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

Мастерская знаний

Воронцов Константин Вячеславович

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

Научный семинар «Проблемы управления знаниями»
Институт Проблем Управления им. В.А.Трапезникова РАН

21 февраля 2024 г.

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

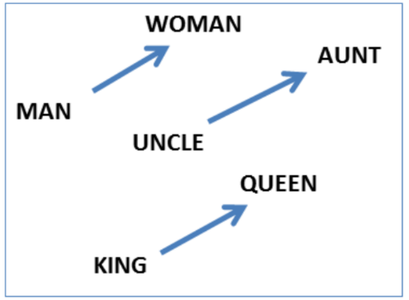
Эволюция подходов в обработке естественного языка

Декомпозиция задач по уровням «пирамиды NLP»

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

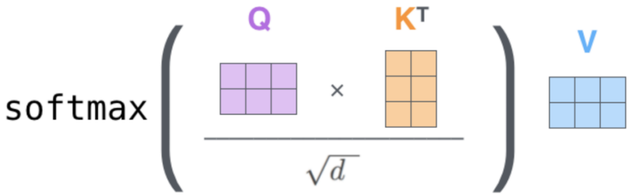
Модели векторных представлений слов (эмбедингов)

- 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],…

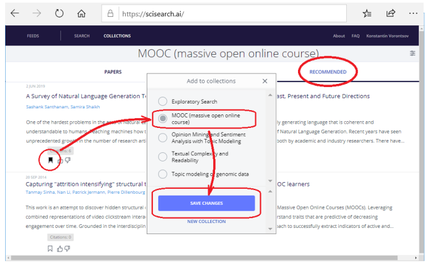


Функции «Мастерской знаний»

Подборка текстов – поисковый интерес и рабочее пространство пользователя или группы

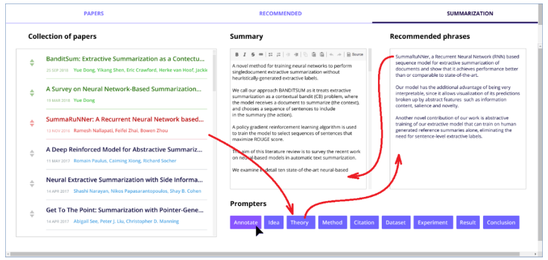
Поисково-рекомендательные сервисы:

- **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



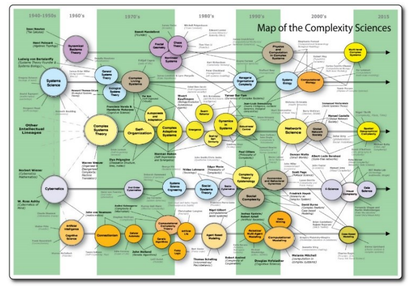
Аналитические сервисы:

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



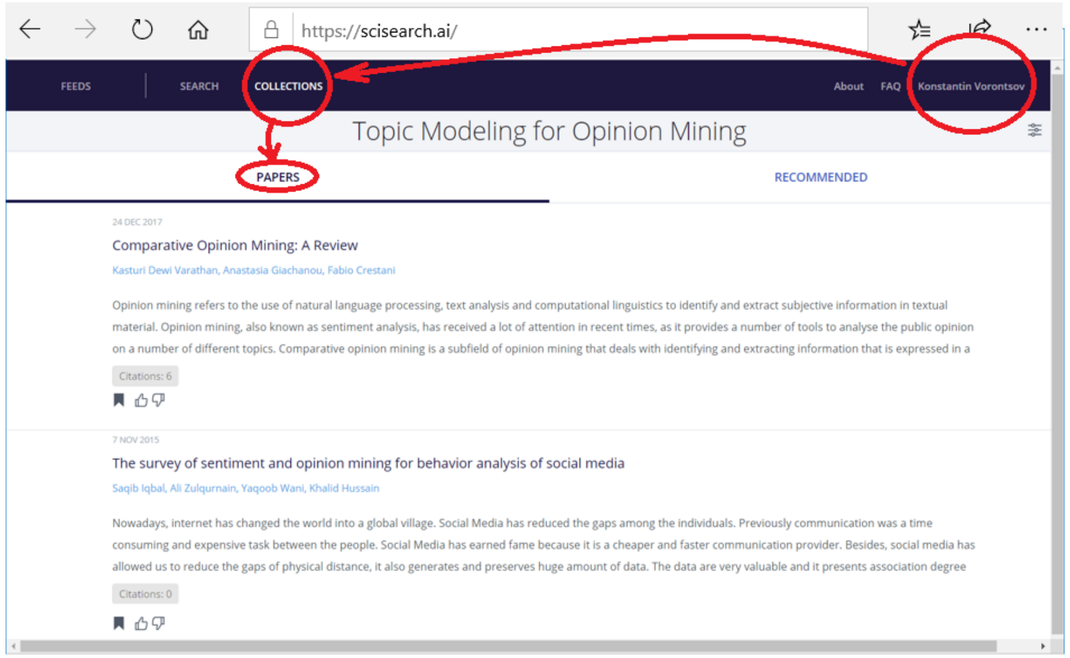
Коммуникативные сервисы:

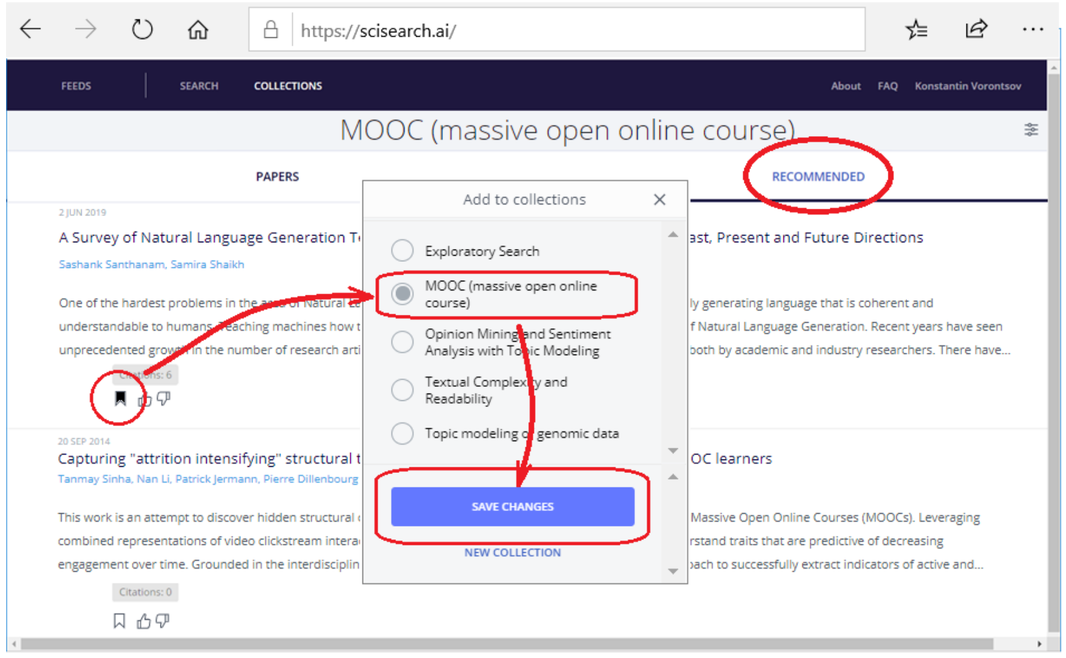
- совместное составление, обсуждение, использование подборок
- интерактивная визуализация и инфографика по подборке



Поиск и рекомендации (прототип SciSearch.ai)

Подборка играет роль поискового запроса и поисковой выдачи одновременно





Технология тематического поиска

Схема эксперимента:

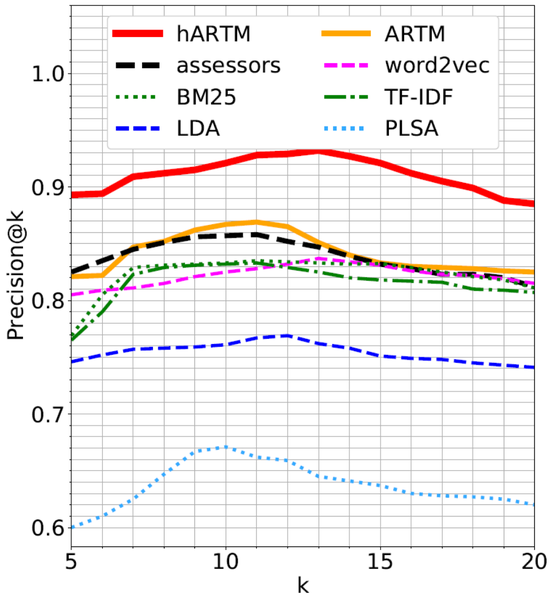
- длинные запросы (1 стр. A4)
- 100 запросов на коллекцию
- 3 асессора на каждый запрос
- от 10 до 60 минут на запрос
- разметка на Яндекс.Толока
- две коллекции техно-новостей:

(170K Russian docs)

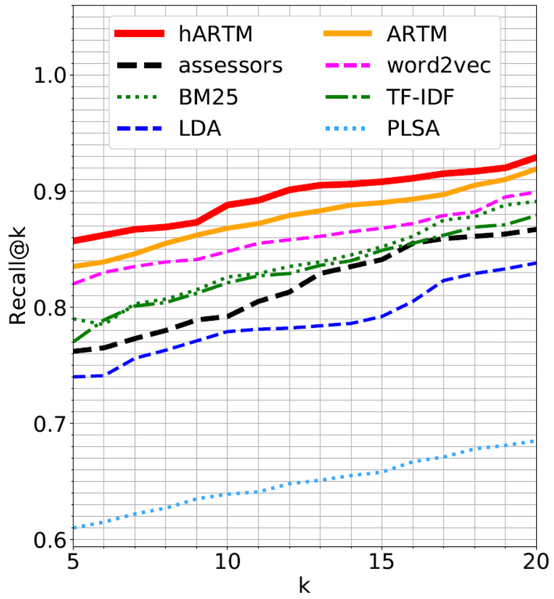
(750K English docs)

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.

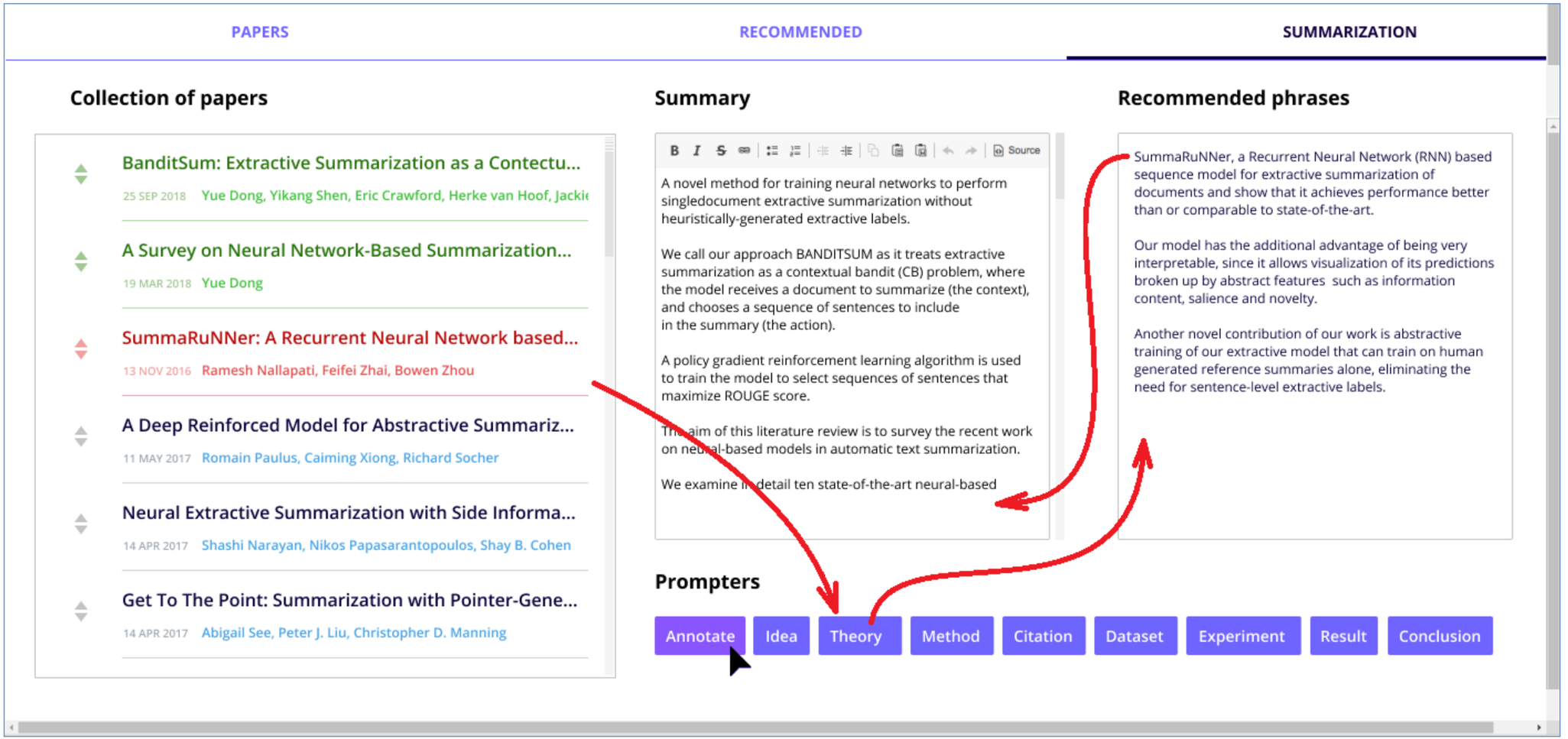
Полуавтоматическое реферирование подборки

Концепция MAHS (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

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

Полуавтоматическое реферирование подборки



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

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

Полуавтоматическое реферирование подборки

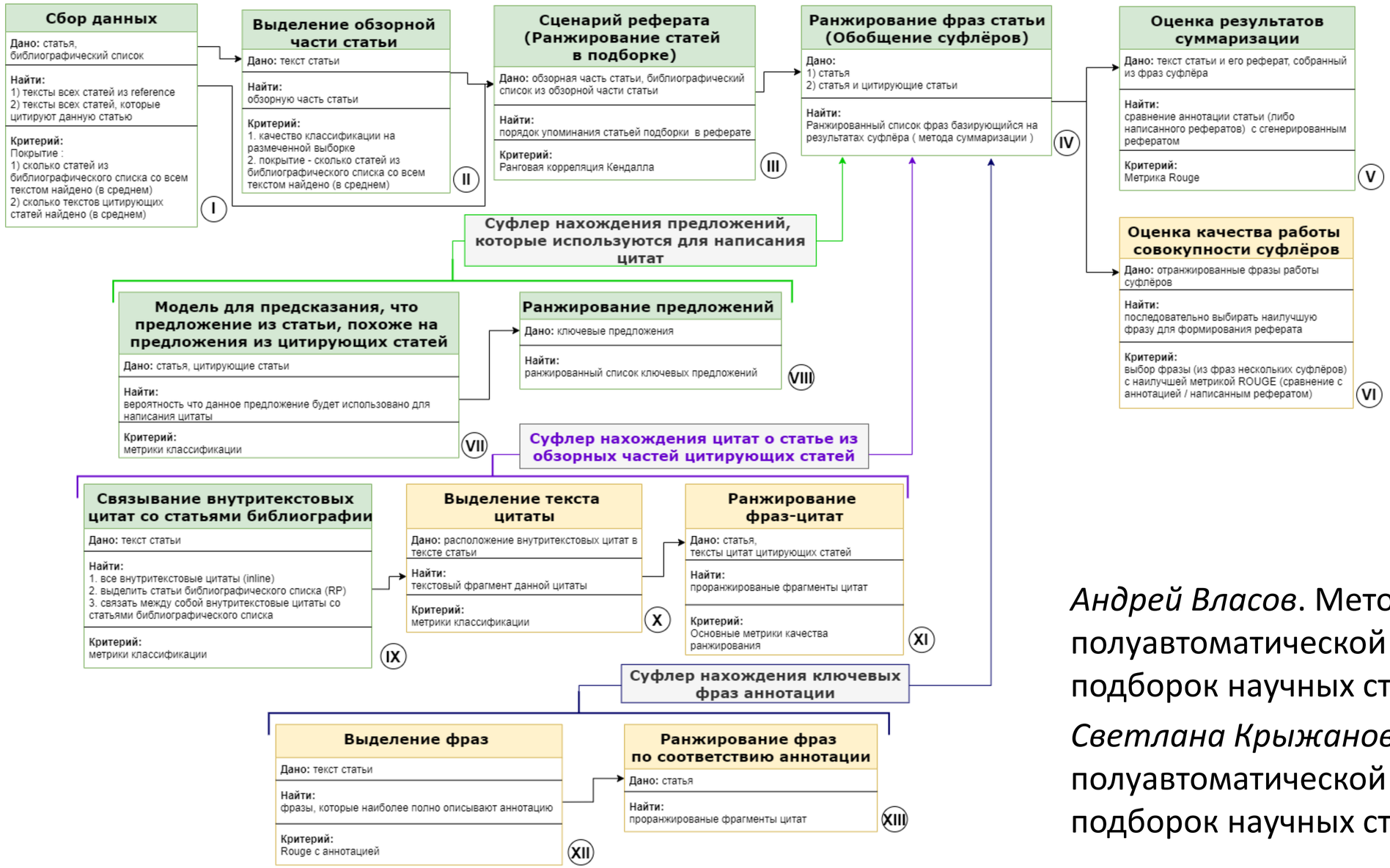
Основные задачи машинного обучения:

- 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



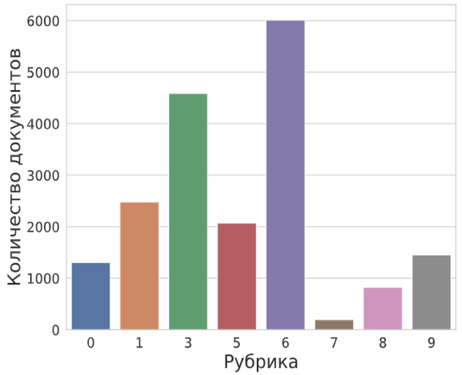
Андрей Власов. Методы полуавтоматической суммаризации подборок научных статей. МФТИ, 2020

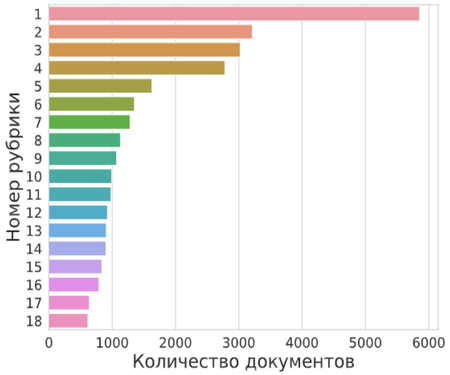
Светлана Крыжановская. Технология полуавтоматической суммаризации подборок научных статей. МГУ, 2022

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

94%
Точность поиска

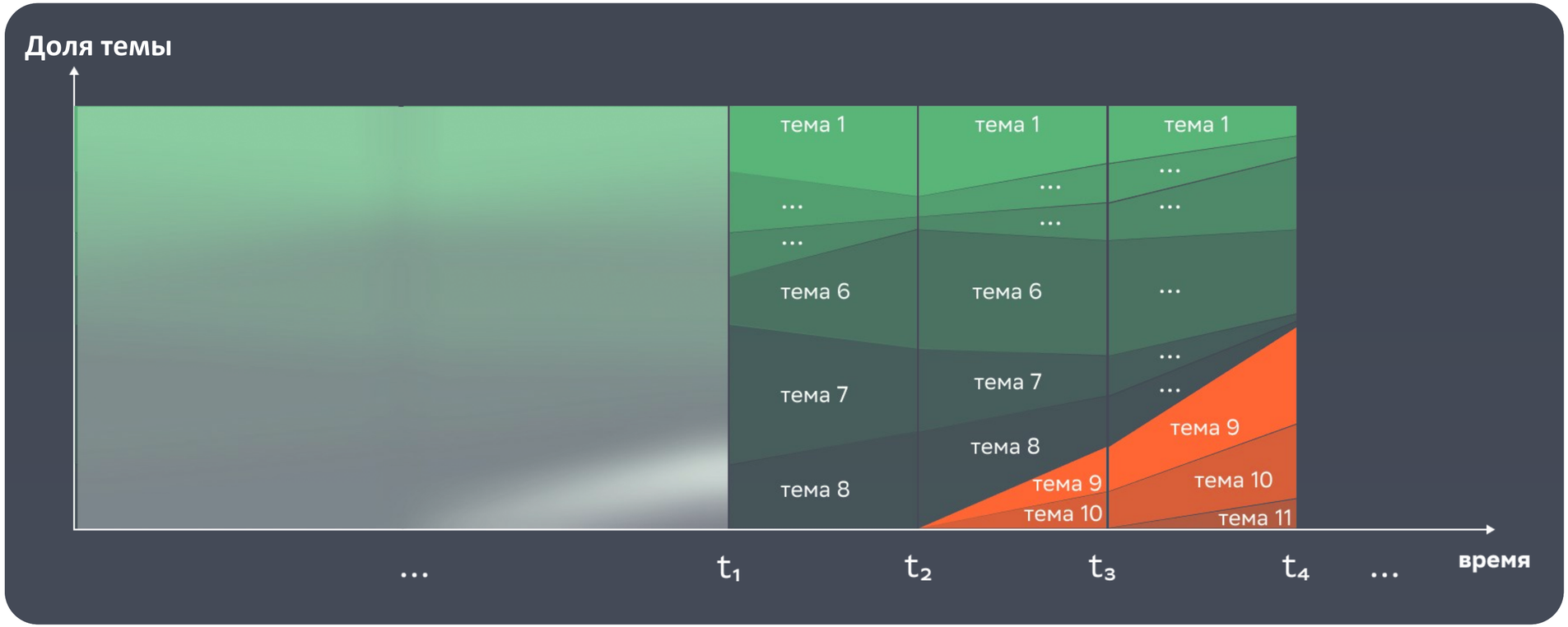
Особенности решения:

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

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

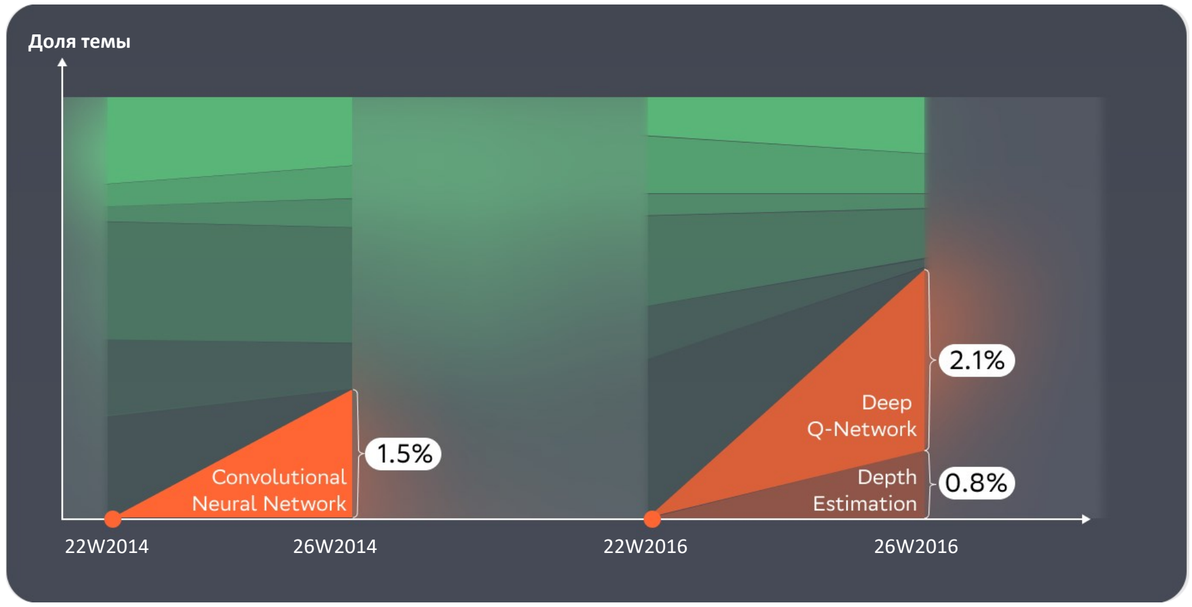


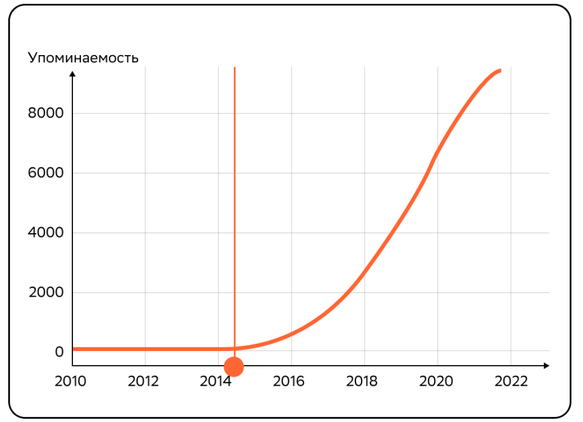
Поиск научных трендов

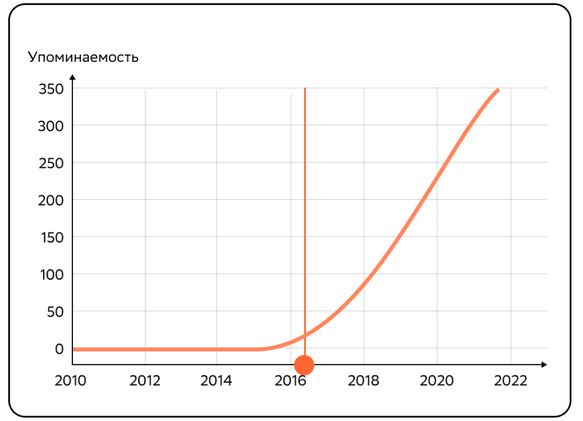
Трендовая тема:

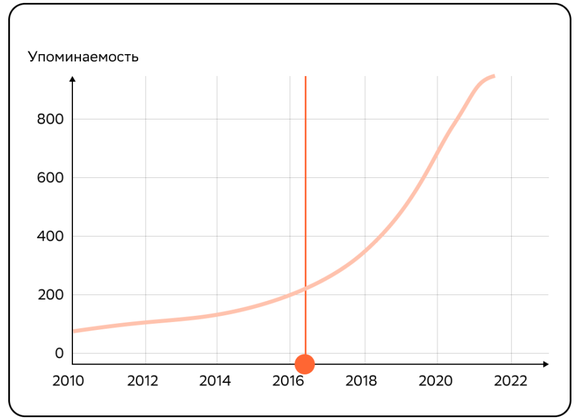
- наличие семантического ядра
- наличие быстрого (обычно экспоненциального) роста

Примеры: динамика упоминаний трендовых тем









Поиск научных трендов: примеры тем

Topic modeling	Speech recognition	Collaborative filtering	Machine translation
latent variable	prosodic feature	web page	word alignment
mixture model	speech signal	search result	target language
topic model	eye gaze	recommender system	bleu score
mixture component	audio signal	collaborative filtering	parallel corpus
Gibbs sampling	spontaneous speech	word sense	source sentence
multinomial distribution	topic segmentation	ranking model	translation model
Gibbs sampler	acoustic feature	web search	machine translation
generative process	ASR output	user preference	sentence pair
Dirichlet distribution	switchboard corpus	user profile	source language
Dirichlet process	audio data	ranking score	best list

Поиск научных трендов: примеры тем

StyleGAN	Meta Learning	NERF
stylegan	meta model	neural radiance field
latent code	meta train	accurate depth estimation
mapping network	meta optimization	additional qualitative result
ablation study	meta update	novel loss function
text generation	meta testing	optical flow prediction
generation quality	training task	image reconstruction loss
generator architecture	continual learning	monocular depth prediction
mask	previous task	geometric consistency loss
encoder	catastrophic forgetting	depth estimation method
gan model	ablation study	optical flow network

Технология тематического моделирования BigARTM

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Онлайновый 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

3.7M статей Википедии, 100K 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 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**

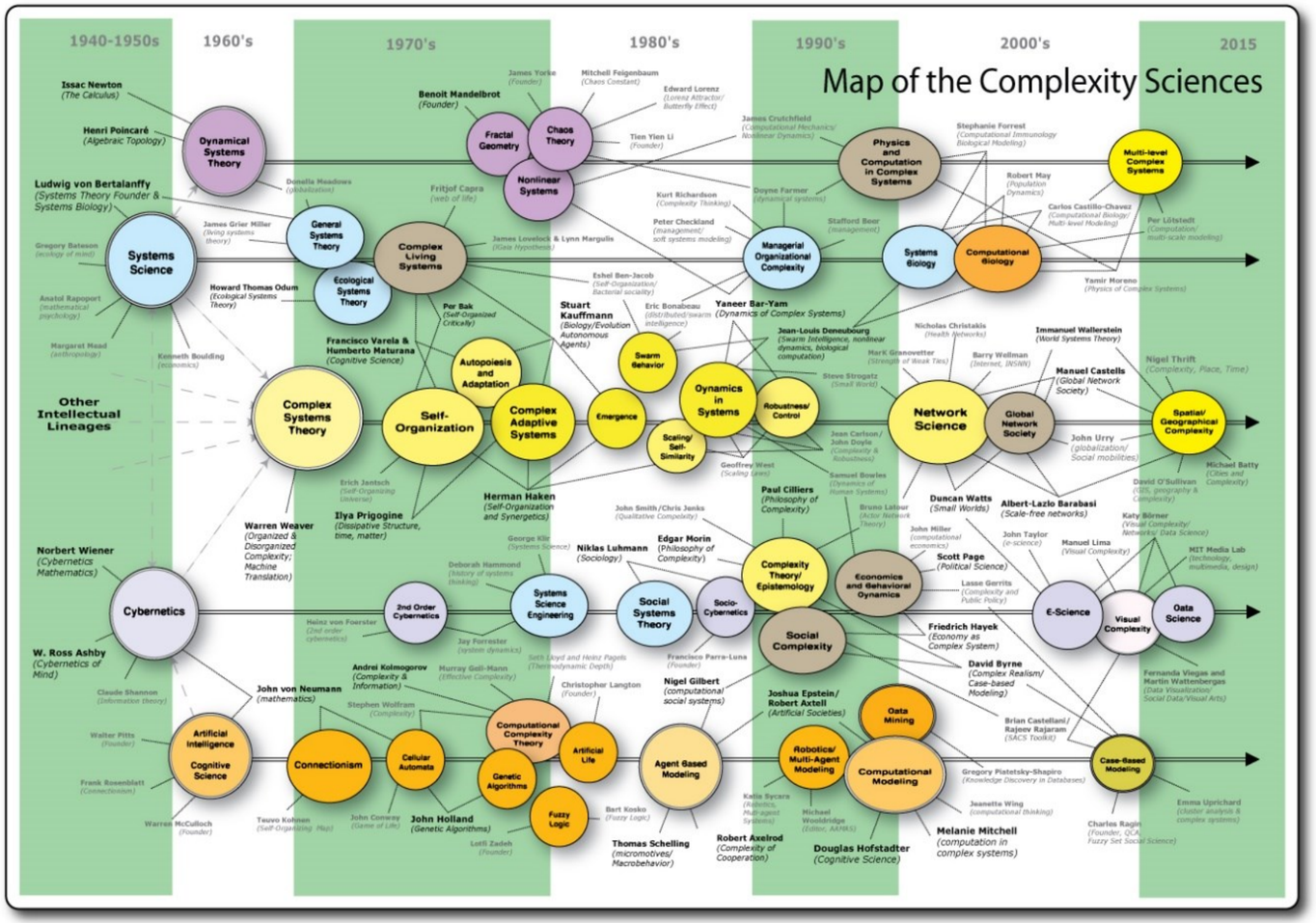
Стат < Син < Син+Стат < Тем < Стат+Тем / Син+Тем < Стат+Син+Тем

- 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

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