векторные модели локальных контекстов

- рекуррентные нейронные сети: LSTM, GRU,...
- «end-to-end» модели внимания и трансформеры: машинный перевод, BERT [2018], GPT-3 [2020], GPT-4 [2023]

$$\text{softmax} \left(\frac{\begin{matrix} \mathbf{Q} & \mathbf{K}^T \\ \begin{matrix} \square & \square \\ \square & \square \end{matrix} & \begin{matrix} \square & \square \\ \square & \square \end{matrix} \end{matrix}}{\sqrt{d}} \right) \mathbf{V}$$

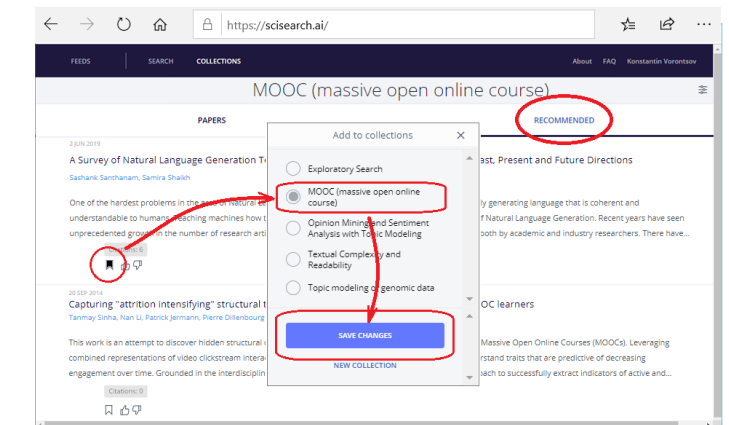
The diagram shows a matrix multiplication of a query matrix Q (purple) and a key matrix K^T (orange), followed by a softmax function and a value matrix V (blue). The result is a vector of size d.

Функции «Мастерской знаний»

Подборка текстов – поисковый интерес и рабочее пространство пользователя или группы

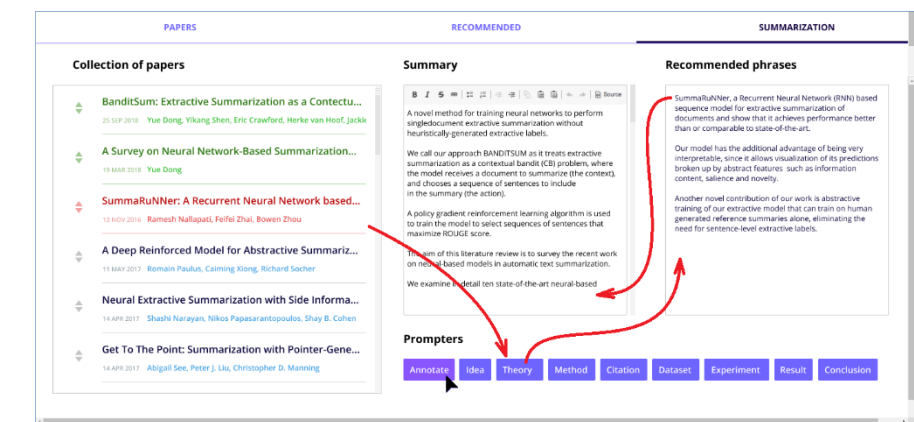
Поисково-рекомендательные сервисы:

- поиск тематически близких документов по **подборке**
- мониторинг новых документов по тематике **подборки**
- выявление новых научных трендов по тематике **подборки**
- контекстные рекомендации в тексте документа из **подборки**



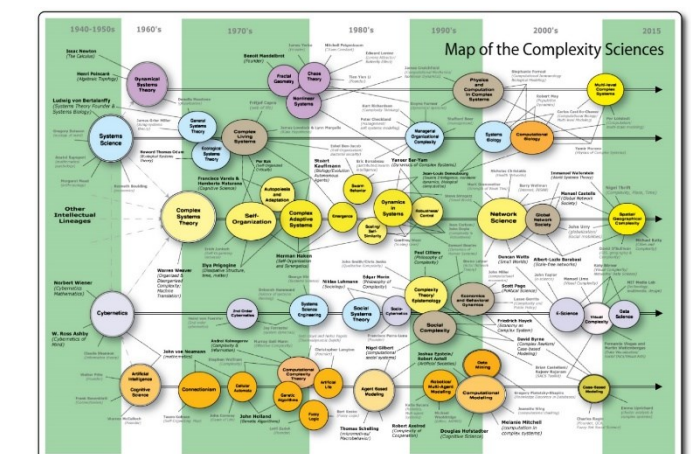
Аналитические сервисы:

- полуавтоматическая суммаризация **подборки**
- систематизация и картирование идей (mindmap) по **подборке**
- рекомендация порядка чтения документов **подборки**



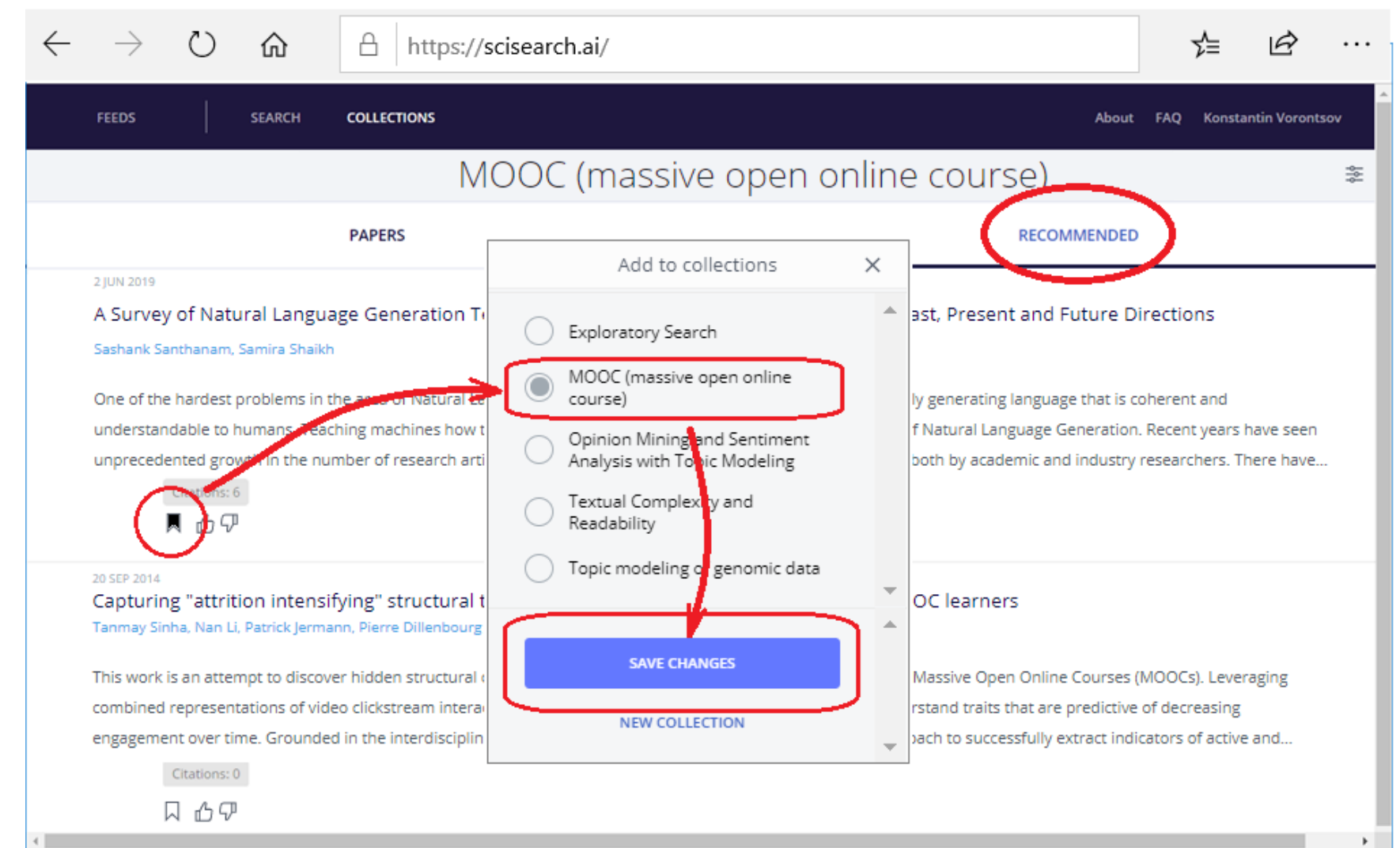
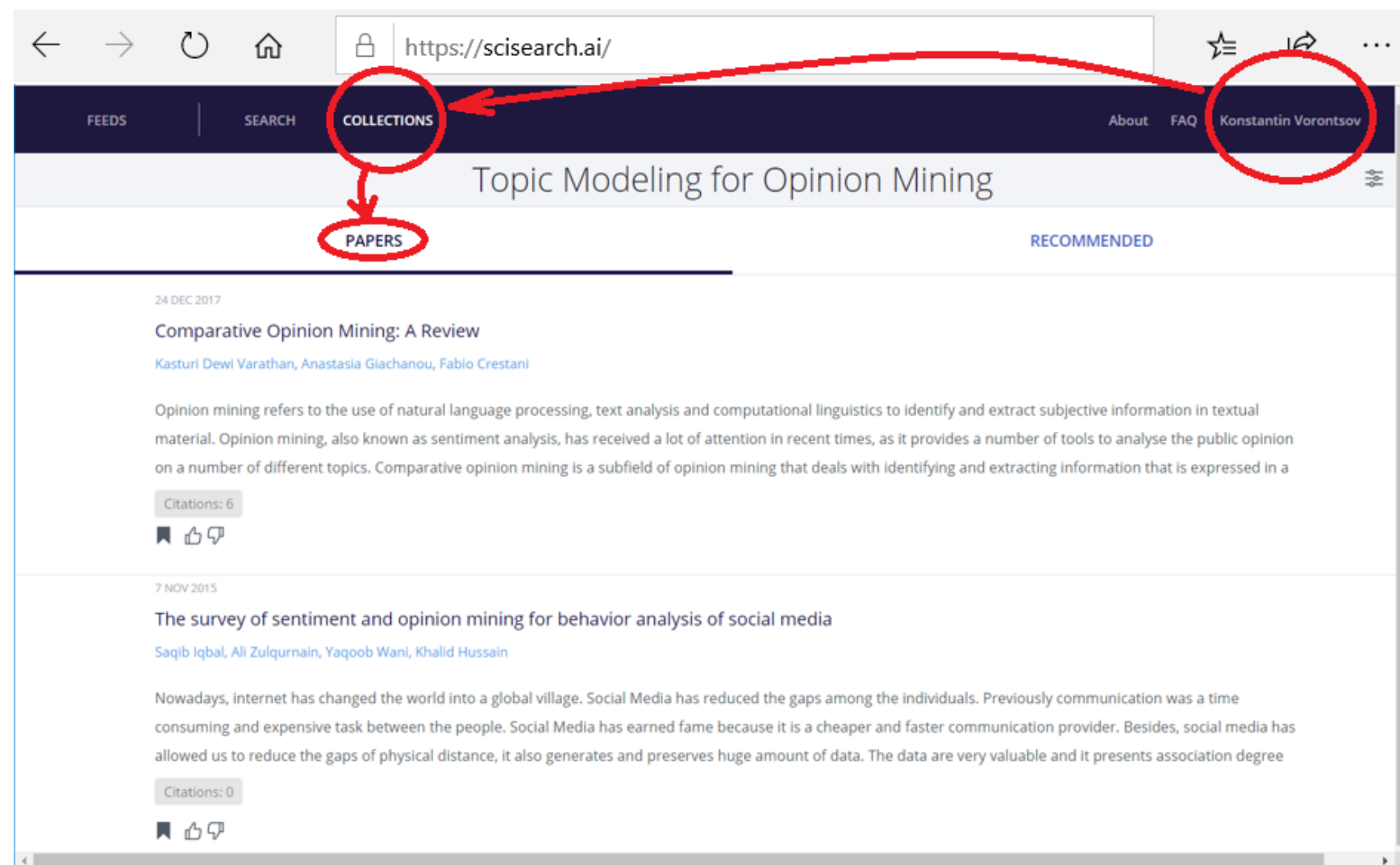
Коммуникативные сервисы:

- совместное составление, обсуждение, использование **подборок**
- интерактивная визуализация и инфографика по **подборке**

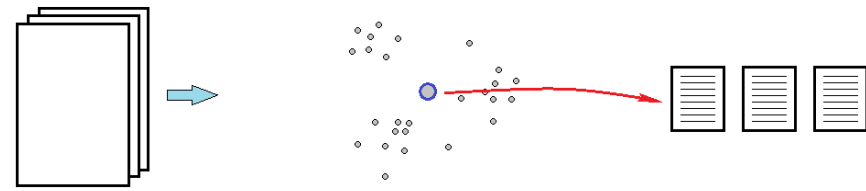


Поиск и рекомендации (прототип SciSearch.ru)

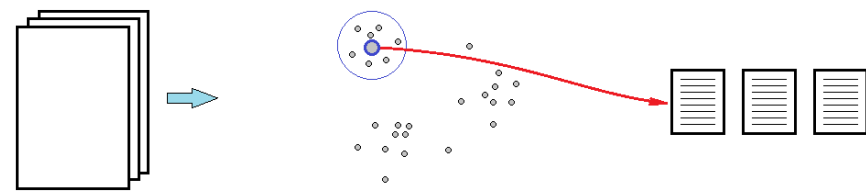
Подборка играет роль поискового запроса и поисковой выдачи одновременно



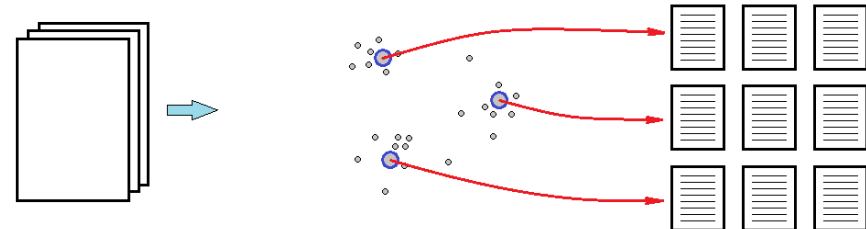
Стратегии векторного документного поиска



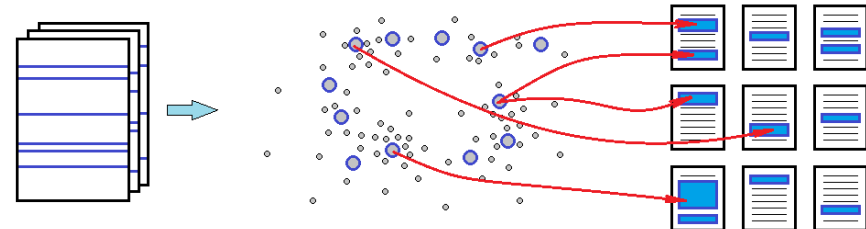
1. Поиск по среднему вектору **подборки** (неудачная стратегия):



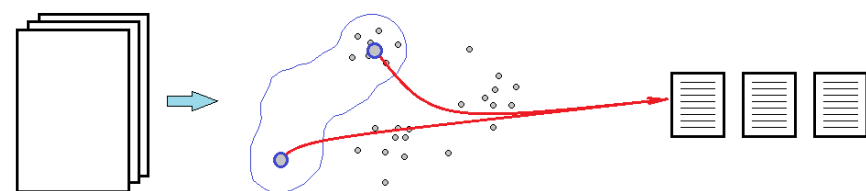
2. Поиск по части **подборки**, близкой к выбранному документу:



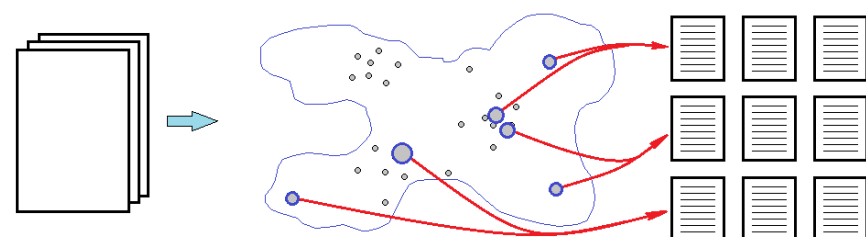
3. Поиск по тематике кластеров, на которые делится **подборка**:



4. Поиск по тематике сегментов документов **подборки**:



5. Поиск по тематике, смежной для части **подборки**:



6. Поиск по тематике, смежной для всей **подборки**:

SciRus: LLM для научных текстов. Как учим?

Датасеты

eLIBRARY.RU

~14M аннотаций (3B токенов)
8.5M ru, 5.2M en



SEMANTIC SCHOLAR

~30M аннотаций (12B токенов)
85% en, 2% ru, ...



- Учим только на научных данных для оптимизации и повышения точности
- Наибольшая ценность – специфический контрастив на ссылках/терминах/etc

Как узнать, что получилось хорошо?

Бенчмарки SciDocs и ruSciBench

SciDocs (Semantic Scholar) – 6 задач

SciDocs MAG	SciDocs MeSH	SciDocs Cite	SciDocs CoView	SciDocs CoCite	SciDocs CoRead
классификация	классификация	поиск (ранжирование)	поиск (ранжирование)	поиск (ранжирование)	поиск (ранжирование)

ruSciBench (IAI MSU) – 7 задач (пока)

elibrary_oecd_full	elibrary_grnti_full	elibrary_oecd_ru	elibrary_grnti_ru	elibrary_oecd_en	elibrary_grnti_en	translation_search
классификация	классификация	классификация	классификация	классификация	классификация	поиск (ранжирование)

- Качество эмбедингов оценивается через качество решения на их базе прикладных задач
- Правильные ответы для всех задач известны, а прикладные методы зафиксированны (например, метод опорных векторов)

Метрики ruSciBench. Путь к state-of-the-art (SOTA)



model_name	Model size	elibrary_oecd_fu ll	translation_search	
		macro_f1	ru_en recall@1	en_ru recall@1
e5-mistral-7b-instruct	7.11B	67,28	3,65	18,11
multilingual-e5-large	560M	63,70	99,19	99,37
scirus-tiny3	23M	61,13	94,83	95,81
scirus-tiny2	23M	60,86	96,7	95,11
multilingual-e5-base	278M	62	97	98
LaBSE	471M	60,21	98,31	97,20
LaBSE-en-ru	128M	60,05	98,26	96,93
paraphrase-multilingual-mpnet-base-v2		60,03	66,33	78,18
FRED-T5-large	360M	59,80	22,25	0,79
distiluse-base-multilingual-cased-v1		58,69	92,04	90,83
paraphrase-multilingual-MiniLM-L12-v2		56,48	72,87	77,49
mfaq		54,84	86,75	90,11
scirus-tiny	23M	54,83	88	88

- Метрики выросли >10%, разрыв сократился в >3 раза

Метрики SciDocs. Путь к state-of-the-art (SOTA)



Model name	Model size	Avg
all-mpnet-base-v2	110M	91,03
scincl	110M	90,84
e5-large-v2	335M	88,70
e5-base	109M	88,58
e5-base-v2	109M	88,43
multilingual-e5-large	560M	87,53
e5-small-v2	33.4M	86,99
multilingual-e5-base	278M	86,91
scirus-tiny3 <small>NEW</small>	23M	86,83
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

- Метрики выросли >20%, разрыв сократился в >3 раза

Полуавтоматическое реферирование подборки

Концепция MANS (Machine Aided Human Summarization)

1. Система рекомендует *сценарий реферата*, то есть в каком порядке процитировать статьи из подборки
2. Пользователь корректирует план в соответствии со своими целями
3. В цикле по ранжированным статьям подборки:
 - пользователь вызывает (кликает кнопку) одного из *суфлёров* по статье: «как другие авторы обычно ссылаются на эту статью», «основная идея статьи», «метод», «достоинство», «недостаток», «результат», «вывод» и т.д.
 - система строит ранжированный список фраз
 - пользователь выбирает фразу из ранжированного списка
 - пользователь корректирует фразу и контекст в соответствии со своими целями

Полуавтоматическое реферирование подборки

The interface is divided into three main sections: **PAPERS**, **RECOMMENDED**, and **SUMMARIZATION**.

Collection of papers (PAPERS section):

- BanditSum: Extractive Summarization as a Contextu... (25 SEP 2018, Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, Jacki)
- A Survey on Neural Network-Based Summarization... (19 MAR 2018, Yue Dong)
- SummaRuNNer: A Recurrent Neural Network based...** (13 NOV 2016, Ramesh Nallapati, Feifei Zhai, Bowen Zhou)
- A Deep Reinforced Model for Abstractive Summariz... (11 MAY 2017, Romain Paulus, Caiming Xiong, Richard Socher)
- Neural Extractive Summarization with Side Informa... (14 APR 2017, Shashi Narayan, Nikos Papasarantopoulos, Shay B. Cohen)
- Get To The Point: Summarization with Pointer-Gener... (14 APR 2017, Abigail See, Peter J. Liu, Christopher D. Manning)

Summary (RECOMMENDED section):

A novel method for training neural networks to perform single-document extractive summarization without heuristically-generated extractive labels.

We call our approach BANDITSUM as it treats extractive summarization as a contextual bandit (CB) problem, where the model receives a document to summarize (the context), and chooses a sequence of sentences to include in the summary (the action).

A policy gradient reinforcement learning algorithm is used to train the model to select sequences of sentences that maximize ROUGE score.

The aim of this literature review is to survey the recent work on neural-based models in automatic text summarization.

We examine in detail ten state-of-the-art neural-based

Recommended phrases (SUMMARIZATION section):

SummaRuNNer, a Recurrent Neural Network (RNN) based sequence model for extractive summarization of documents and show that it achieves performance better than or comparable to state-of-the-art.

Our model has the additional advantage of being very interpretable, since it allows visualization of its predictions broken up by abstract features such as information content, salience and novelty.

Another novel contribution of our work is abstractive training of our extractive model that can train on human generated reference summaries alone, eliminating the need for sentence-level extractive labels.

Promoters

Annotate Idea Theory Method Citation Dataset Experiment Result Conclusion

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

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

Полуавтоматическое реферирование подборки

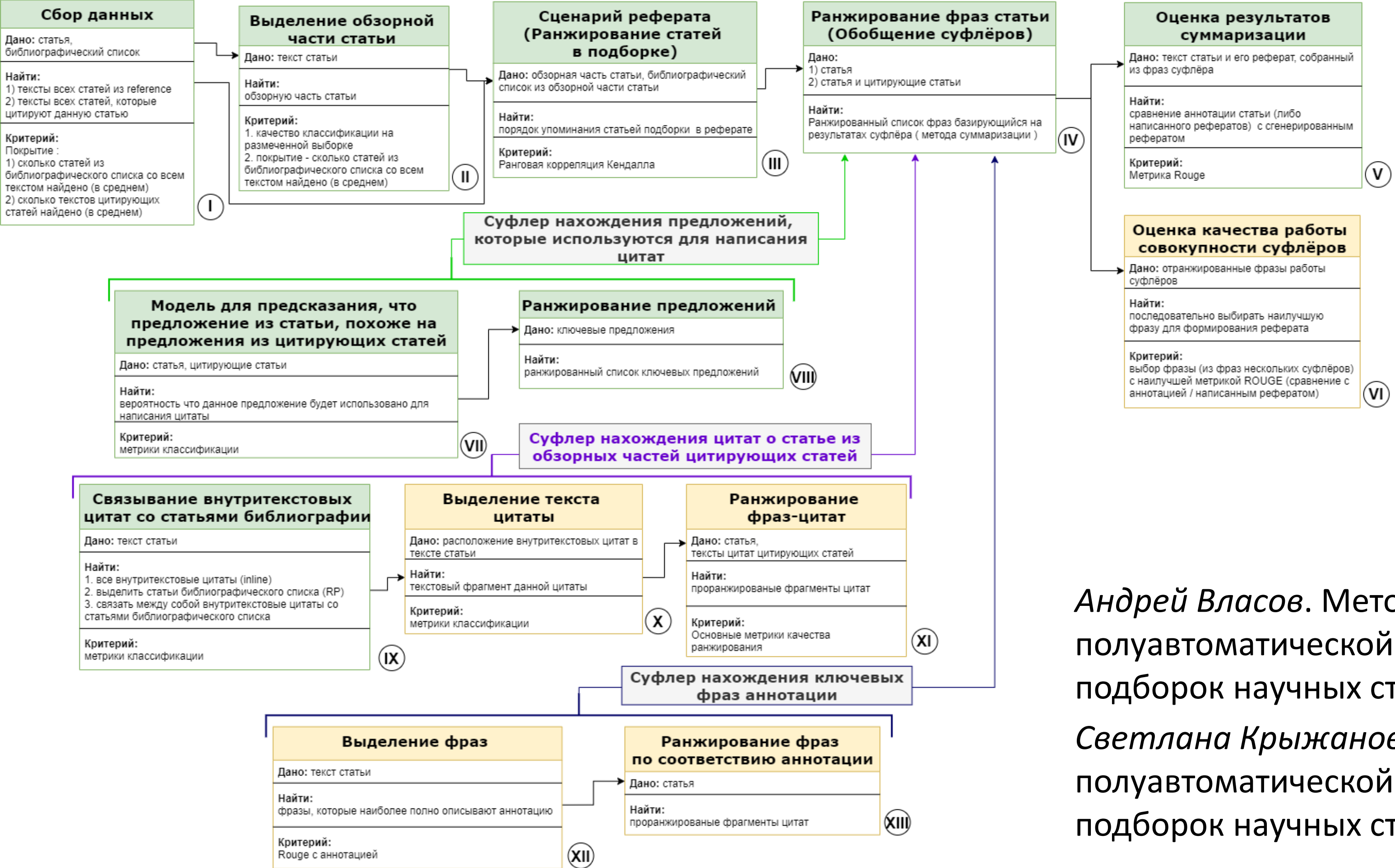
Основные задачи машинного обучения:

- Формирование обучающей выборки: **paper** → (**refs, survey**)
- Ранжирование статей для сценария реферата
- Выбор релевантных фраз из текста статьи для каждого суфлёра
- Ранжирование выбранных фраз для каждого суфлёра
- Выбор релевантного контекста по данной ссылке, например:

Few contextual citation graphs are publicly available. The ACL Anthology Network (AAN) (Radev et al., 2009) is one such contextual citation graph built from the ACL Anthology corpus (Bird et al., 2008), consisting of 24.6K papers manually augmented with citation information. CiteSeer (Giles et al., 1998) provides a large corpus consisting of 1.0M papers with full text and bibliography entries parsed from PDFs. Saier and Farber (2019) introduces a contextual citation graph of approximately 1.0M arXiv papers with full text LaTeX parses where citations are linked to papers in the Microsoft Academic Graph.

M.Yasunaga, J.Kasai, R.Zhang, A.Fabbri, I.Li, D.Friedman, D.Radev. ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. 2019.

Полуавтоматическое реферирование подборки



Андрей Власов. Методы полуавтоматической суммаризации подборок научных статей. МФТИ, 2020
 Светлана Крыжановская. Технология полуавтоматической суммаризации подборок научных статей. МГУ, 2022

Технология тематического моделирования BigARTM

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Онлайн-параллельный мультимодальный ARTM
- Встроенная библиотека регуляризаторов и мер качества

Сообщество:

- Открытый код <https://github.com/bigartm>
(discussion group, issue tracker, pull requests)
- Документация <http://bigartm.org>



Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

3.7М статей Википедии, 100К слов: время min (перплексия)

проц.	T	Gensim	Vowpal Wabbit	BigARTM	BigARTM асинхрон
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

Ianina A., Golitsyn L., Vorontsov K. [Multi-objective topic modeling for exploratory search in tech news](#). AINL 2017.

Vorontsov K. Rethinking Probabilistic Topic Modeling from the Point of View of Classical Non-Bayesian Regularization. 2023.

<http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>

Технология тематического поиска

Схема эксперимента:

- длинные запросы (1 стр. А4)
- 100 запросов на коллекцию
- 3 ассессора на каждый запрос
- от 10 до 60 минут на запрос
- разметка на Яндекс.Толока
- две коллекции техно-новостей:



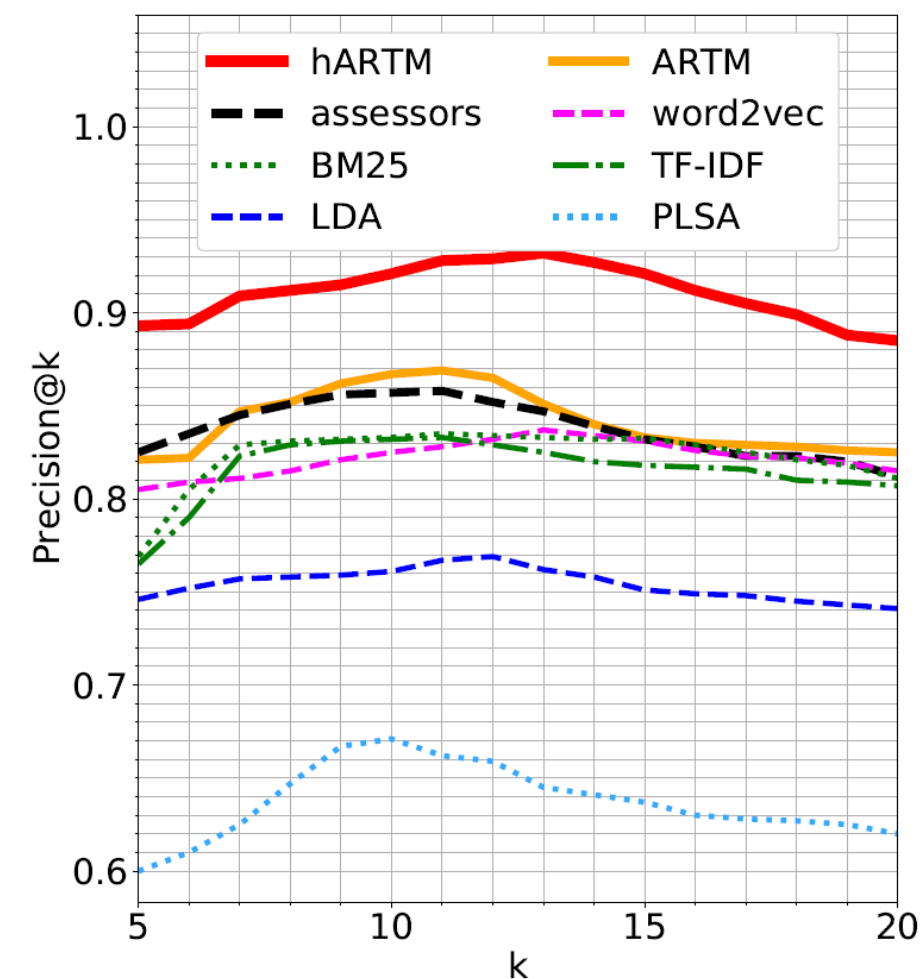
(170K Russian docs)



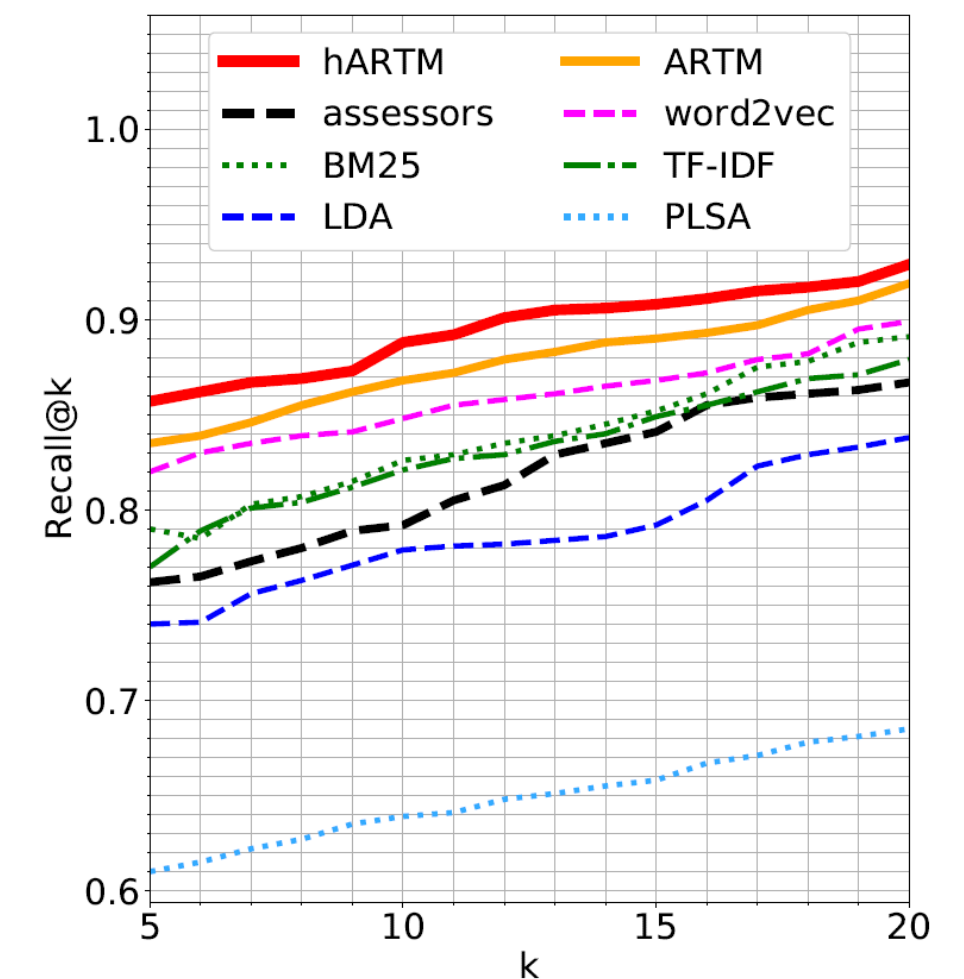
(750K English docs)

Оценки качества поиска:

точность (precision@k)



полнота (recall@k)



Ianina A., Golitsyn L., Vorontsov K. [Multi-objective topic modeling for exploratory search in tech news](#). AINL 2017.

Ianina A., Vorontsov K. [Regularized multimodal hierarchical topic model for document-by-document exploratory search](#). 2019.

А.Янина. Тематические и нейросетевые модели языка для разведочного информационного поиска. Дисс. к.ф.-м.н., МФТИ. 2022.

Мультиязычный тематический поиск и категоризация

Данные:

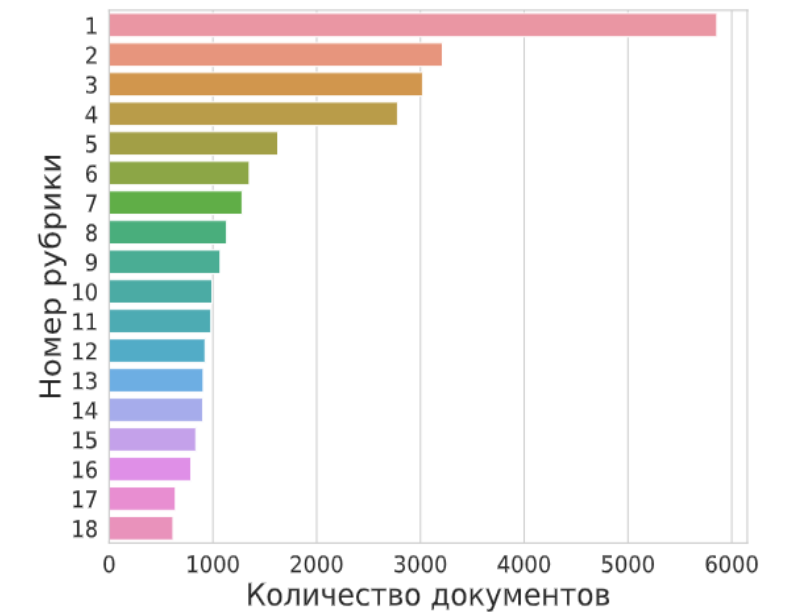
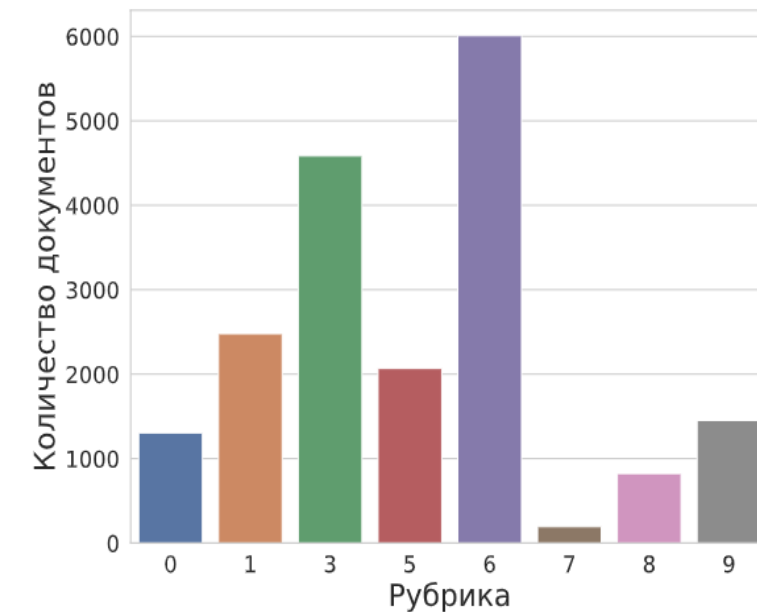
- научные статьи eLibrary и статьи Wikipedia (100 языков)
- рубрики ГРНТИ, ВАК, УДК, ОЭСР

Две задачи, одна модель:

- тематический поиск документов по документам
- категоризация документов

Особенности решения:

- модальности: языки, рубрики
- редукция словарей (VPE-токенизация)
до 11 тыс. токенов на каждый язык
- сокращение модели с 128 Гб до 4.8 Гб

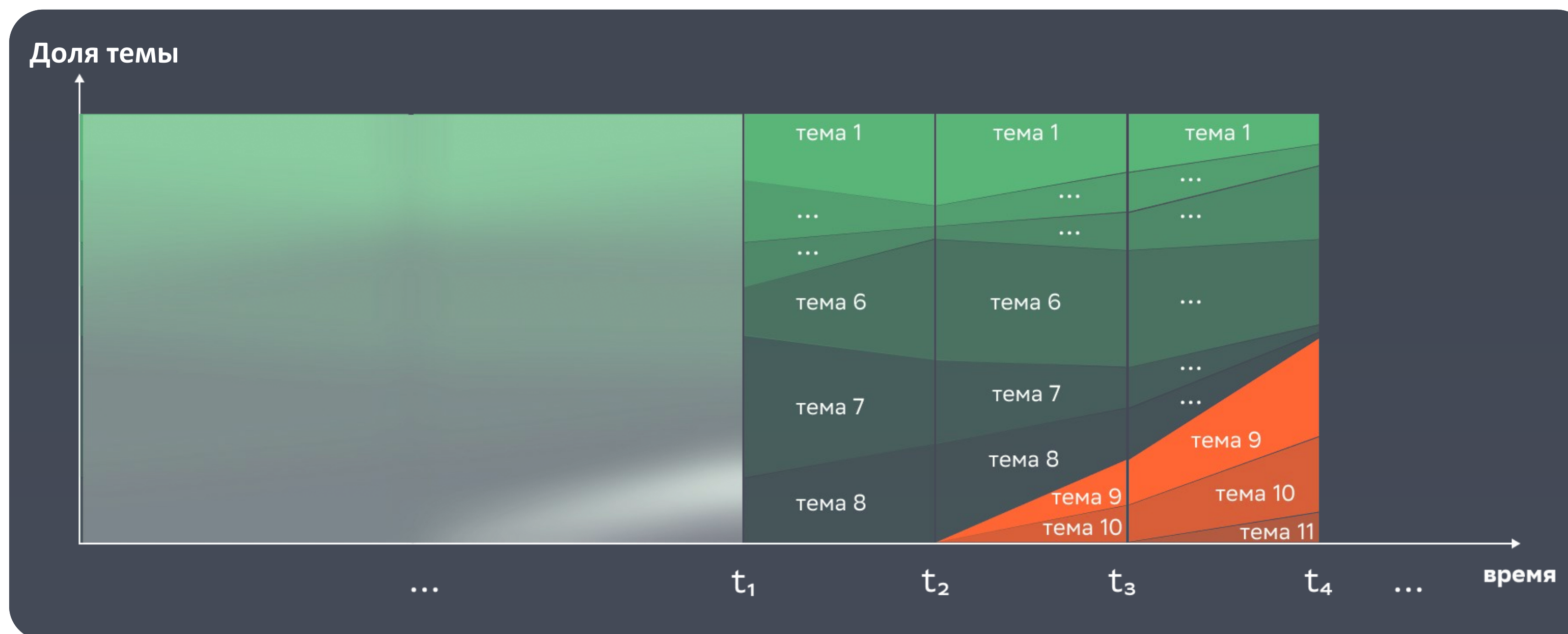


94%
Точность
поиска

Классификатор	ГРНТИ	ВАК	УДК	ОЭСР
Точность	81%	70%	86%	80%

Поиск научных трендов

- *Темпоральная тематическая модель* последовательно дообучается на статьях, вышедших за 30 дней
- Удаётся детектировать >60% из 87 трендовых тем (из области Data Science), выделенных экспертами в течение года после появления темы

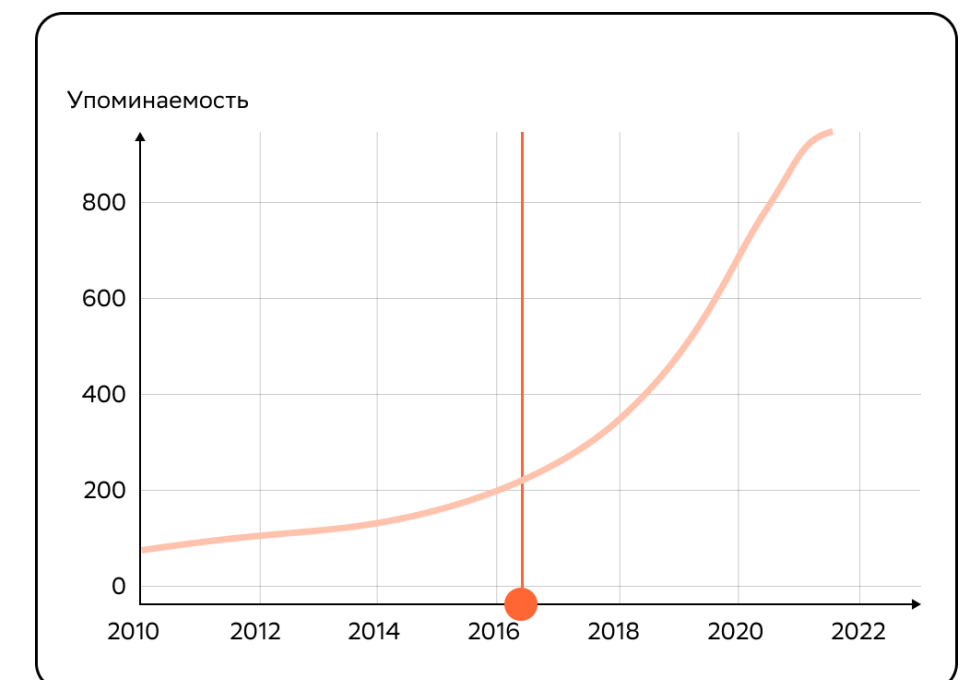
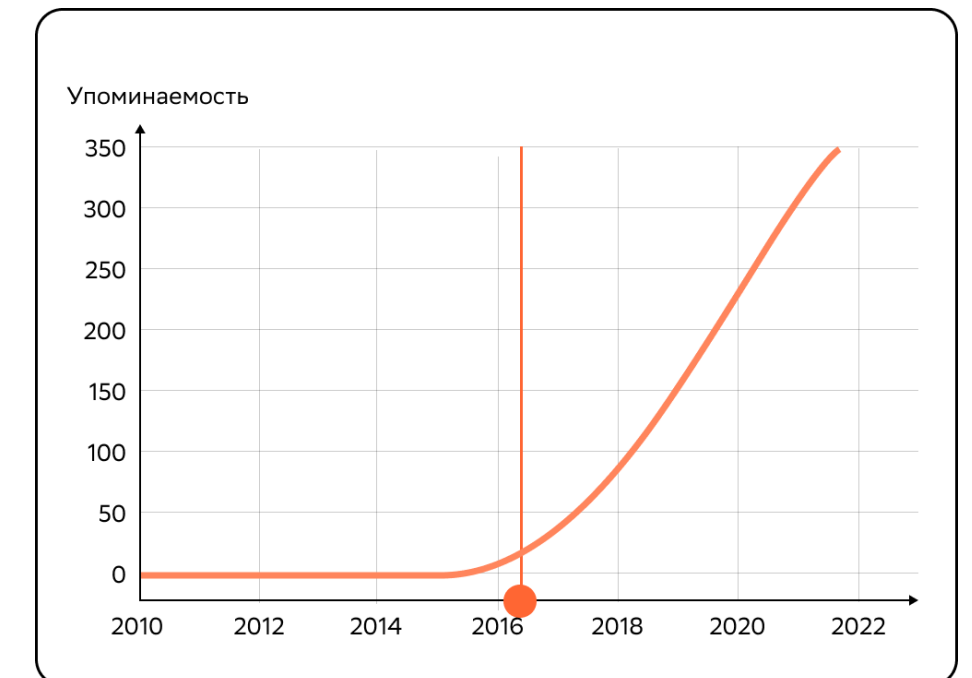
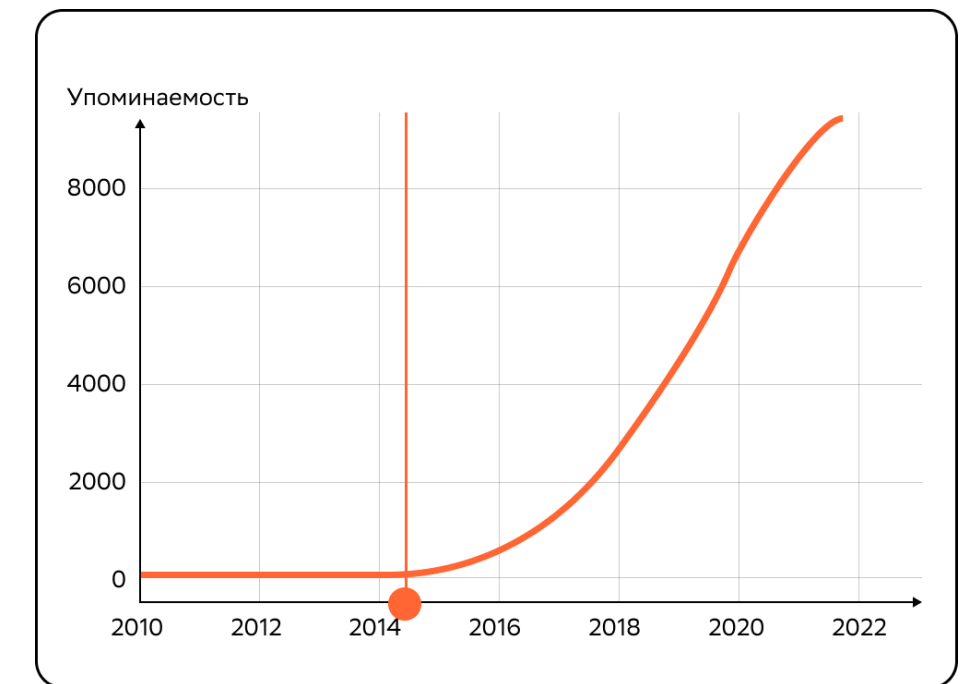
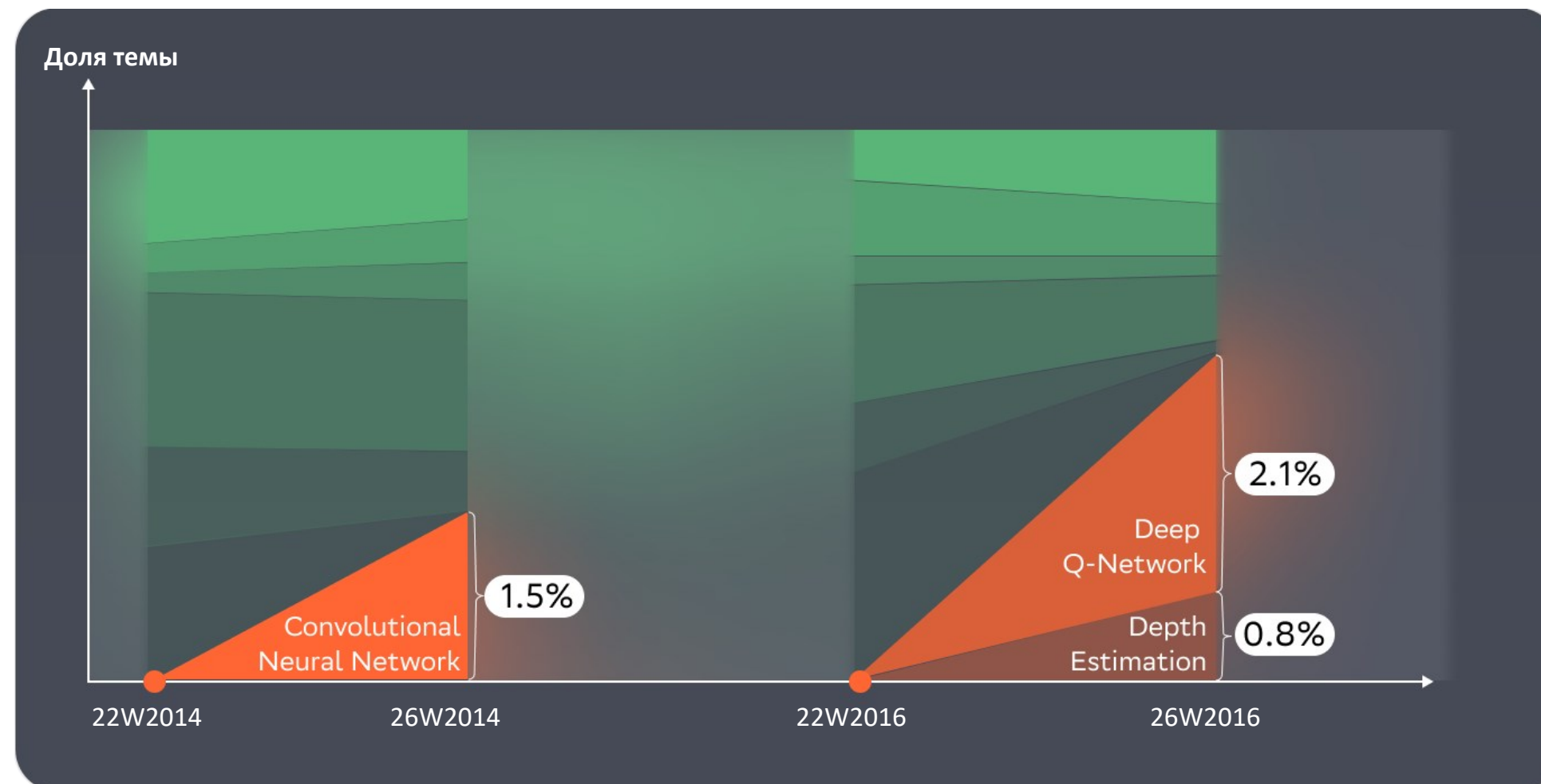


Поиск научных трендов

Трендовая тема:

- наличие семантического ядра
- наличие быстрого (обычно экспоненциального) роста

Примеры: динамика упоминаний трендовых тем



Поиск научных трендов: примеры тем

Topic modeling

latent variable

mixture model

topic model

mixture component

Gibbs sampling

multinomial distribution

Gibbs sampler

generative process

Dirichlet distribution

Dirichlet process

Speech recognition

prosodic feature

speech signal

eye gaze

audio signal

spontaneous speech

topic segmentation

acoustic feature

ASR output

switchboard corpus

audio data

Collaborative filtering

web page

search result

recommender system

collaborative filtering

word sense

ranking model

web search

user preference

user profile

ranking score

Machine translation

word alignment

target language

bleu score

parallel corpus

source sentence

translation model

machine translation

sentence pair

source language

best list

Поиск научных трендов: примеры тем

StyleGAN

stylegan

latent code

mapping network

ablation study

text generation

generation quality

generator architecture

mask

encoder

gan model

Meta Learning

meta model

meta train

meta optimization

meta update

meta testing

training task

continual learning

previous task

catastrophic forgetting

ablation study

NERF

neural radiance field

accurate depth estimation

additional qualitative result

novel loss function

optical flow prediction

image reconstruction loss

monocular depth prediction

geometric consistency loss

depth estimation method

optical flow network

Технология автоматического выделения терминов

Объединение трёх технологий: TopMine & SyntaxNet & BigARTM

- Коллекция $|D| = 3200$ аннотаций статей NIPS (Neural Information Processing Systems), $n = 500\,000$ слов
- Ручная разметка небольшого случайного подмножества (2000 n -грамм) на термины / не-термины
- Train : Test = 1000 : 1000
- 7 статистических признаков из TopMine
- 2 синтаксических признака из SyntaxNet
- 3 тематических признака из BigARTM, 30 тем
- две модели классификации:
логистическая регрессия, градиентный бустинг

Группа признаков			Линейная модель			Градиентный бустинг		
Синт	Стат	Тем	AUC	Точность	Полнота	AUC	Точность	Полнота
+			0.83	0.20	0.91	0.83	0.20	0.91
	+		0.71	0.09	0.94	0.73	0.11	0.90
		+	0.92	0.32	1.00	0.95	0.32	1.00
+	+		0.88	0.22	0.91	0.88	0.24	0.91
+		+	0.91	0.36	0.91	0.95	0.34	0.99
	+	+	0.93	0.29	0.94	0.98	0.34	1.00
+	+	+	0.95	0.38	0.91	0.97	0.41	0.99

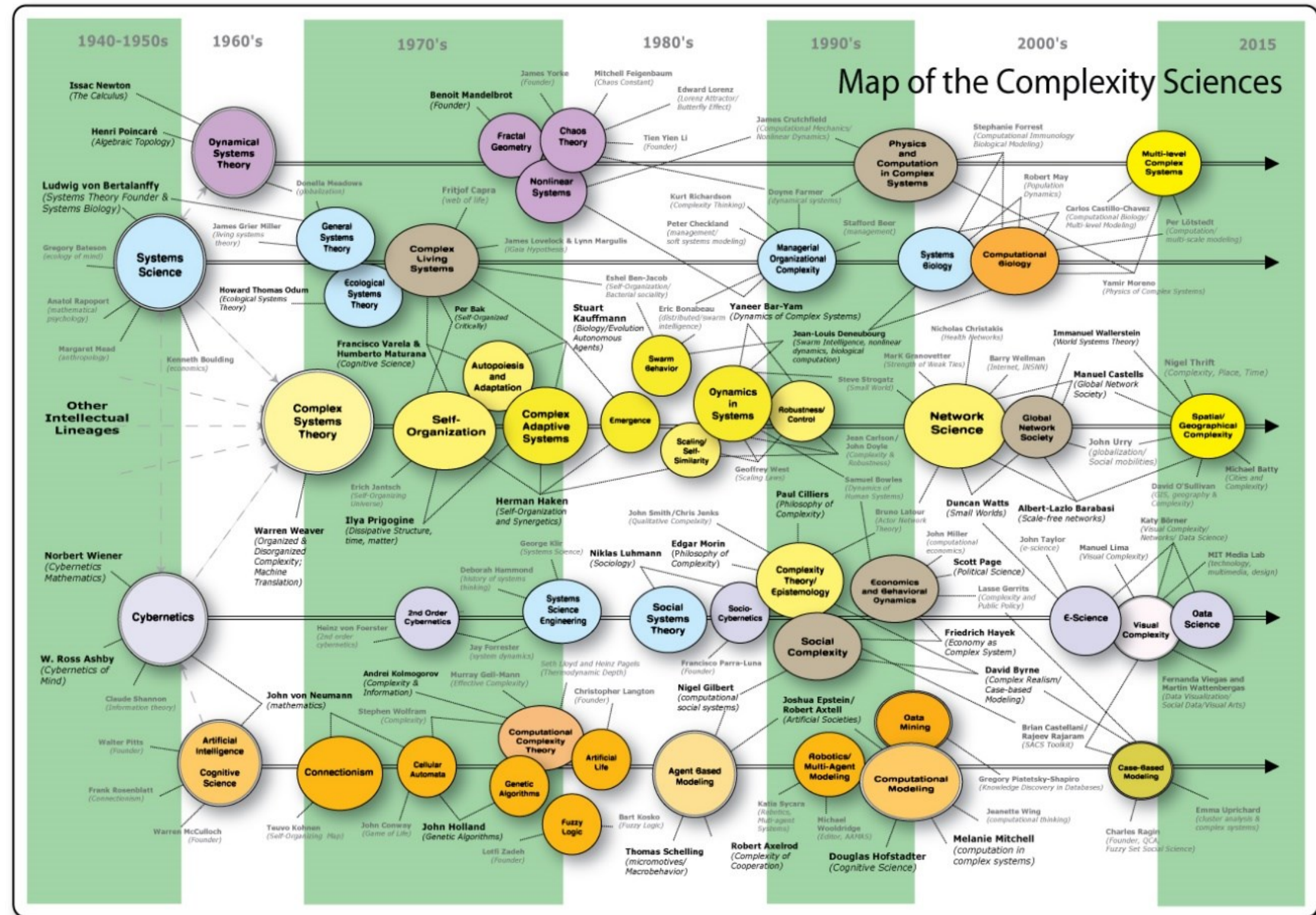
$$\boxed{\text{Стат}} < \boxed{\text{Син}} < \boxed{\text{Син+Стат}} < \boxed{\text{Тем}} < \boxed{\begin{matrix} \text{Стат+Тем} \\ \text{Син+Тем} \end{matrix}} < \boxed{\text{Стат+Син+Тем}}$$

- Тематические признаки существенно повышают качество
- Синтаксические признаки можно не использовать

Визуализация и дистантное чтение (distant reading)

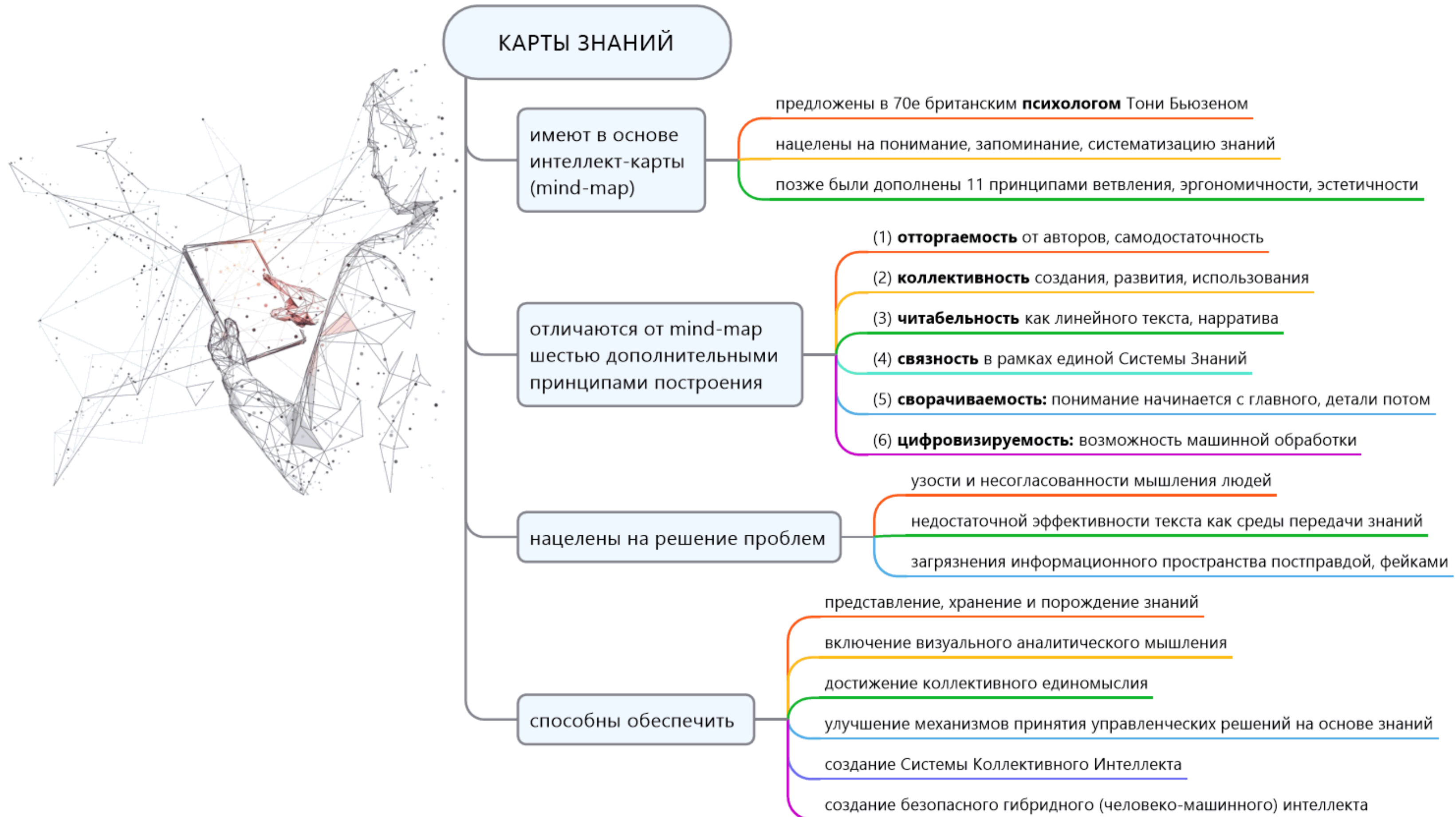
Осями на карте могут быть:

- время
- спектр тем
- сложность
- обзорность
- актуальность
- «хайповость»
- цитируемость



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

Автоматическое построение карт знаний



Спасибо за внимание!

Воронцов Константин Вячеславович
д.ф.-м.н., профессор РАН,
зав. лабораторией МОСА ИИИ МГУ,
зав. кафедрой ММП ВМК МГУ

voron@mlsa-iai.ru

<http://www.MachineLearning.ru/wiki?title=User:Vokov>