Задачи и методы автоматического анализа текстов в разведочном информационном поиске

Воронцов Константин Вячеславович

(Лаборатория машинного интеллекта МФТИ • ФИЦ ИУ РАН)

Москва • 26-29 ноября 2019

Содержание

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*

От поиска информации к «Мастерской знаний»

Обычный поиск:

- «нашёл и забыл»

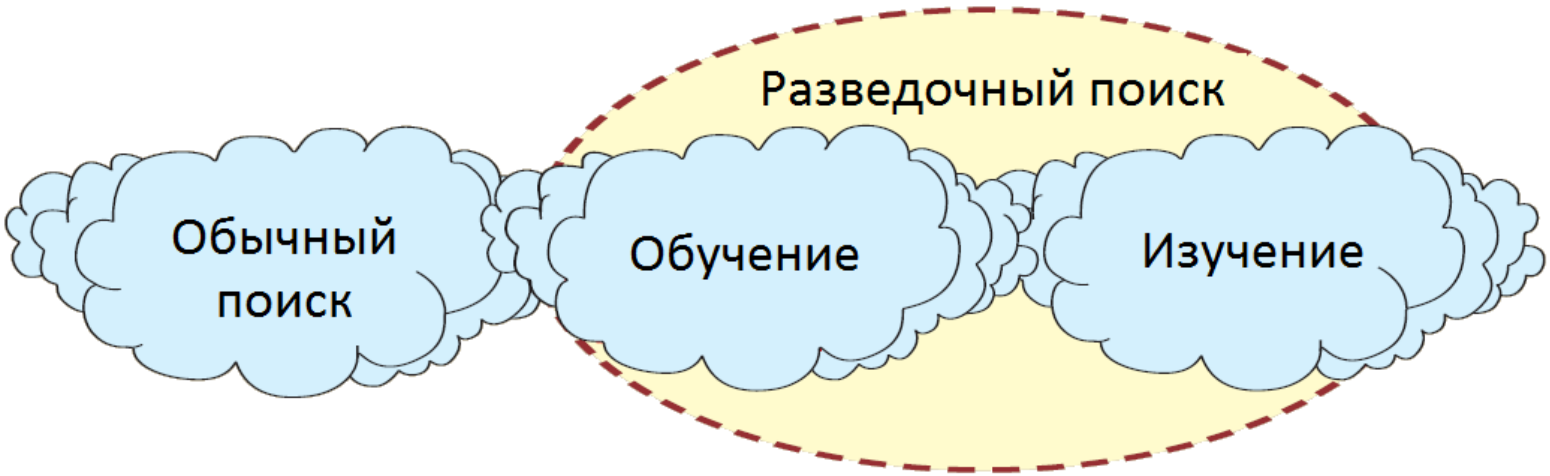
Мастерская знаний — инструментарий для автоматизации последующих этапов работы с профессиональными знаниями:

- ищу – чтобы накапливать
- накапливаю – чтобы анализировать
- анализирую – чтобы понимать
- понимаю – чтобы передавать

Эти задачи связаны с *автоматическим анализом текстов* (применение знаний остаётся за пределами системы).

Концепция разведочного поиска (exploratory search)

- пользователь может не знать ключевых терминов
- запросом может быть текст произвольной длины
- информационная потребность — систематизация знаний



навигация в сети,	самообразование,	исследование,
поиск фактов,	тематический поиск	экспертиза,
упоминаний,	систематизация	реферирование,
конкретных ответов	знаний	мониторинг тем

Gary Marchionini. Exploratory Search: from finding to understanding. 2006.

Определения и модели разведочного поиска

Определение *разведочного поиска* через 11 его свойств:

1. **An evolving search process**
   разведочный поиск – это многошаговый процесс
   каждый шаг – переформулировка или дополнение запроса
2. **An anomalous state of knowledge**
   в начале поиска у пользователя есть лишь мотивации,
   но нет знаний и нет определённого плана, как их получать
3. **Multiple targets / goals of search**
   нет конкретной, точно определённой цели поиска
   есть лишь общий интерес и эволюционирующие подцели

E.Palagi et al. A Survey of Definitions and Models of Exploratory Search. 2017.

Определения и модели разведочного поиска

Свойства *неопределённости* процесса разведочного поиска

4. **Multiple possible answers**
   возможных правильных ответов может быть много
5. **Not an expected exact answer**
   не существует единственного правильного ответа
6. **A serendipitous attitude**
   любой шаг может давать неожиданные новые знания
7. **An evolving information need**
   на любом шаге цели и стратегии поиска могут измениться
8. **Uncertainty is fluctuating**
   в процессе поиска неопределённость уменьшается,
   но изменение цели может снова её увеличить

E.Palagi et al. A Survey of Definitions and Models of Exploratory Search. 2017.

Определения и модели разведочного поиска

Свойства *разветвлённости* процесса разведочного поиска

9. **Multifaceted search**
   при поиске используются различные фильтры (фасеты), например, по авторам, тематике, свежести, сложности
10. **Several one-off pinpoint searches**
    многократные точечные одноразовые ответвления поиска, например, чтобы уточнить понятие, первоисточник, и т.п.
11. **An open-ended search activity which can occur over time**
    процесс поиска никогда на заканчивается
    пользователь может вернуться после долгого перерыва

E.Palagi et al. A Survey of Definitions and Models of Exploratory Search. 2017.

Концепция сервиса тематического разведочного поиска

Подборка — долгосрочный поисковый интерес пользователя

Поисково-рекомендательные функции:

- поиск семантически близких документов по подборке
- мониторинг новых документов для подборки

Аналитические функции:

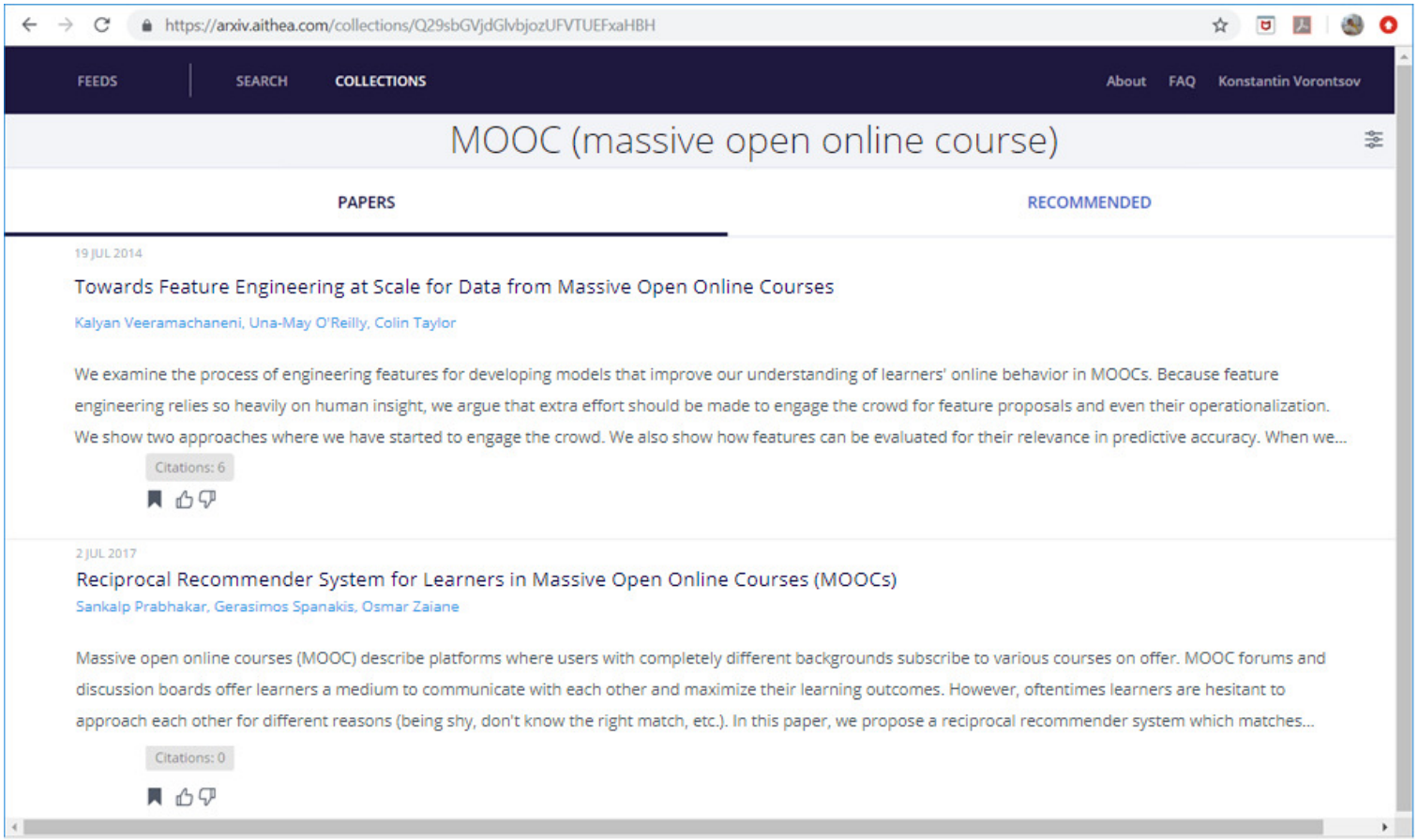
- рекомендация порядка чтения внутри подборки
- автоматизация реферирования подборки
- кластеризация тем, идей, мнений во всей подборке
- выделение ключевых понятий, фактов, идей из документа

Коммуникативные функции:

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

Поисково-рекомендательная система arXiv.AITHEA.com

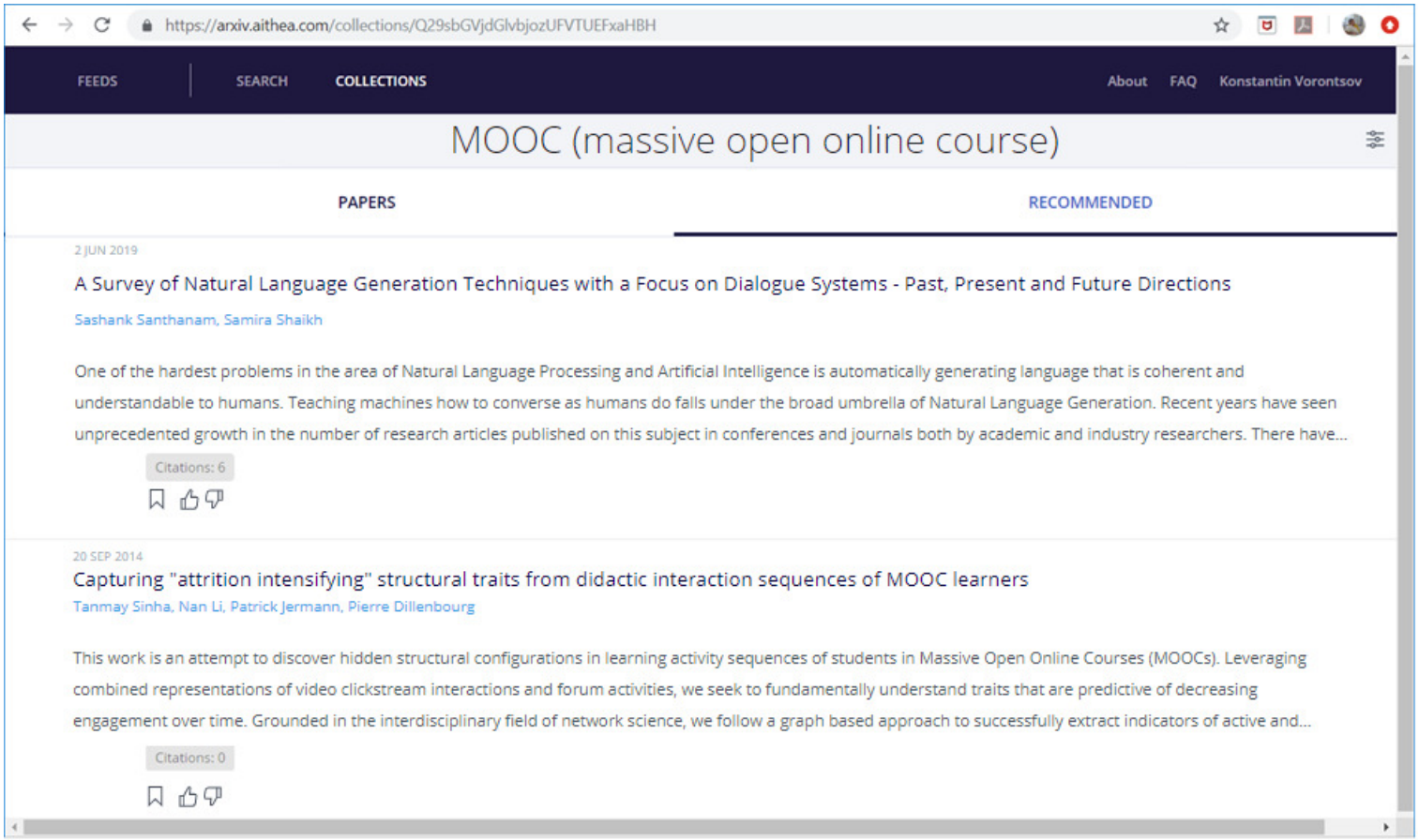
Тематическая подборка пользователя:



Разработка компаний AITHEA и Digital Decisions (http://ddecisions.ai)

Поисково-рекомендательная система arXiv.AITHEA.com

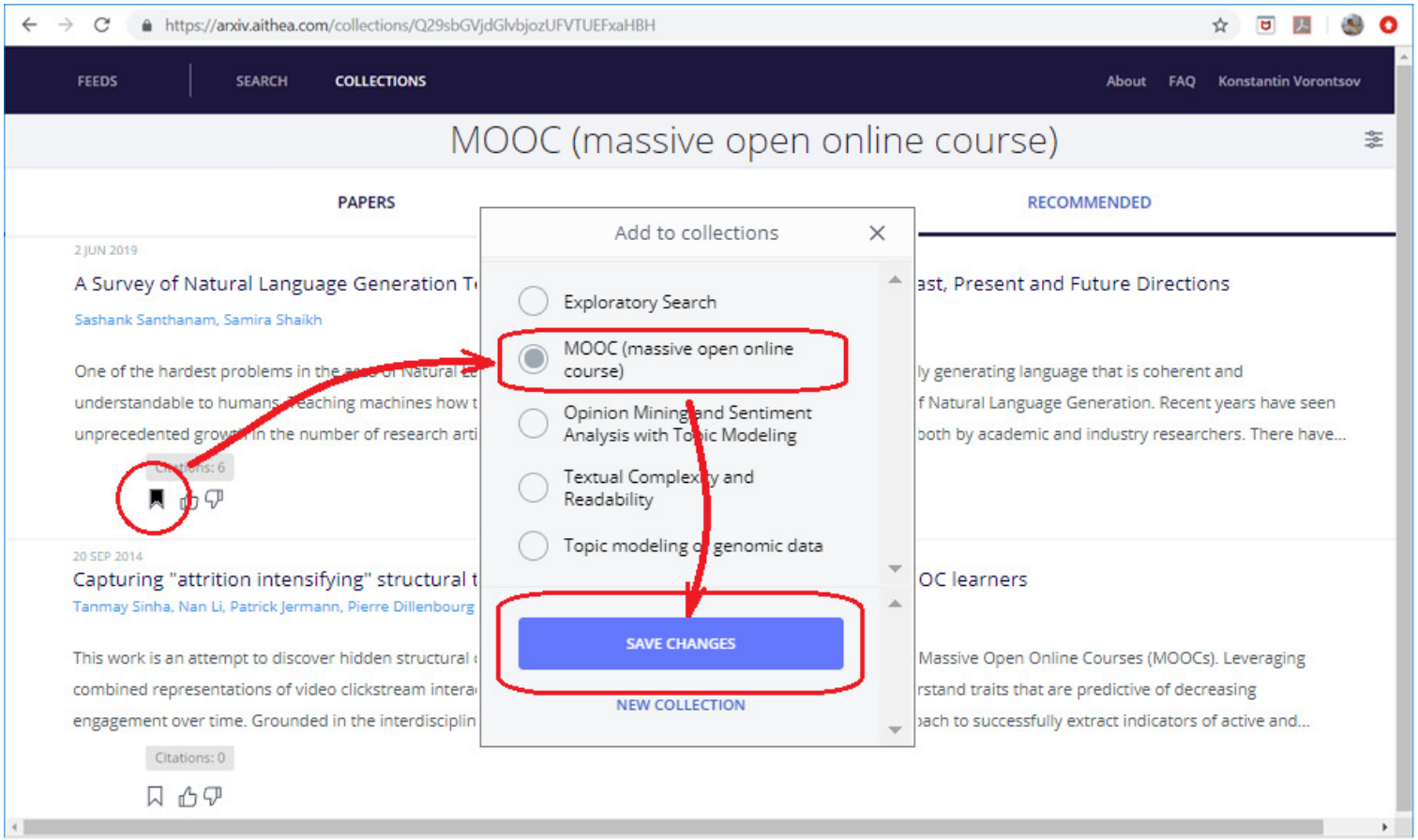
Список статей, рекомендуемых для добавления в подборку:



Разработка компаний AITHEA и Digital Decisions (http://ddecisions.ai)

Поисково-рекомендательная система arXiv.AITHEA.com

Добавление статьи из списка рекомендаций в подборку:



Разработка компаний AITHEA и Digital Decisions (http://ddecisions.ai)

Аналитические функции в приложениях

Научные публикации

- *Подбираем* статьи по тематике исследований
- *Разделяем* по подтемам, научным школам, терминологии
- *Находим* тренды, подходы, открытые проблемы

Массовые открытые онлайн-курсы

- *Подбираем* курсы по тематике
- *Разделяем* по уровням, пререквизитам, компетенциям
- *Находим* персональные образовательные траектории

Научно-популярный и просветительский контент

- *Подбираем* статьи по тематике интересов
- *Разделяем* по подтемам, возрасту целевой аудитории
- *Находим* «точки входа» в науку, порядок чтения

Аналитические функции в приложениях

Отзывы и обзоры о потребительских товарах

- *Подбираем* отзывы по назначению товара
- *Разделяем* по потребительским свойствам товара
- *Находим* аргументацию для принятия решения о покупке

Новостные потоки

- *Подбираем* новости по теме, проблеме или событию
- *Разделяем* по тональности, акцентированию, умалчиванию
- *Находим* полярные мнения и их источники

Акты арбитражных судов

- *Подбираем* дела, схожие по существу
- *Разделяем* по исходу дела
- *Находим* аргументацию для суда

Задачи, методы и технологии NLP для «Мастерской знаний»

Основные задачи:

- Семантические векторные представления текста
- Иерархическое тематическое моделирование
- Суммаризация и аннотирование текстов
- Тематическая сегментация текста
- Ранжирование документов в порядке чтения
- Визуализация больших текстовых коллекций

Вспомогательные задачи:

- Морфологический и синтаксический парсинг
- Выделение терминов, фактов, тональности
- Кластеризация и классификация текстов
- Оценивание когнитивной сложности текста
- Модели обучения ранжированию

Задача тематического моделирования

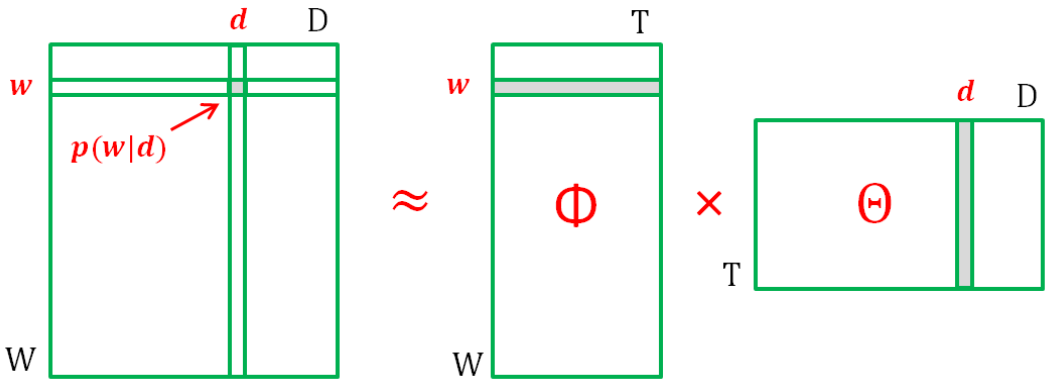
Дано: коллекция текстовых документов, $p(w|d) = \frac{n_{dw}}{n_d}$

Вероятностная тематическая модель:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

Найти: параметры модели $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$

Это задача стохастического матричного разложения:



Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

ARTM: аддитивная регуляризация тематических моделей

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

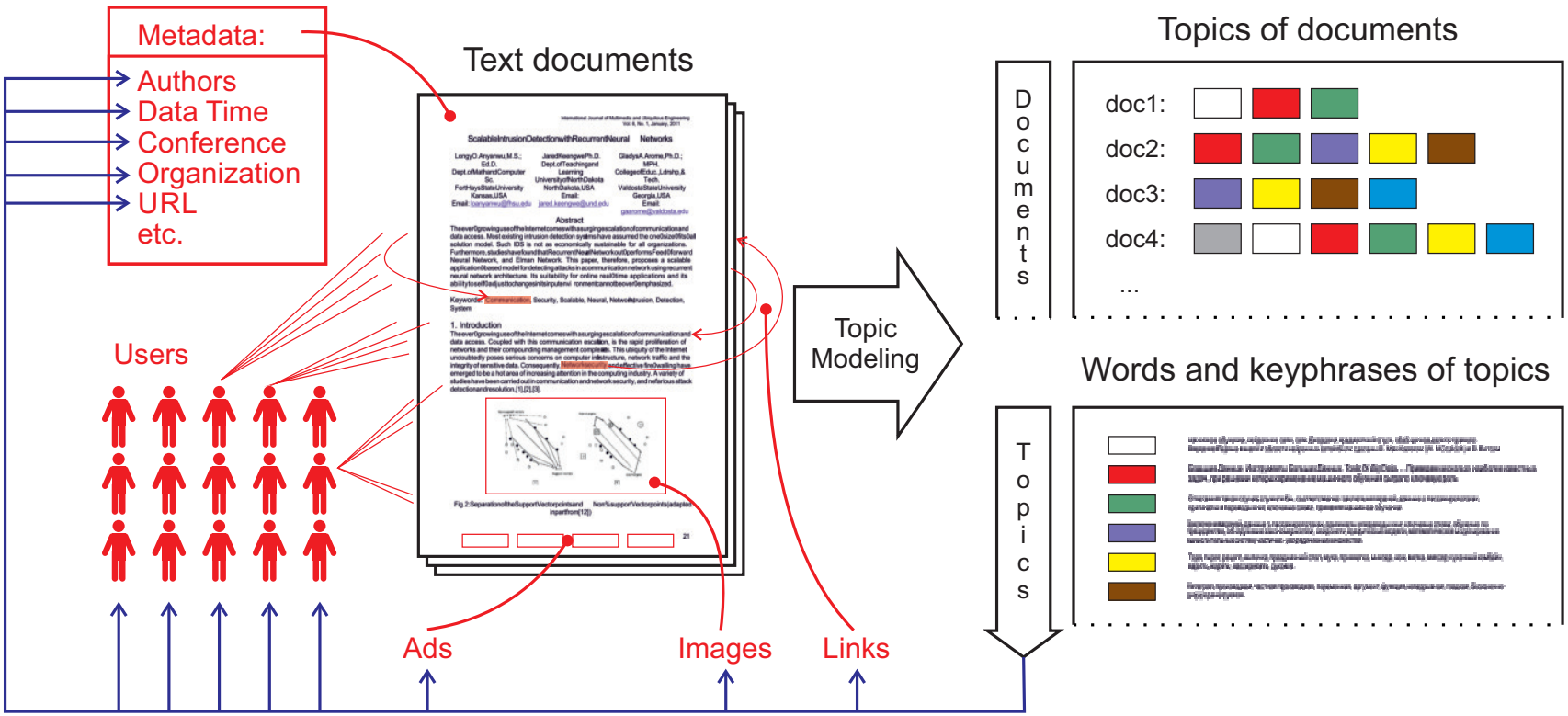
$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_d n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_w n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

где $\underset{t\in T}{\text{norm}}(x_t) = \dfrac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_s,0\}}$ — операция нормирования вектора.

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014.

Задачи мультимодального тематического моделирования

Темы определяют распределения не только терминов $p(w|t)$, но и других *модальностей*: $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{элемент\_изображения}|t)$, $p(\text{пользователь}|t), \ldots$



Мультимодальная ARTM

W^m — словарь токенов m-й модальности, $m \in M$

Максимизация суммы log правдоподобий с регуляризацией:

$$\sum_{m\in M} \tau_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W^m}{\text{norm}}\left(\sum_{d\in D} \tau_{m(w)} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

K.Vorontsov, O.Frei, M.Apishev et al. Non-bayesian additive regularization for multimodal topic modeling of large collections. CIKM TM workshop, 2015.

Модальность биграмм улучшает интерпретируемость тем

Коллекция 1000 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели // Магистерская диссертация, МФТИ, 2015.

Иерархическая тематическая модель: послойное построение

Шаг 1. Строим модель с небольшим числом тем.

Шаг *k*. Пусть модель с множеством тем T уже построена. Строим множество дочерних тем S (subtopics), $|S| > |T|$.

Родительские темы приближаются смесями дочерних тем:

$$\sum_{t\in T} n_{wt} \ln p(w|t) = \sum_{t\in T} n_{wt} \ln \sum_{s\in S} p(w|s)p(s|t) \;\to\; \max_{\Phi,\Psi},$$

где $p(s|t) = \psi_{st}$, $\Psi = (\psi_{st})_{S\times T}$ — *матрица связей*.

Родительская $\Phi^p \approx \Phi\Psi$, отсюда регуляризатор матрицы Φ:

$$R(\Phi,\Psi) = \tau \sum_{t\in T}\sum_{w\in W} n_{wt} \ln \sum_{s\in S} \phi_{ws}\psi_{st} \;\to\; \max.$$

Родительские темы t — *псевдо-документы* с частотами слов n_{wt}.

Дистрибутивная гипотеза и векторные представления слов

- Words that occur in the same contexts tend to have similar meanings [Harris, 1954].
- You shall know a word by the company it keeps [Firth, 1957].

Дано: n_{uw} — частота пары слов u, w в окне $\pm h$ слов
Найти: векторные представления слов x_w и контекстов y_u
Модель: вероятность слова w в контексте слова u:

$$p(w|u) = \underset{w \in W}{\mathrm{SoftMax}} \langle x_w, y_u \rangle = \underset{w \in W}{\mathrm{norm}} \big(\exp \langle x_w, y_u \rangle \big)$$

Критерий максимума log-правдоподобия:

$$\sum_{w,u \in W} n_{wu} \ln p(w|u) \;\to\; \max_{\{x_w, y_u\}}$$

Z.Harris. Distributional structure. 1954.
J.R.Firth. A synopsis of linguistic theory 1930-1955. Oxford, 1957.

Модели векторных представлений текста

word2vec: *эмбединги* (векторные представления) слов
T.Mikolov et al. Efficient estimation of word representations in vector space. 2013.

paragraph2vec: эмбединги фрагментов или документов
Q.Le, T.Mikolov. Distributed representations of sentences and documents. 2014.

sent2vec: эмбединги предложений
M.Pagliardini et al. Unsupervised learning of sentence embeddings using compositional n-gram features. 2017.

FastText: эмбединги символьных *n*-грамм
`https://github.com/facebookresearch/fastText`

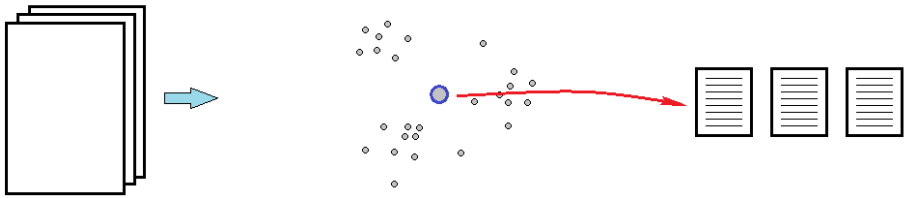
StarSpace: эмбединги чего угодно от Facebook AI Research
L.Wu, A.Fisch, S.Chopra, K.Adams, A.B.J.Weston. StarSpace: embed all the things! 2018.

BERT: эмбединги фраз и предложений от Google AI Language
J.Devlin et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018.

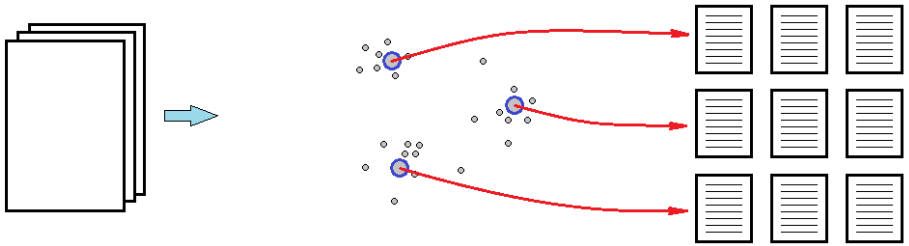
Преимущества тематических векторных представлений:
координаты соответствуют темам, интерпретируемые, разреженные, могут быть сделаны иерархическими

Поиск в пространстве семантических векторных представлений

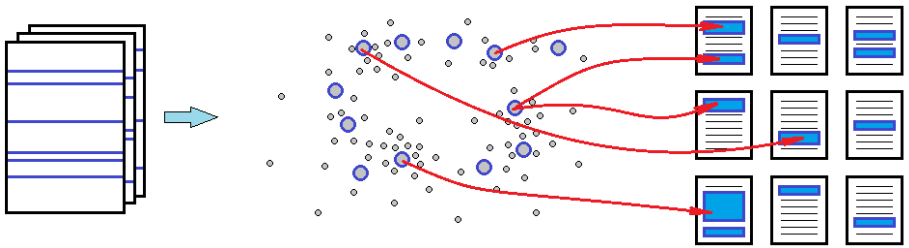
Поиск по среднему вектору подборки (неудачная стратегия):



Поиск по кластерам, построенным по документам подборки:

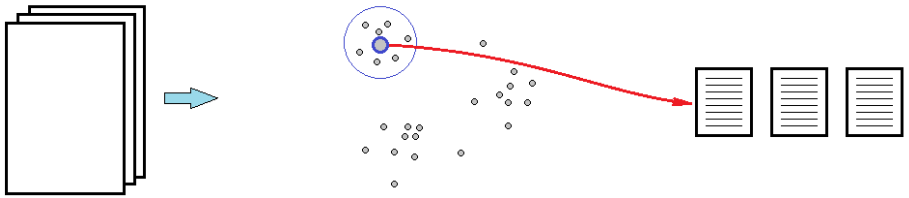


Поиск по кластерам, построенным по сегментам документов:

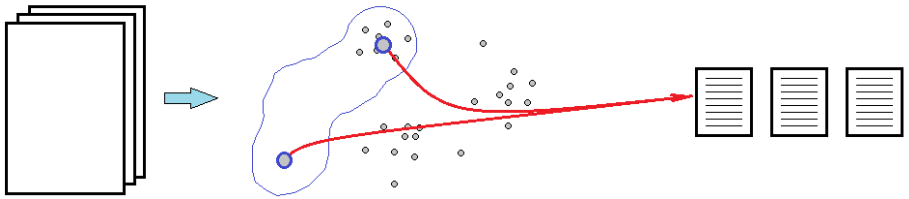


Поиск в пространстве семантических векторных представлений

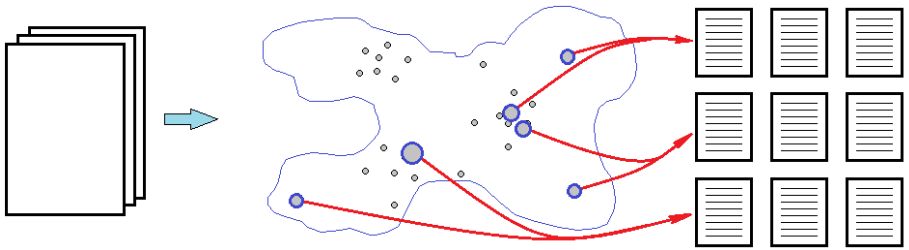
Поиск по части подборки или по отдельному документу:



Междисциплинарный поиск по смежным темам части подборки:



Междисциплинарный поиск по смежным темам всей подборки:



Задача суммаризации (аннотирования, реферирования) текста

Автоматическая суммаризация — *краткий* текст, построенный по одному или нескольким документам и *наиболее полно* передающий их содержание.

Полуавтоматическая — HAMS, human aided machine summarization

Основные типы задач суммаризации:

- *one-document* — на входе один документ $d \in D$
- *multi-document* — на входе набор документов $D' \subseteq D$
- ⊕ *topic* — на входе набор сегментов темы $p(d, s|t)$

Основные подходы к суммаризации:

- *extractive* — выбор некоторых предложений целиком
- *abstractive* — генерация текста на естественном языке

H.P.Luhn. The automatic creation of literature abstracts. 1958.
Juan-Manuel Torres-Moreno. Automatic Text Summarization. 2014.

Задача многокритериальной дискретной оптимизации

S_d — множество предложений документа d
подмножество $a \subset S_d$ — искомая суммаризация

Покрытие терминологии документа (lexicon coverage):

$$\mathrm{WCov}(a) = \mathrm{KL}\big(p(w|d)\|p(w|a)\big) \to \min_{a \subset S_d}$$

Покрытие тематики документа (topic coverage):

$$\mathrm{TCov}(a) = \mathrm{KL}\big(p(t|d)\|p(t|a)\big) \to \min_{a \subset S_d}$$

Избыточность суммаризации (redundancy):

$$\mathrm{Red}(a) = \sum_{s,s' \in a} B_{ss'} \to \min_{a \subset S_d}, \qquad B_{ss'} = \mathrm{sim}\big(p(w|s), p(w|s')\big),$$

где sim — одна из мер сходства: cos, JS, Jaccard и т.п.

Marina Litvak, Natalia Vanetik, Chunlei Liu, Lemin Xiao, Onur Savas. Improving Summarization Quality with Topic Modeling. 2015.

Сведение дискретной оптимизационной задачи к непрерывной

Метод релаксации: вместо $a \subset S_d$ ищем $\pi_s = p(s|a)$, где $s \in S_d$.
В релаксированной задаче:

$$p(w|a) = \sum_{s \in d} p(w|s)p(s|a) = \sum_{s \in d} \frac{n_{ws}}{n_s} \pi_s$$

$$p(t|a) = \sum_{s \in d} p(t|s)p(s|a) = \sum_{s \in d} \theta_{ts} \pi_s$$

Сумма трёх критериев $\mathrm{WCov}(a) + \tau_1 \mathrm{TCov}(a) + \tau_2 \mathrm{Red}(a)$:

$$\sum_{w \in d} n_{dw} \ln \sum_{s \in d} \frac{n_{ws}}{n_s} \pi_s + \tau_1 \sum_{t \in T} \theta_{td} \ln \sum_{s \in d} \theta_{ts} \pi_s - \tau_2 \sum_{s,s' \in d} B_{ss'} \pi_s \pi_{s'} \to \max_{\{\pi\}}$$

Максимизация покрытия — это максимизация правдоподобия!

Можно добавить регуляризатор разреживания:

$$R(\pi) = -\tau_3 \sum_{s \in S_d} \ln \pi_s \to \max_{\{\pi\}}$$

Последние достижения абстрактивной суммаризации

Abstract

We present a method to produce abstractive summaries of long documents that exceed several thousand words via neural abstractive summarization. We perform a simple extractive step before generating a summary, which is then used to condition the transformer language model on relevant information before being tasked with generating a summary. We show that this extractive step significantly improves summarization results. We also show that this approach produces more abstractive summaries compared to prior work that employs a copy mechanism while still achieving higher rouge scores. *Note: The abstract above was not written by the authors, it was generated by one of the models presented in this paper.*

S.Subramanian, R.Li, J.Pilault, C.Pal. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. 2019.

Автоматизация реферирования в стиле гибридного интеллекта

Научные статьи имеют ссылки и списки литературы.

Основной цикл в процессе реферирования подборки статей:

- пользователь выбирает суфлёра
- суфлёр выдаёт ранжированный список подсказок
- пользователь выбирает из него фразу
- фраза добавляется в редактируемый текст реферата

Виды суфлёров:

- какую статью упомянуть следующей
- какое предложение взять про данную статью
- какую цитату взять из данной статьи

Формирование выборок для обучения суфлёров

Идея из self-supervised learning: автоматическое формирование обучающей выборки из текстов научных статей

- *Подборка* x_i — статьи из списка литературы статьи i
- *Реферат* y_i — предложения из обзорных частей статьи i

Выделение обзорных частей статьи

- решение задачи сегментации
- отбор сегментов с широкой тематикой и частыми ссылками

Критерий качества суфлёра

- ROUGE@k — среднее по всем предложениям всех рефератов значение метрики качества суммаризации ROUGE для k-го предложения в ранжированном списке подсказок.

Методы сегментации TextTiling, TopicTiling

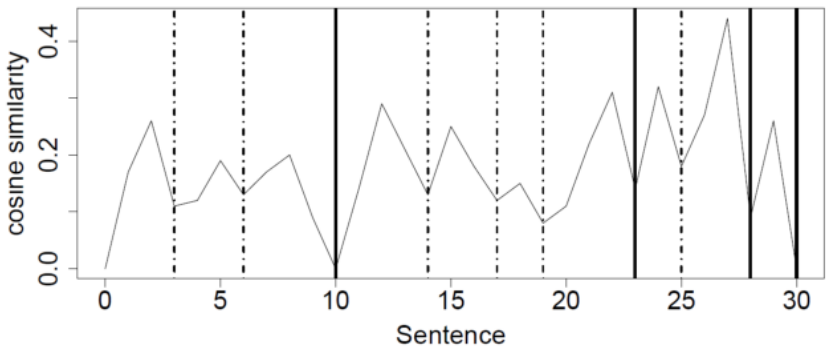
$(s_j)_{j=1}^{k_d}$ — последовательность предложений документа d

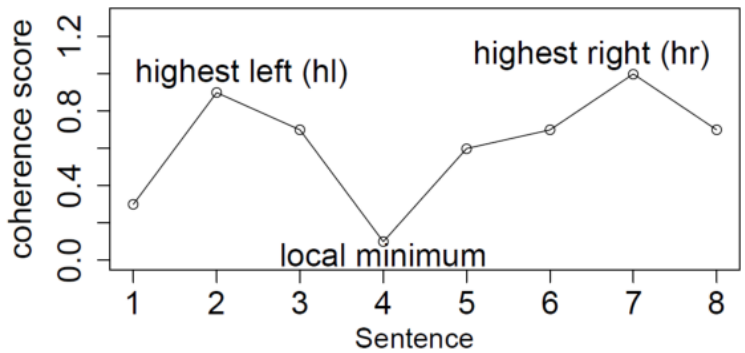
$v_s[t] = \frac{1}{|s|} \sum_{w \in s} v_w[t]$ — векторное представление предложения s

$v_w[t]$ — эмбединги слов (word2vec, тематические $p(t|d,w)$ и т.п.)

$c_j = \cos(v_{j-1}, v_j)$ — *coherence score*, оценка близости соседних предложений (чем глубже провал, тем чётче граница)

$d_j = \frac{1}{2}(\mathrm{hl}_j + \mathrm{hr}_j - 2c_j)$ — *depth score*, оценка глубины провала





Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Пример: кластеризация новостной темы на полярные мнения

Формализация понятие «полярное мнение»:

- набор тональных оценок ключевых объектов темы
- набор часто упоминаемых фактов
- набор умалчиваемых фактов
- набор семантических ролей ключевых объектов темы

Для оценивания качества классификации используется выборка новостей, размеченных экспертами.

Текущий уровень точности классификации около 70%.

Т.Садекова. Выделение мнений в тематических моделях новостных потоков. ВМК МГУ, 2018.

Д.Фельдман. Использование фактов для поиска мнений в новостях. МФТИ, 2018.

Задача оценивания когнитивной сложности текста

Основные предположения:

- *уровни языка*: фонетический, морфологический, лексический, синтаксический, дискурсивный
- на уровне i текст может быть представлен в виде последовательности *токенов* алфавита A_i
- *сложность текста* на уровне i — это доля токенов, имеющих аномально высокую частоту
- частота токена *аномально высокая*, если она превышает 95%-ю квантиль его частоты в референтном корпусе
- *референтный корпус* — тексты, не являющиеся сложными для выбранной читательской аудитории

M.Eremeev, K.Vorontsov. Lexical quantile-based text complexity measure. RANLP-2019.

От ближнего чтения (close reading) к дальнему (distant reading)

Концепция дальнего чтения Франко Моретти

«*Дальнее чтение* — не ограничение, а способ представления знаний: меньше элементов, чётче понимание их взаимосвязей, акцент на формах, отношениях, структурах, моделях»

Мантра Шнейдермана

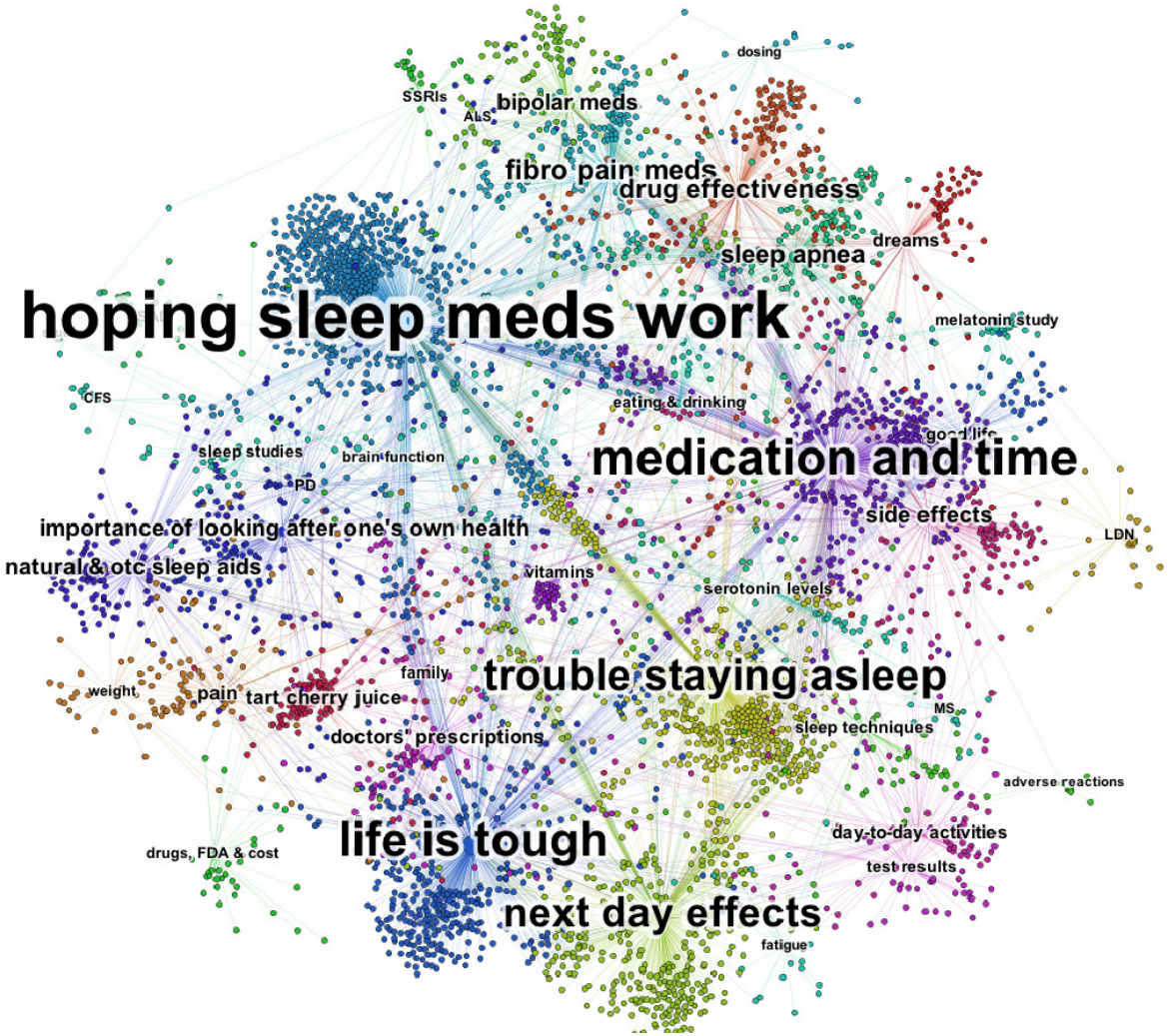
«Сначала крупный план, затем масшабирование и фильтрация, детали по требованию»

B.Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Visual Languages, 1996.

F.Moretti. Graphs, Maps, Trees: Abstract Models for a Literary History. 2005.

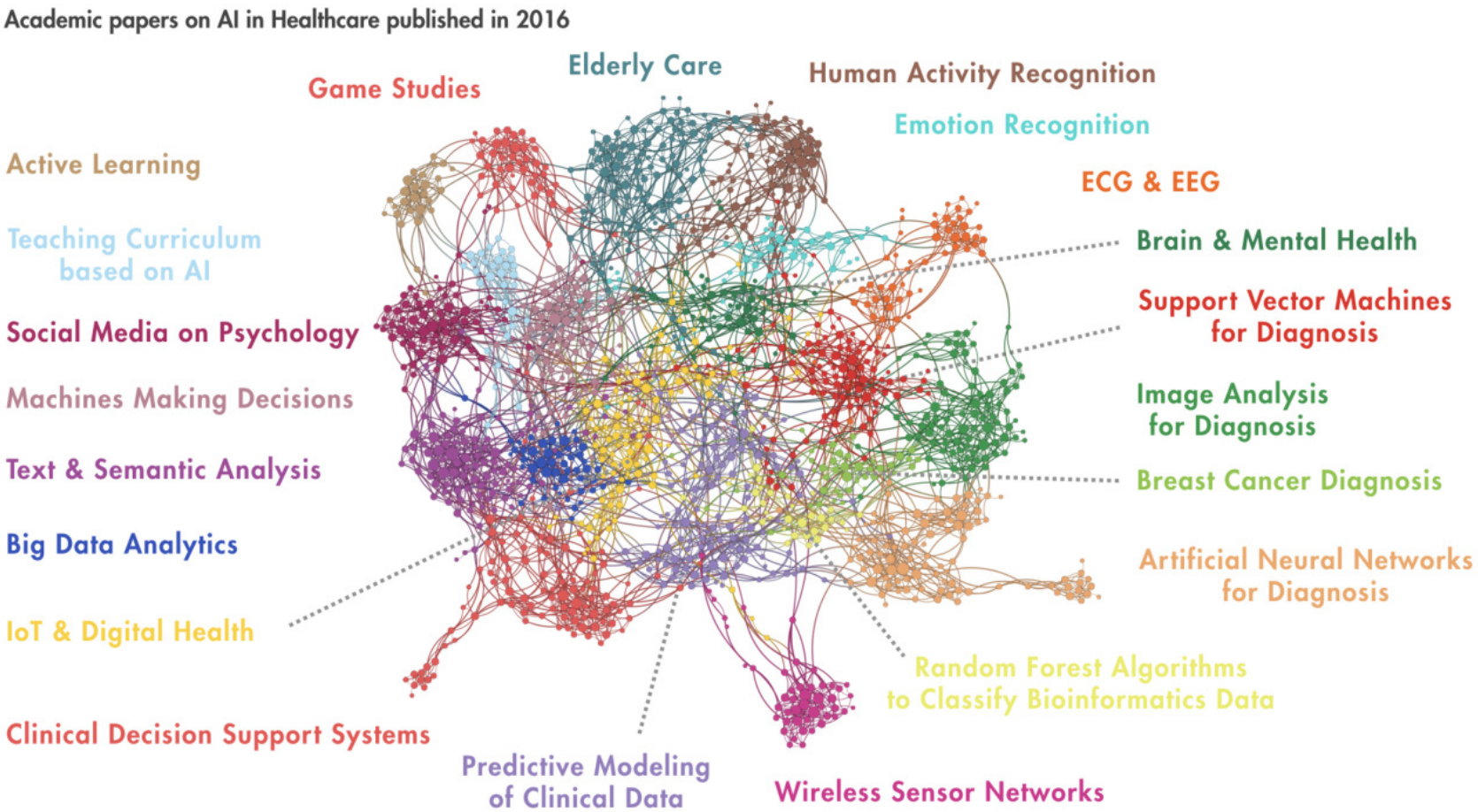
S.Janicke, G.Franzini, M.F.Cheema, G.Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis, 2015.

Пример: тематика обсуждений на www.PatientsLikeMe.com



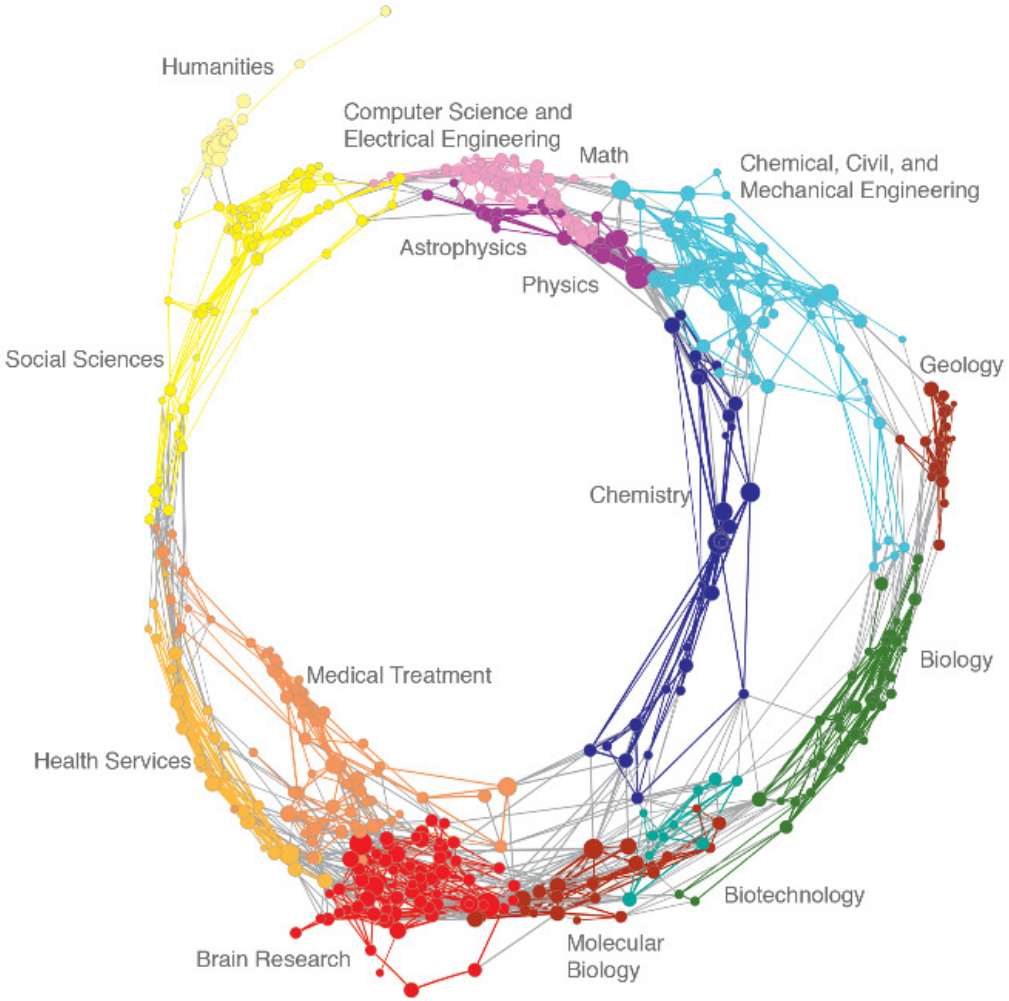
Chen A., Eichler G. Topic Modeling and Network Visualization to Explore Patient Experiences. 2013.

Ещё одна карта тематической кластерной структуры



C.Folgar, J.McCuan. The 3 most-cited studies in healthcare and AI. Quid, 2017.

Ещё один пример карты науки



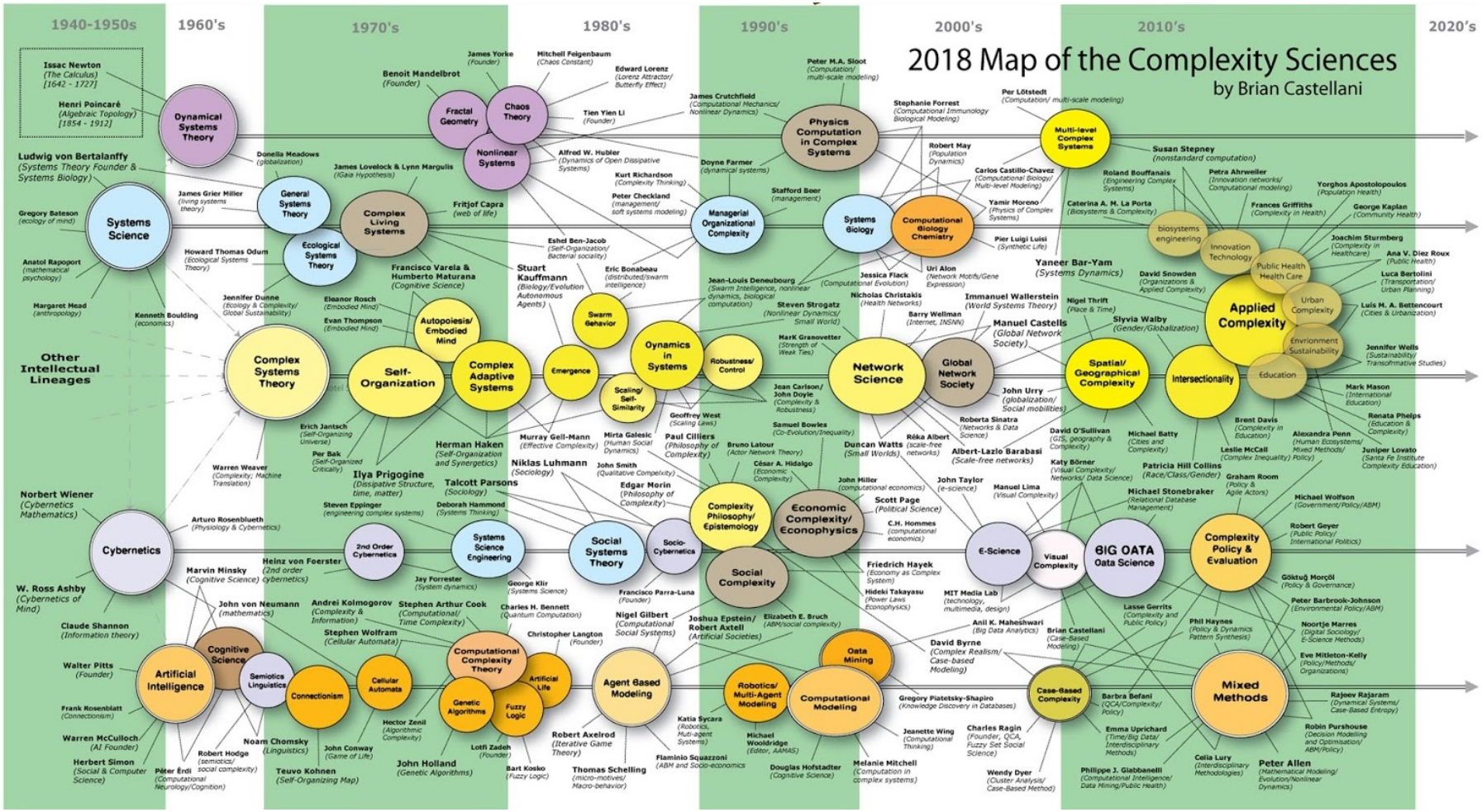
Важное наблюдение:
области знания самопроизвольно располагаются по кругу, значит,
их можно располагать и вдоль прямой линии.

Недостатки:

- оси не имеют интерпретации
- искажение сходства при двумерном проецировании

http://scimaps.org

Пример карты предметной области, построенной вручную



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

Визуализация тематического разведочного поиска (концепт)

- **Интерпретация осей:** время, темы, сложность, и т.д.
- **Иерархичность:** темы делятся на подтемы
- **Спектр тем:** гуманитарные $\rightarrow$ естественные $\rightarrow$ точные
- **Интерактивность:** реализация мантры Шнейдермана
- **Суммаризация:** карта любого масштаба заполнена текстом

Построение спектра тем. Постановка задачи

Тематический спектр — такая перестановка тем $t_1, \ldots, t_{|T|}$, что сумма расстояний между соседними темами минимальна:

$$\sum_{i=2}^{|T|} \rho(t_i, t_{i-1}) \to \min$$

Функция расстояния $\rho(t, t')$ между темами, примеры:

- Манхэттенское: $\rho(t, t') = \sum_{w \in W} \left| \phi_{wt} - \phi_{wt'} \right|$
- Хеллингера: $\rho^2(t, t') = \frac{1}{2} \sum_{w \in W} \left(\sqrt{\phi_{wt}} - \sqrt{\phi_{wt'}} \right)^2$
- Жаккара: $\rho(t, t') = 1 - \frac{|W_t \cap W_{t'}|}{|W_t \cup W_{t'}|}$, $W_t = \left\{ w : \phi_{wt} > \frac{1}{|W|} \right\}$

Построение спектра тем — это задача коммивояжёра

Задача TSP (traveling salesman problem)

Найти путь минимальной суммарной стоимости, соединяющий T городов так, чтобы в каждом городе побывать один раз.

Алгоритм Лина–Кернигана в реализации Хельсгауна — лучший для решения задачи TSP, по данным *Encyclopedia of operations research* на 2013 год.
Вычислительная сложность $T^{2.2}$.

Другие алгоритмы оказались не только медленнее, но и хуже по качеству тематических спектров.

Keld Helsgaun. An effective implementation of the Lin–Kernighan traveling salesman heuristic. EJOR, 2000.

Дмитрий Федоряка. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация. МФТИ, 2017.

Резюме

Технологии «Мастерской знаний» — машинное обучение, численная оптимизация, математическая статистика

- семантические векторные представления текста
- иерархическое тематическое моделирование
- тематическая сегментация текста
- обучение ранжированию для различных стратегий поиска
- обучение ранжированию фраз для суммаризации
- обучение ранжированию документов в порядке чтения
- выявление понятий, фактов, идей, мнений, выводов
- оценивание относительной когнитивной сложности текста
- оптимальное линейное ранжирование тем
- оптимальное размещение текстовых данных на картах

Литература

- *К.В.Воронцов.* Обзор вероятностных тематических моделей. 2017. – NEW!
  http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf
- *К.В.Воронцов.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
- *K.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.
- *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.
- *N.Chirkova, K.Vorontsov.* Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.
- *A.Ianina, K.Vorontsov.* Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.
- *A.Potapenko, A.Popov, K.Vorontsov.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.
- *V.Alekseev, V.Bulatov, K.Vorontsov.* Intra-Text Coherence as a Measure of Topic Models Interpretability. Dialogue, 2018.
- *A.Belyy, M.Seleznova, A.Sholokhov, K.Vorontsov.* Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue, 2018.
- *N.Skachkov, K.Vorontsov.* Improving topic models with segmental structure of texts. Dialogue, 2018.