Вероятностные тематические модели Лекция 4. Тематические иерархии и разведочный информационный поиск

К. В. Воронцов
vokov@forecsys.ru

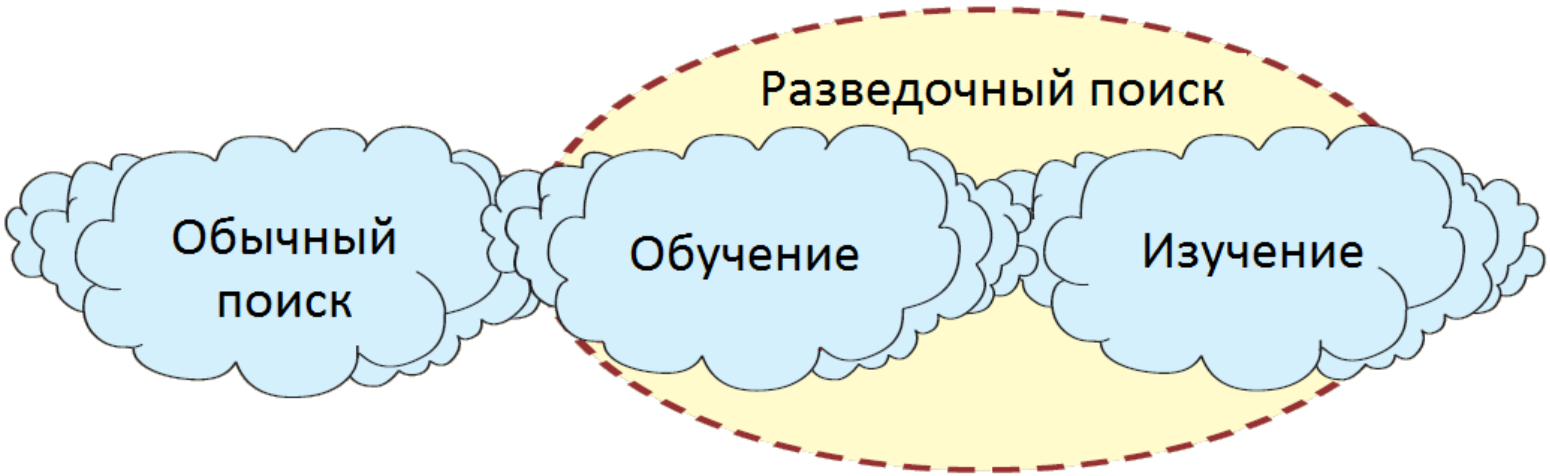
Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • 7 марта 2019

Содержание

Концепция разведочного поиска (exploratory search)

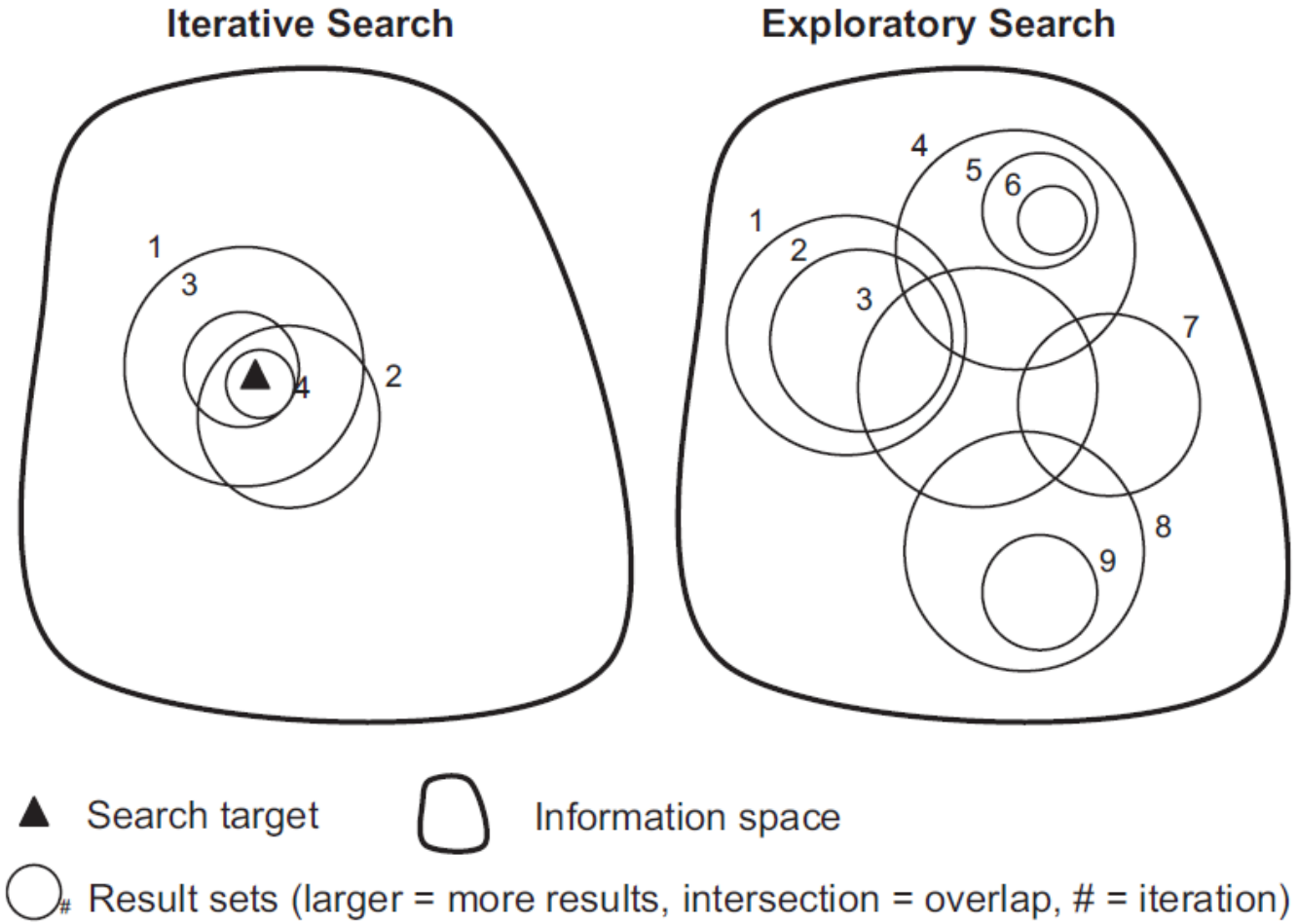
- пользователь может не знать ключевых терминов
- запросом может быть текст произвольной длины
- информационная потребность — систематизация знаний



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.

От итераций «query-browse-refine» к разведочному поиску



R.W.White, R.A.Roth. Exploratory Search: beyond the Query-Response paradigm. San Rafael, CA: Morgan and Claypool, 2009.

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

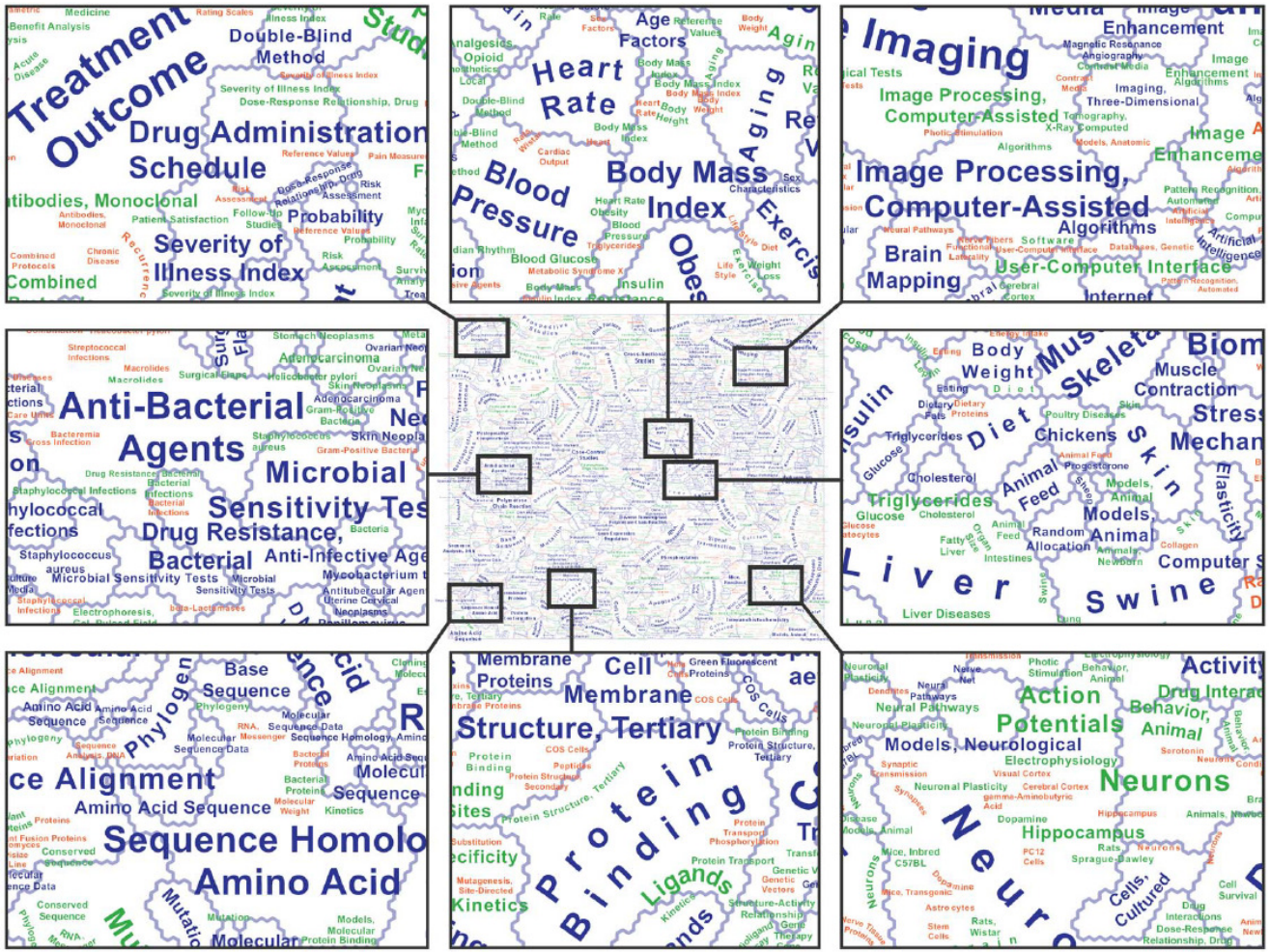
«Сначала крупный план, затем масшабирование и фильтрация, детали по требованию»

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.

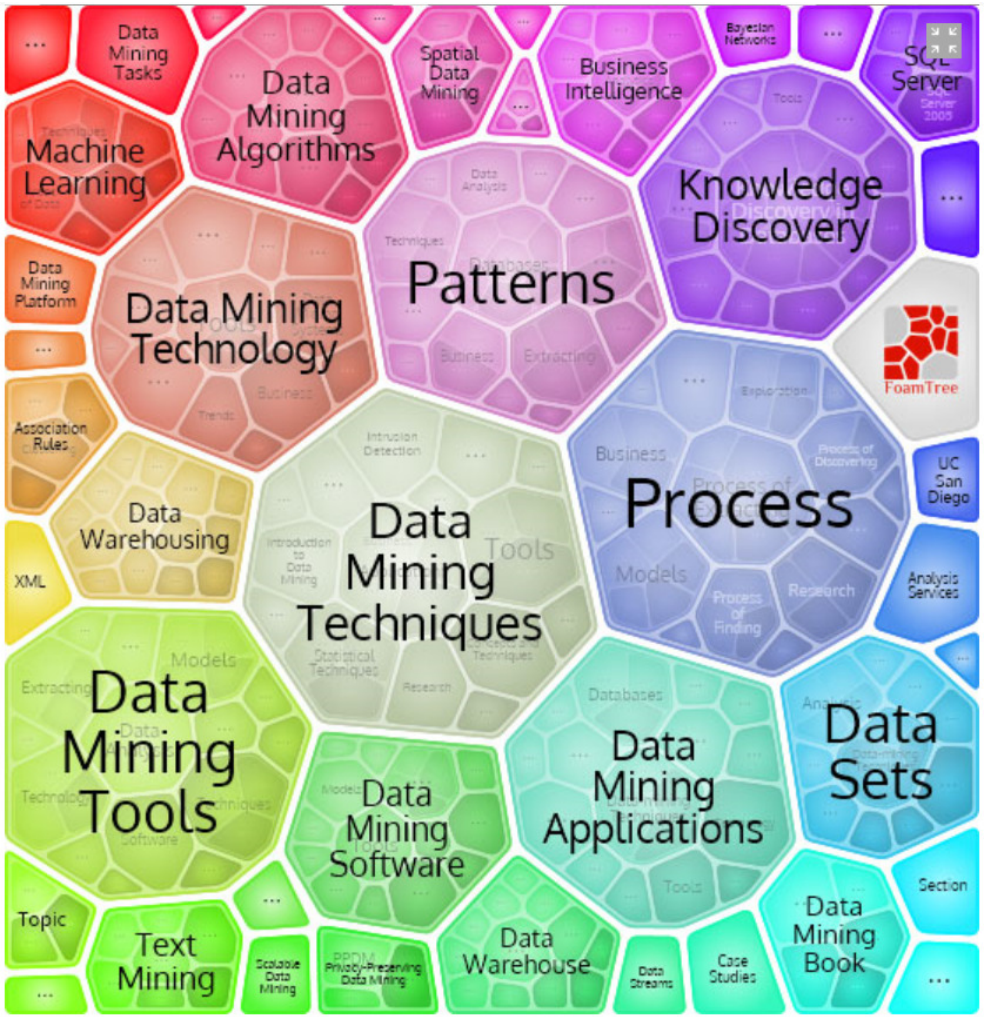
S.Janicke, G.Franzini, M.F.Cheema, G.Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis, 2015.

Пример карты медицинских знаний



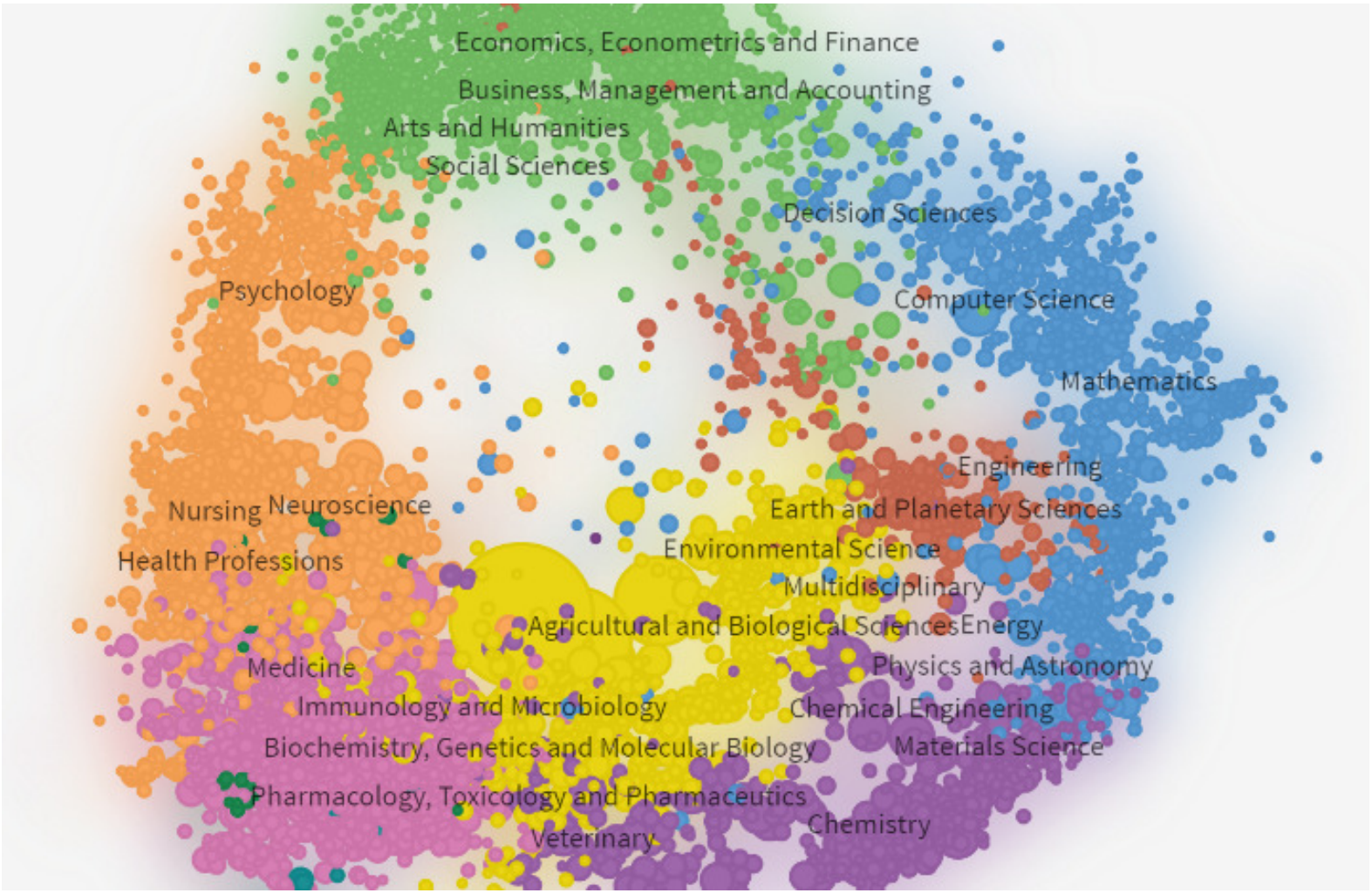
Skupin, Biberstine, Borner. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS ONE, 2013.

Пример иерархической карты области *Data Mining*



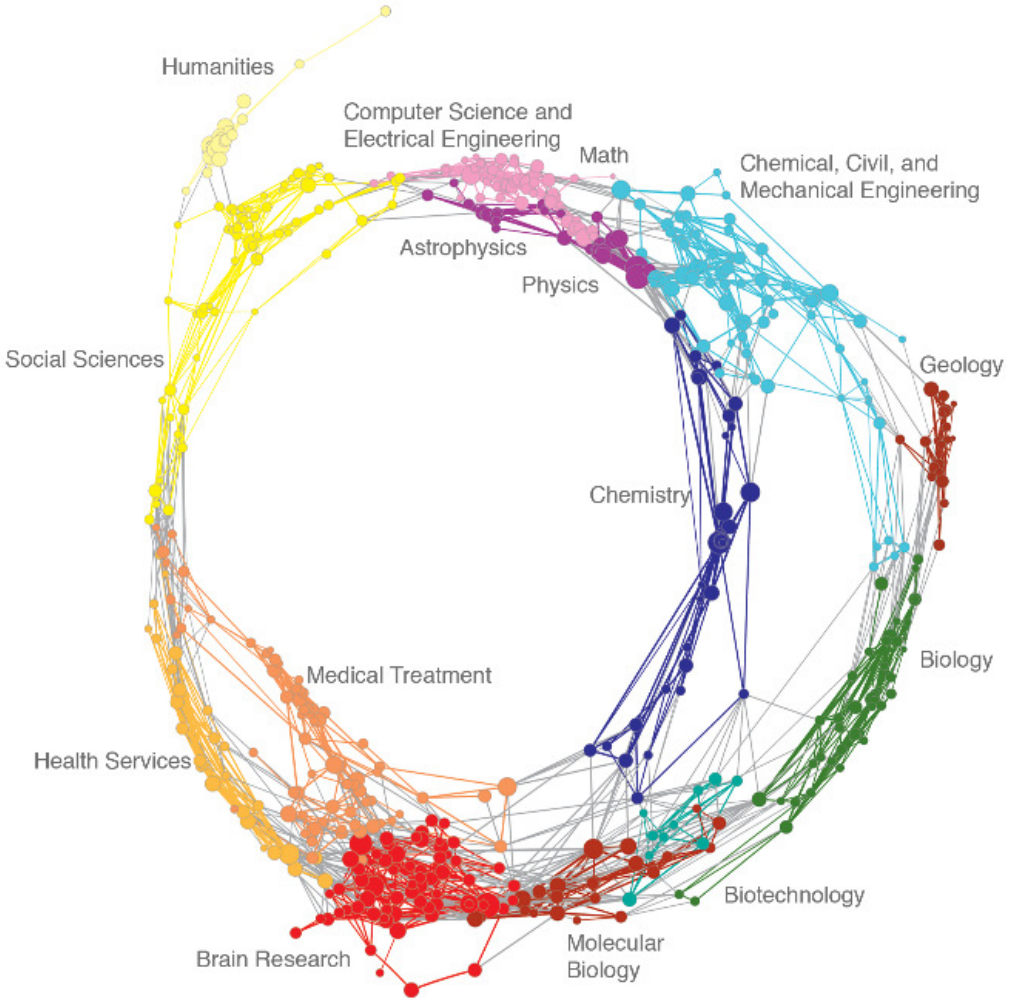
FoamTree: https://carrotsearch.com/foamtree

Пример карты науки



http://onlinelibrary.wiley.com/browse/subjects

Ещё один пример карты науки



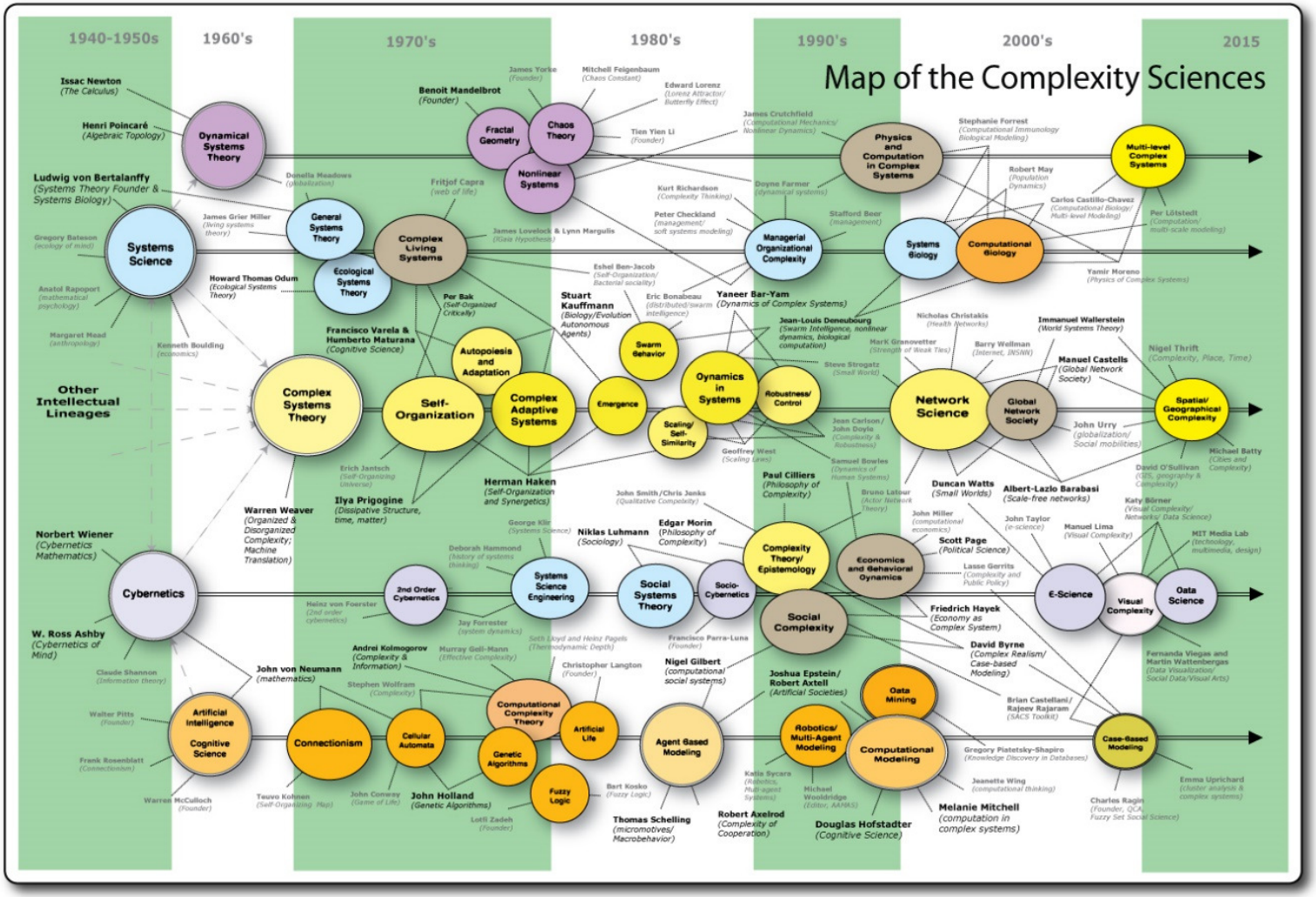
Важное наблюдение:
области знания самопроизвольно располагаются по кругу, значит,
их можно располагать и вдоль прямой линии.

Недостатки:

- оси не имеют интерпретации
- искажение сходства при двумерном проецировании

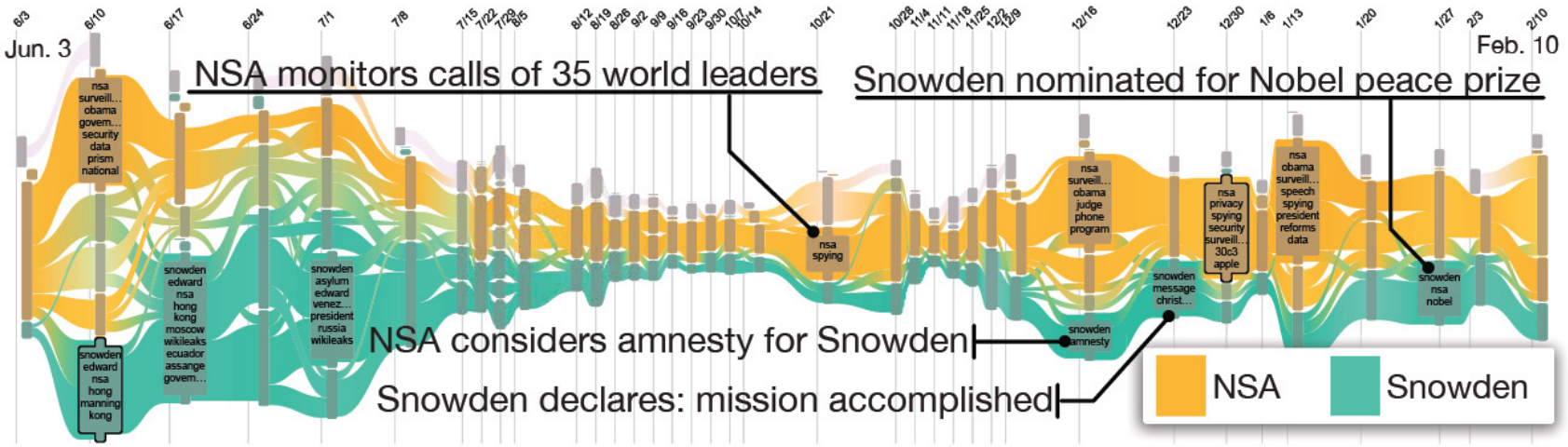
http://scimaps.org

Пример карты предметной области, построенной вручную



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

Динамика тем: эволюция предметной области



Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

Стратегия визуализации в системах TextFlow и RoseRiver:

- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- получает сгенерированный отчёт с инфографикой.

Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. 2014.

Визуализация тематического разведочного поиска (концепт)

- **Интерпретация осей:** время–темы или сложность–темы
- **Иерархичность:** темы делятся на подтемы
- **Спектр тем:** гуманитарные $\rightarrow$ естественные $\rightarrow$ точные
- **Интерактивность:** реализация мантры Шнейдермана
- **Суммаризация:** на карте любого масштаба много текста

http://textvis.lnu.se

Интерактивный обзор 430 средств визуализации текстов

Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // JMLDA, 2015.

Возможный сценарий разведочного поиска

Поисковый запрос:

- документ любой длины или даже коллекция документов

Цели поиска:

- к каким темам относится мой запрос?
- какова тематическая структура предметной области?
- в каком порядке читать, чтобы лучше разобраться в теме?
- какие области являются смежными?
- какие фрагменты текста наиболее релевантны теме?

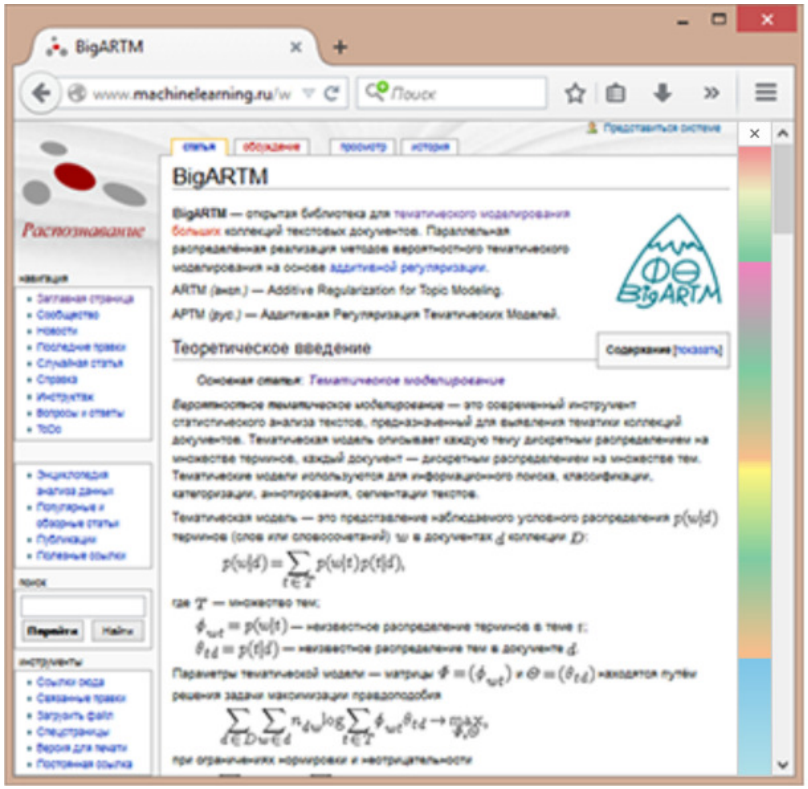
Сценарий поиска:

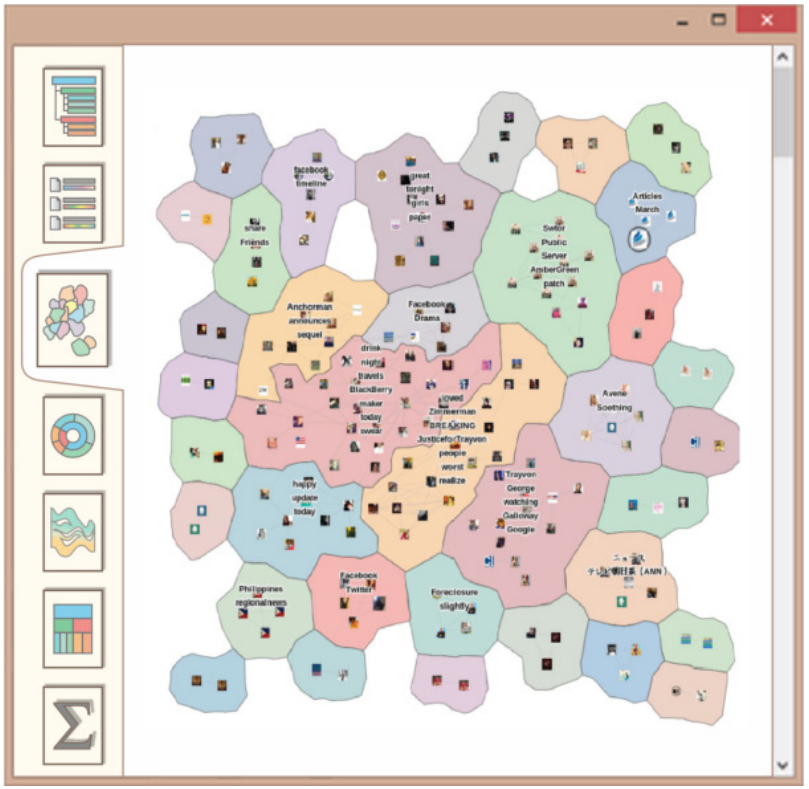
1. имея текст под рукой,
2. получить карту содержащихся в нём тем-подтем
3. и карту предметной области в целом

Документ-запрос и результат тематического поиска (концепт)

Тематическая сегментация: структура документа-запроса

Тематическая карта: кластеризация релевантных документов





Технологические элементы разведочного поиска

По всем технологиям имеются готовые решения:

1. интернет-краулинг
2. фильтрация контента
3. тематическое моделирование
4. инвертированный индекс
5. ранжирование
6. визуализация
7. персонализация

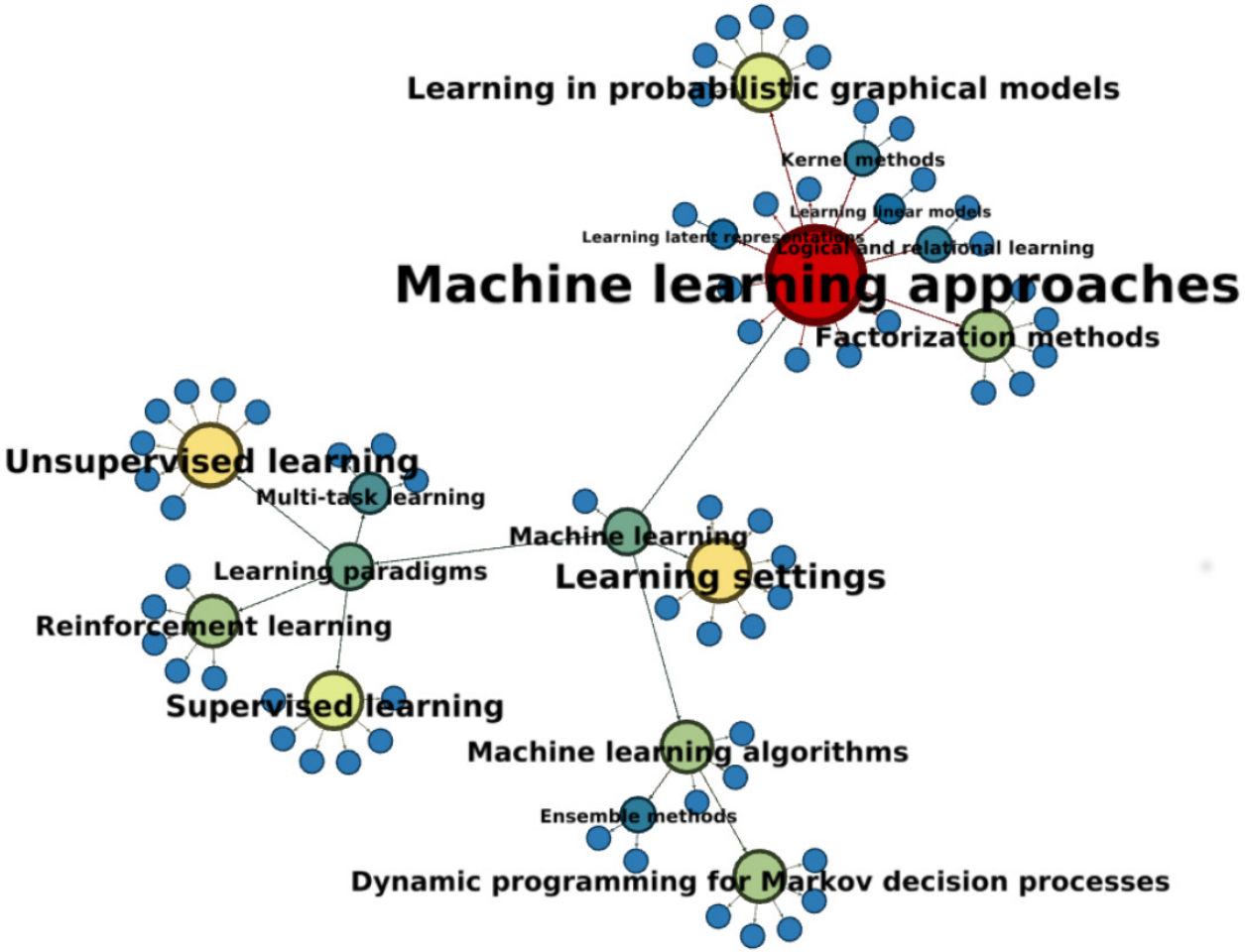
Тематическая модель — ключевое звено разведочного поиска

ARTM позволяет строить тематические модели с заданным набором требуемых свойств

Тематическая модель для разведочного поиска должна быть...

1. **Интерпретируемая**: каждая тема понятна людям
2. **Иерархическая**: выявление иерархических связей тем
3. **Динамическая**: прослеживание истории развития тем
4. **Мультимодальная**: авторы, связи, теги, пользователи,...
5. **Мультиграммная**: термины-словосочетания неразрывны
6. **Мультиязычная**: для кросс- и много-языкового поиска
7. **Сегментирующая**: выделение тем внутри документа
8. **Обучаемая** по оценкам асессоров и логам пользователей
9. **Определяющая число тем** автоматически
10. **Создающая и именующая новые темы** автоматически
11. **Онлайновая**: обрабатывающая коллекцию за 1 проход
12. **Параллельная, распределённая** для больших коллекций

Пример древовидной тематической иерархии



Georgeta Bordea. Domain adaptive extraction of topical hierarchies for Expertise Mining. 2013.

Пример тематической иерархии

Тексты научно-просветительского ресурса Postnauka.ru:
2976 документов, 43196 слов, 1799 тэгов

Chirkova N.A., Vorontsov K.V. Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.

Belyy A.V., Seleznova M.S., Sholokhov A.K., Vorontsov K.V. Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue 2018.

Иерархические тематические модели

- структура иерархии: дерево / многодольный граф
- направление: снизу вверх / сверху вниз / одновременно
- наращивание: поверхинное / послойное

Открытые проблемы:

- "Despite recent activity in the field of HPTMs, determining the hierarchical model that best fits a given data set, in terms of the structure and size of the learned hierarchy, still remains a challenging task and an open issue."
- "The evaluation of hierarchical PTMs is also an open issue."

Zavitsanos E., Paliouras G., Vouros G. A. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes. 2011.

Регуляризатор Ф: родительские темы как псевдо-документы

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:

$$\sum_{t\in T} n_t \,\mathrm{KL}_w\Big(p(w|t) \,\Big\|\, \sum_{s\in S} p(w|s)p(s|t)\Big) \;\to\; \min_{\Phi,\tilde{\Psi}},$$

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

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

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

Родительские темы t — «документы» с частотами термов n_{wt}.

Регуляризатор Θ: родительские темы как модальность

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:

$$\sum_{d\in D} n_d \,\mathrm{KL}_t\Big(p(t|d) \,\Big\|\, \sum_{s\in S} p(t|s)p(s|d)\Big) \;\to\; \min_{\Theta,\Psi},$$

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

Родительская $\Theta^p \approx \Psi\Theta$, отсюда регуляризатор матрицы Θ:

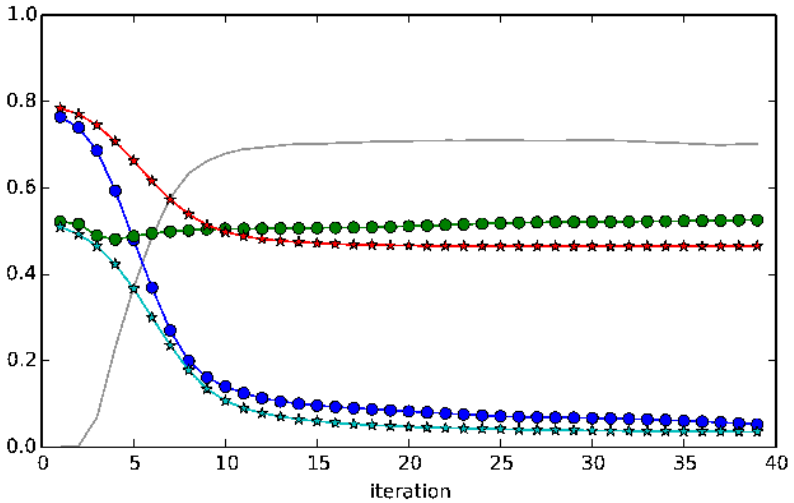
$$R(\Theta,\Psi) = \tau \sum_{d\in D}\sum_{t\in T} n_{td} \ln \sum_{s\in S} \psi_{ts}\theta_{sd} \;\to\; \max.$$

Родительские темы t — модальность с частотами термов n_{td}.

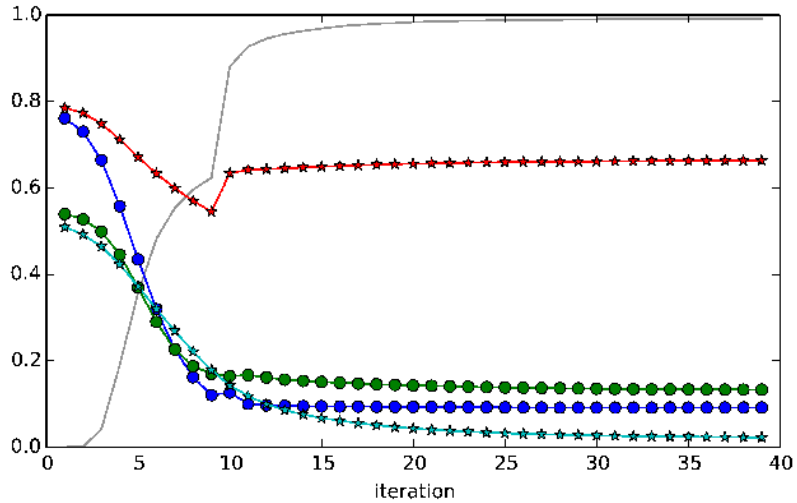
Эксперимент на коллекции ММРО-ИОИ

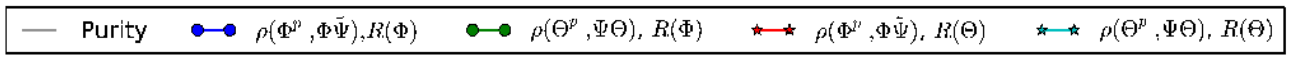
Среднее расстояние Хеллингера $\rho(\Phi^p, \Phi\tilde{\Psi})$ и $\rho(\Theta^p, \Psi\Theta)$ для регуляризаторов Φ и Θ при переходе между уровнями $1 \to 2$:

Разреживание Φ с 1-й итерации



Разреживание Φ с 10-й итерации





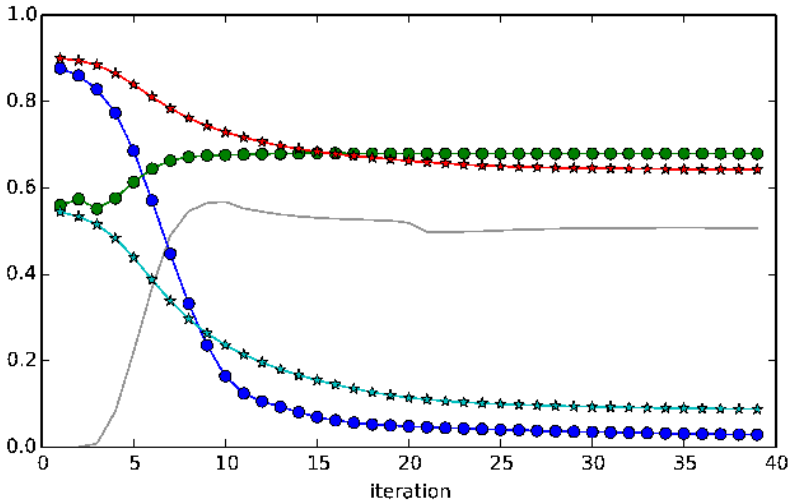
Вывод. Регуляризатор Θ плохо приближает Φ^p.

Chirkova N. A., Vorontsov K. V. Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.

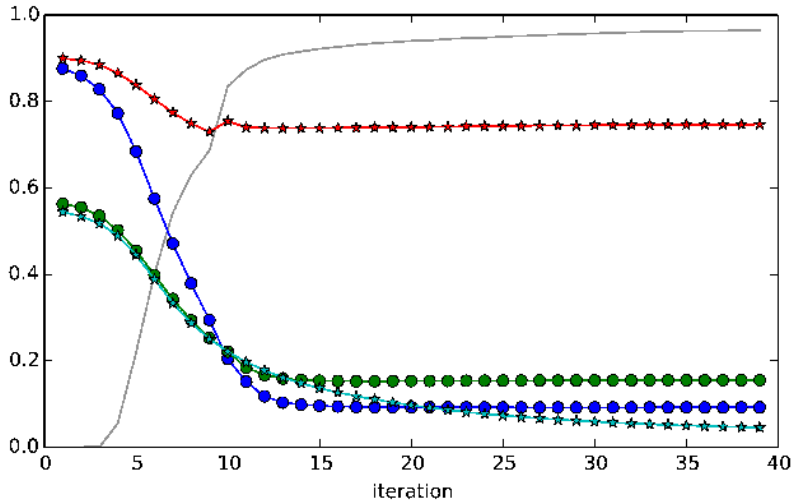
Эксперимент на коллекции ММРО-ИОИ

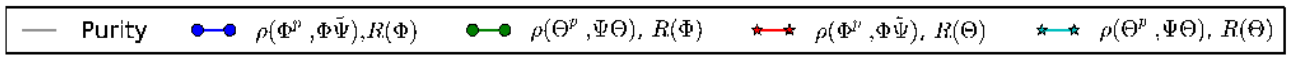
Среднее расстояние Хеллингера $\rho(\Phi^p, \Phi\tilde{\Psi})$ и $\rho(\Theta^p, \Psi\Theta)$ для регуляризаторов Φ и Θ при переходе между уровнями $2 \to 3$:

Разреживание Φ с 1-й итерации



Разреживание Φ с 10-й итерации





Вывод. Регуляризатор Θ плохо приближает Φ^p.

Chirkova N. A., Vorontsov K. V. Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.

Выводы

- Регуляризатор Φ приближает $\Phi^p \approx \Phi\tilde{\Psi}$ и $\Theta^p \approx \Psi\Theta$
- Регуляризатор Θ приближает только $\Theta^p \approx \Psi\Theta$
- Сильное разреживание $\psi_{ts} \in \{0, 1\}$ даёт иерархию–дерево
- Нельзя допускать вырождения $\psi_{ts} = p(t|s) \equiv 0$

Дальнейшие задачи:

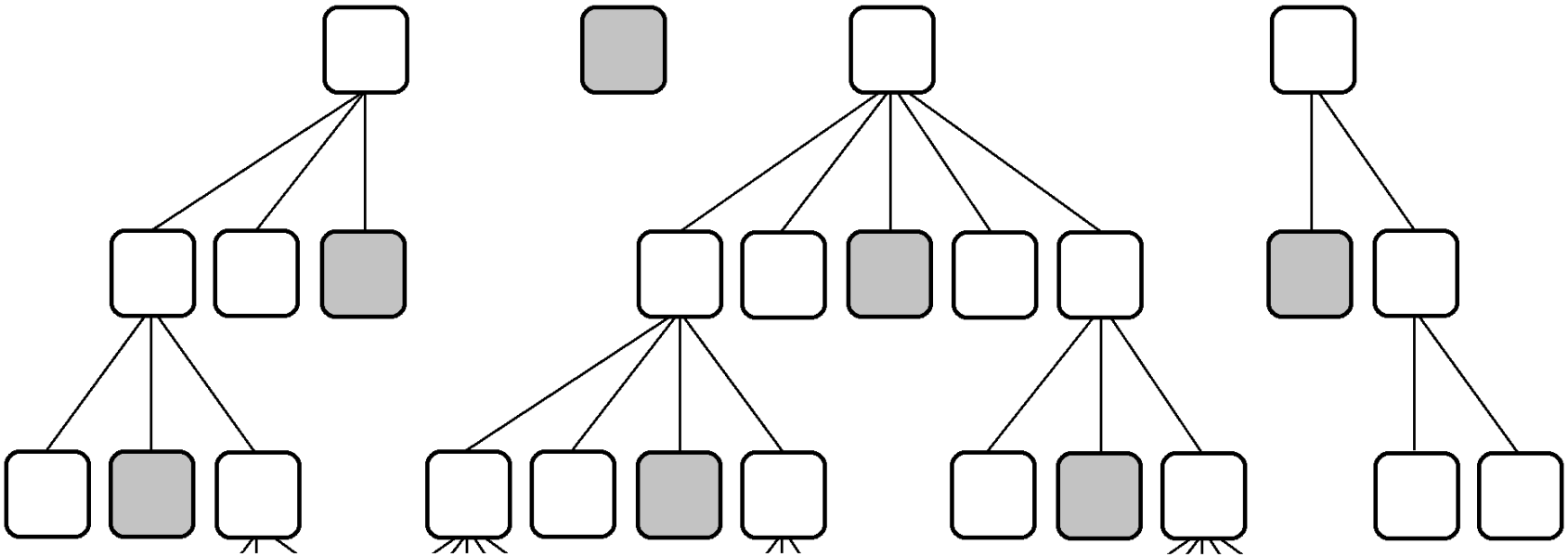
- Согласованная регуляризация: $\tilde{\psi}_{st} p(t) = \psi_{ts} p(s)$

$$\tau_1 \sum_{t,w} n_{wt} \ln \sum_s \phi_{ws}\psi_{ts}\tfrac{n_s}{n_t} + \tau_2 \sum_{d,t} n_{td} \ln \sum_s \psi_{ts}\theta_{sd} \;\to\; \max_{\Phi,\Psi,\Theta}$$

- Иерархии с темами различной глубины:
  — наращивание уровня для подмножества $T' \subseteq T$
  — критерий неоднородности темы для включения её в T'

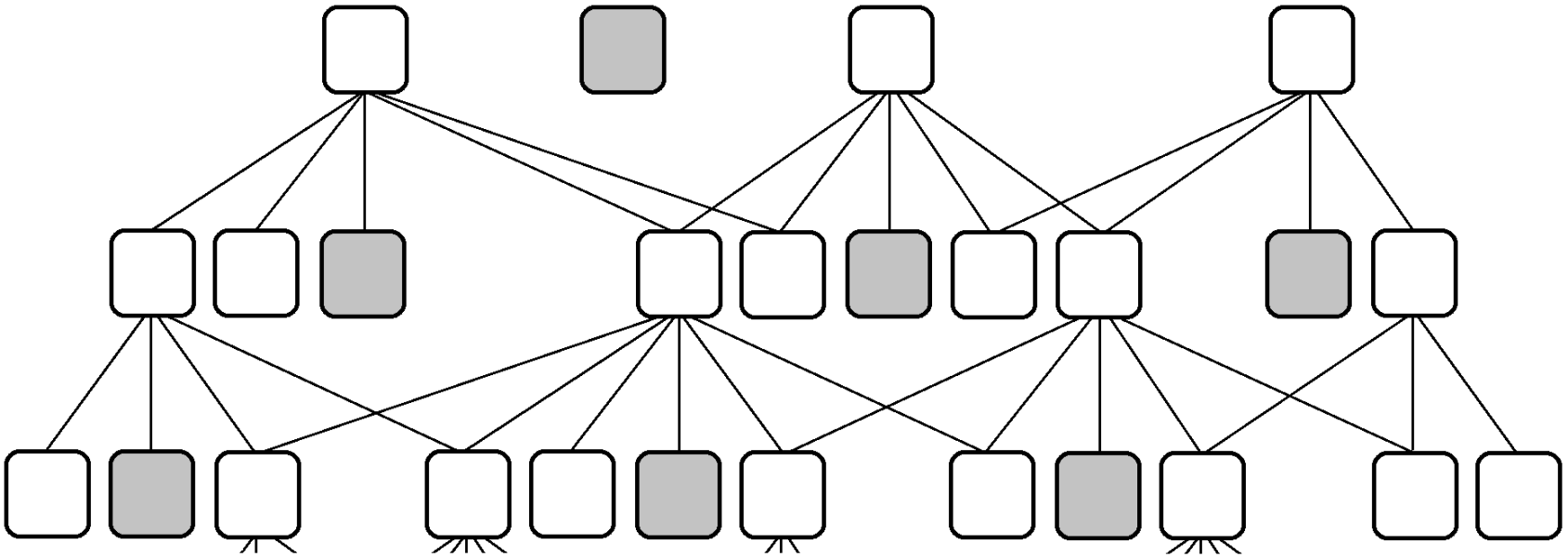
Иерархии с темами различной глубины

- На каждом уровне расщепляются не все темы (допускается вырожденность: $p(s|t) \equiv 0$ для некоторых t)
- Расщепляемая тема может иметь дочернюю фоновую, в которой собирается общая лексика родительской темы
- При максимальном разреживании $p(t|s) \in \{0, 1\}$ иерархия является деревом (корень не показан)



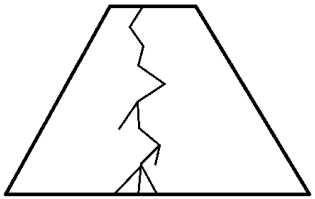
Иерархии с темами различной глубины

- На каждом уровне расщепляются не все темы (допускается вырожденность: $p(s|t) \equiv 0$ для некоторых t)
- Расщепляемая тема может иметь дочернюю фоновую, в которой собирается общая лексика родительской темы
- При умеренном разреживании $p(t|s)$ у вершины может быть несколько родителей (корень не показан)

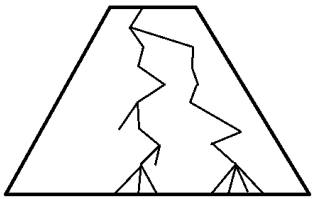


Иерархии с темами различной глубины

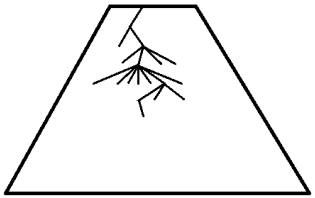
След документа в тематической иерархии определяет степень его специализации, назначение, аудиторию



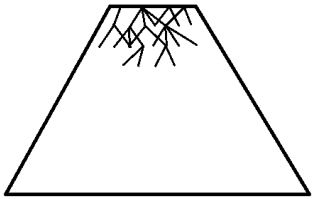
узко специализированный,
для профессионалов



междисциплинарное исследование,
для профессионалов



обзорный,
для ознакомления с предметной областью



популярный или энциклопедиченский,
для расширения кругозора

Способы оценивания качества тематических иерархий

- *Перплексия* или правдоподобие: приводит ли постепенное дробление тем к более точному разложению
- *Полезность*: сколько шагов делает пользователь, чтобы найти документ по иерархии
- *Когерентность*: как часто слова темы и её подтемы совместно встречаются рядом в тексте
- *Метод интрузий*: правильно ли асессоры определяют чужую тему, внедрённую в список дочерних тем
- *Сравнение с «золотым стандартом»*: насколько иерархия похожа на имеющуюся категоризацию документов

Что такое «спектр тем» и зачем он нужен

Визуализация иерархии тем во времени (концепт):

- Интерпретируемые оси «время–темы»
- Близкие темы должны находиться рядом
- *Тематический спектр* — одномерная линейная проекция (например, науки: гуманитарные $\rightarrow$ естественные $\rightarrow$ точные)

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.

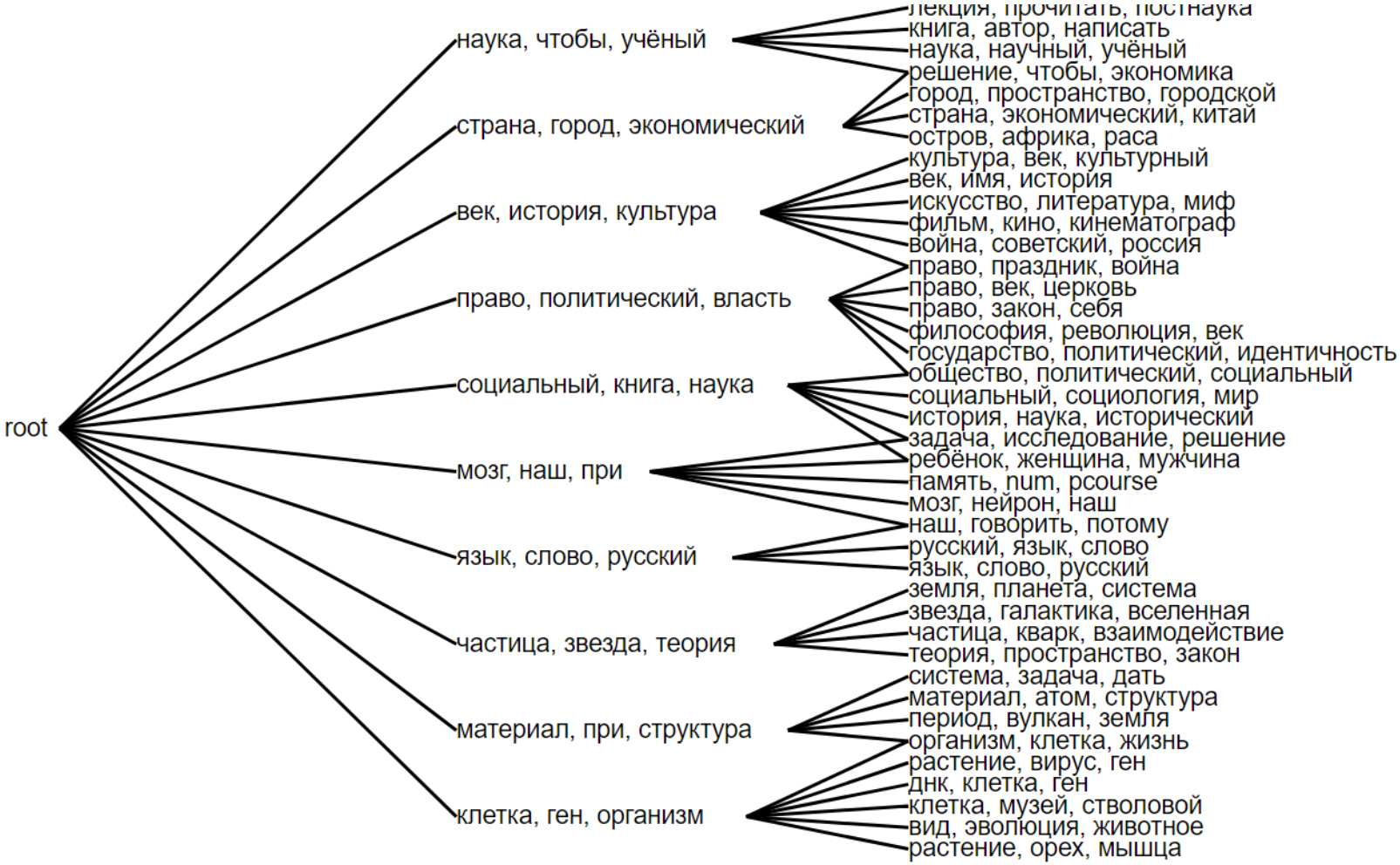
Пример спектра (коллекция postnauka.ru)

1. остров, земля, период, там, территория, океан, где, более, вид, найти, вулкан, находиться, южный
2. растение, япония, раса, при, более, чем, например, исследование, вид, страна, население
3. вид, эволюция, самец, мозг, самка, животное, отбор, ген, более, птица, наш, между, чтобы, чем, друг
4. мозг, нейрон, при, заболевание, наш, пациент, состояние, система, болезнь, сон, исследование
5. клетка, музей, стволовой, ткань, организм, чтобы, опухоль, система, использовать, технология
6. клетка, ген, днк, организм, молекула, геном, белок, белка, бактерия, система, процесс, жизнь
7. система, материал, задача, структура, метод, компьютер, дать, при, химический, область, химия
8. квантовый, свет, волна, атом, информация, фотон, сигнал, использовать, два, при, частота, состояние
9. частица, энергия, кварк, взаимодействие, магнитный, электрон, масса, физика, бозон, протон, модель
10. звезда, галактика, земля, планета, вселенная, дыра, чёрный, объект, солнце, масса, наш, система
11. теория, пространство, вселенная, закон, физика, математический, уравнение, число, два, мир, система
12. наш, сеть, информация, дать, объект, культура, задача, например, образ, память, слово, разный
13. язык, слово, русский, например, говорить, словарь, речь, разный, языковой, текст, два, лингвист
14. наука, учёный, научный, потому, чтобы, лекция, хороший, университет, сейчас, наш, заниматься
15. экономический, экономика, страна, чтобы, более, рынок, компания, цена, решение, деньга, работа, чем
16. страна, война, государство, политический, россия, советский, власть, политика, германия, стать
17. ребёнок, женщина, мужчина, жизнь, культура, общество, себя, семья, социальный, советский, женский
18. город, пространство, социальный, городской, общество, место, культурный, жизнь, более, современный
19. исследование, социальный, поведение, группа, решение, and, the, теория, проблема, наука
20. социальный, социология, мир, теория, объект, социологический, действие, событие, социолог, наука
21. политический, философия, идея, наука, свобода, понятие, революция, история, философ, век, себя
22. право, власть, закон, король, век, римский, бог, себя, церковь, правовой, политический, суд, два
23. век, история, русский, исторический, имя, традиция, христианский, культура, историк, текст, уже
24. себя, искусство, литература, говорить, потому, мир, сам, миф, жизнь, слово, текст, роман, век
25. книга, фильм, автор, кино, pcourse, num, читатель, посвятить, тема, история, исследование, работа

Пример спектра (коллекция lenta.ru)

1. спортсмен, допинг, олимпиада, рио, де, россия, проба, жанейро, wada, олимпийский_игра, соревнование
2. команда, матч, счёт, клуб, победа, чемпионат, турнир, минута, футболист, встреча, летний, футбол
3. евро, евровидение, страна, россия, конкурс, франция, болельщик, англия, украина, футбол, певец
4. пройти, мероприятие, россия, акция, фестиваль, москва, фильм, участник, картина, театр, музей
5. фильм, сериал, продукт, актёр, компания, продукция, процент, россия, книга, товар, картина, сезон
6. россия, москва, турист, процент, россиянин, страна, отель, рейс, путешественник, город, тысяча
7. процент, доллар, рубль, нефть, цена, россия, баррель, страна, уровень, вырасти, рынок, рост
8. компания, миллиард_рубль, процент, миллиард_доллар, россия, сумма, миллион_доллар, банк, банка
9. закон, законопроект, документ, реклама, использование, деятельность, поправка, внести, организация
10. россия, страна, керченский_пролив, российский, боинг, работа, чайка, ряд, гражданин, аэропорт
11. партия, кандидат, журналист, праймериза, выбор, единый_россия, госдума, выборы
12. россия, украина, крым, решение, киев, депутат, вопрос, отношение, страна, мнение, право, москва
13. россия, страна, турция, сша, ес, евросоюз, москва, санкция, отношение, украина, вопрос, государство
14. россия, сирия, исламский_государство, сша, нато, иго, запретить, террорист, страна, боевик
15. ракета, путин, россия, запуск, глава_государство, союз, спутник, президент
16. учёный, клетка, исследование, исследователь, ген, университет, оказаться, процент, помощь, организм
17. земля, животное, учёный, животный, тысяча, звезда, планета, обнаружить, кошка, территория, жизнь
18. самолёт, километр, машина, борт, пассажир, вертолёт, погибнуть, лайнер, пилот, час, район, яхта
19. полицейский, полиция, мужчина, задержать, автомобиль, улица, москва, пострадать, life
20. статья, убийство, задержать, суд, отношение, ук_рф, подозревать, следствие, обвинять, трамп, часть
21. ребёнок, женщина, мужчина, летний, дом, сын, семья, мальчик, жена, полиция, дочь, школа, врач
22. видео, youtube, ролик, фото, фотография, канал, снимка, auto, instagram, девушка, страница, группа
23. facebook, пользователь, интернет, страница, twitter, пост, написать, соцсеть, вконтакте, аккаунт
24. устройство, смартфон, компания, мотоциклист, игра, байкер, видео, миллион_доллар, робот, молодая
25. бренд, модель, компания, обувь, основать, одежда, релиз, коллекция, редакция, часы, поступить

Иерархический спектр (коллекция postnauka.ru)



Иерархический спектр (коллекция lenta.ru)

Две коллекции новостей про технологии

Habrahabr.ru

175 143 статей на русском
10 552 слов (униграмм)
742 000 биграмм
524 авторов статей
10 000 авторов комментариев
2546 тегов
123 хаба (категории)

TechCrunch.com

759 324 статей на английском
11 523 слов (униграмм)
1.2 млн. биграмм
605 авторов
184 категорий

Предобработка текстов

- отброшены 5% наиболее частотных слов (общая лексика)
- удаление пунктуации
- нижний регистр, ё→е
- лемматизация pymorphy2

Поиск тематически близких документов

$\theta_{tq} = p(t|q)$ — тематический вектор запроса q
$\theta_{td} = p(t|d)$ — тематические векторы документов $d \in D$

Косинусная мера близости документа d и запроса q:

$$\text{sim}(q,d) = \frac{\sum_t \theta_{tq}\theta_{td}}{\left(\sum_t \theta_{tq}^2\right)^{1/2}\left(\sum_t \theta_{td}^2\right)^{1/2}}.$$

Ранжируем документы коллекции $d \in D$ по убыванию $\text{sim}(q,d)$
Выдача тематического поиска — k первых документов.

Реализация: *инвертированный индекс* для быстрого поиска документов d по каждой из тем t запроса

A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Методика оценивания качества разведочного поиска

Поисковый запрос

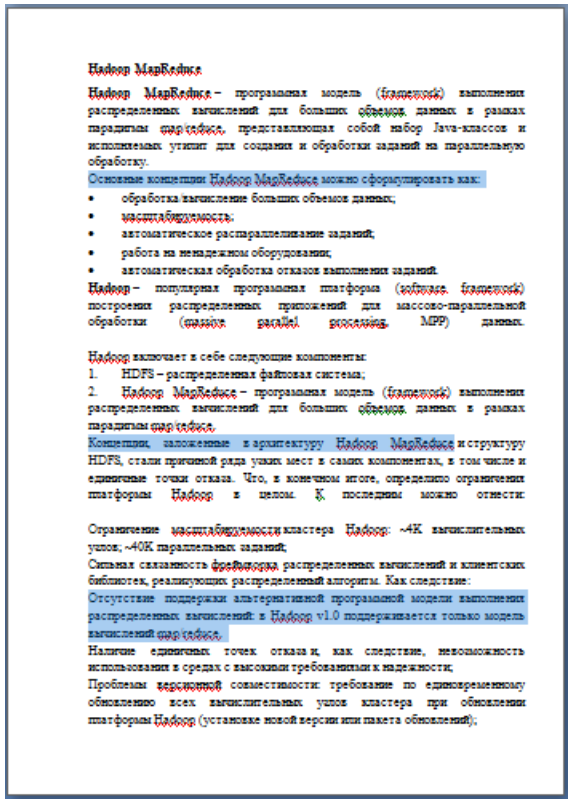
набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

документы d с распределением $p(t|d)$, близким к распределению $p(t|q)$ запроса

Два задания асессорам

1. найти как можно больше статей, пользуясь любыми средствами поиска (и засечь время)
2. оценить релевантность поисковой выдачи на том же запросе



Hadoop MapReduce

Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce, представляющая собой набор Java-классов и исполняемых утилит для создания и обработки заданий на параллельную обработку.

Основные концепции Hadoop MapReduce можно сформулировать как:

- обработка/вычисление больших объемов данных;
- масштабируемость;
- автоматическое распараллеливание заданий;
- работа на ненадежном оборудовании;
- автоматическая обработка отказов выполнения заданий.

Hadoop – популярная программная платформа (software framework) построения распределенных приложений для массово-параллельной обработки (massive parallel processing, MPP) данных.

Hadoop включает в себе следующие компоненты:

1. HDFS – распределенная файловая система;
2. Hadoop MapReduce – программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce.

Концепции, заложенные в архитектуру Hadoop MapReduce и структуру HDFS, стали причиной ряда узких мест в самих компонентах, в том числе и единичные точки отказа. Что, в конечном итоге, определило ограничения платформы Hadoop в целом. К последним можно отнести:

Ограничение масштабируемости кластера Hadoop: ~4K вычислительных узлов; ~40K параллельных заданий;

Сильная связанность фреймворка распределенных вычислений и клиентских библиотек, реализующих распределенный алгоритм. Как следствие:

Отсутствие поддержки альтернативной программной модели выполнения распределенных вычислений: в Hadoop v1.0 поддерживается только модель вычислений map/reduce.

Наличие единичных точек отказа и, как следствие, невозможность использования в средах с высокими требованиями к надежности;

Проблемы версионной совместимости: требование по единовременному обновлению всех вычислительных узлов кластера при обновлении платформы Hadoop (установке новой версии или пакета обновлений);

Пример запроса для разведочного поиска

Пример: фрагмент запроса «Система IBM Watson»

IBM Watson — суперкомпьютер фирмы IBM, оснащённый вопросно-ответной системой искусственного интеллекта, созданный группой исследователей под руководством Дэвида Феруччи. Его создание — часть проекта DeepQA. Основная задача Уотсона — понимать вопросы, сформулированные на естественном языке, и находить на них ответы в базе данных. Назван в честь основателя IBM Томаса Уотсона.

IBM Watson представляет собой когнитивную систему, которая способна понимать, делать выводы и обучаться. Она также позволяет преобразовывать целые отрасли, различные направления науки и техники. Например, предсказывать появление эпидемий или возникновения очагов природных катастроф в различных регионах, вести мониторинг состояния атмосферы больших городов, оптимизировать бизнес-процессы, узнавать, какие товары будут в тренде в ближайшее время.

... ...

Релевантные тексты: примеры сервисов и приложений, основа которых — когнитивная платформа IBM Watson, используемые в IBM Watson технологии, вопрос-ответные системы, сопоставление IBM Watson с Wolfram-Alpha.

Нерелевантные тексты: общие вопросы искусственного интеллекта, другие коммерческие решения на рынке бизнес-аналитики.

Тематика запросов разведочного поиска

Примеры заголовков разведочных запросов к Хабру (объём каждого запроса — около одной страницы A4):

Алгоритмы раскраски графов	Система IBM Watson
Рекомендательная система Netflix	3D-принтеры
Методики быстрого набора текста	CERN-кластер
Космические проекты Илона Маска	AB-тестирование
Технологии Hadoop MapReduce	Облачные сервисы
Беспилотный автомобиль Google car	Контекстная реклама
Криптосистемы с открытым ключом	Марсоход Curiosity
Обзор платформ онлайн-курсов	Видеокарты NVIDIA
Data Science Meetups в Москве	Распознавание образов
Образовательные проекты mail.ru	Сервисы Google scholar
Межпланетная станция New horizons	MIT MediaLab Research
Языковая модель word2vec	Платформа Microsoft Azure

Оценивание качества поиска

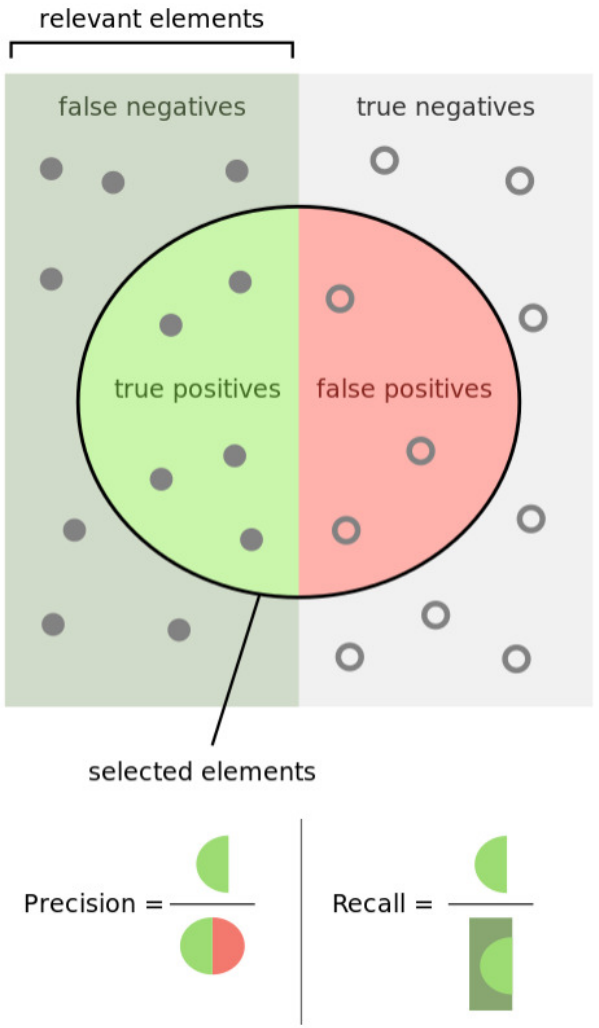
Precision — доля релевантных среди найденных
Recall — доля найденных среди релевантных

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \text{ — точность (precision)}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ — полнота, (recall)}$$

$$F_1 = \frac{P + R}{2PR} \text{ — F1-мера}$$

TP (true positive) — найденные релевантные
FP (false positive) — найденные нерелевантные
FN (false negative) — ненайденные релевантные



Какие модели поиска сравнивались

- **assessors:** результаты поиска, выполненного асессорами
- **TF-IDF, BM25:** сравнение документов по частотам слов
- **word2vec:** нетематические векторные представления слов
- **PLSA:** Probabilistic Latent Semantic Analysis (1999)
- **LDA:** Latent Dirichlet Allocation (2003)
- **ARTM:** тематическая модель с тремя регуляризаторами
- **hARTM:** двухуровневая иерархическая модель ARTM

Задачи регуляризаторов в ARTM и hARTM:

- сделать темы как можно более различными
- сделать векторы $p(t|d)$ как можно более разреженными
- не допустить вырожденности распределений $p(w|t)$

Стратегия регуляризации

Последовательное применение трёх регуляризаторов

1. декоррелирование тем:

$$R(\Phi) = -\tau \sum_{s,t \in T} \sum_{w \in W} \phi_{wt} \phi_{ws}$$

2. разреживание распределений $p(t|d)$:

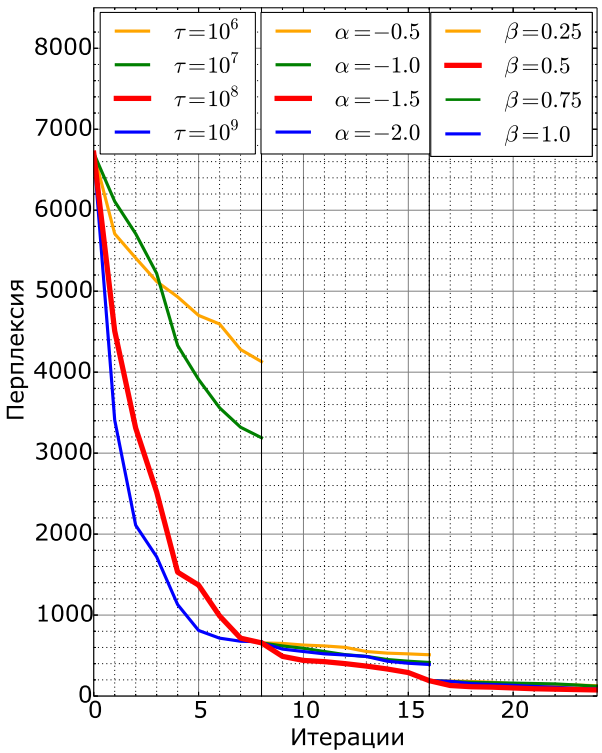
$$R(\Theta) = -\alpha \sum_{d,t} \ln \theta_{td}$$

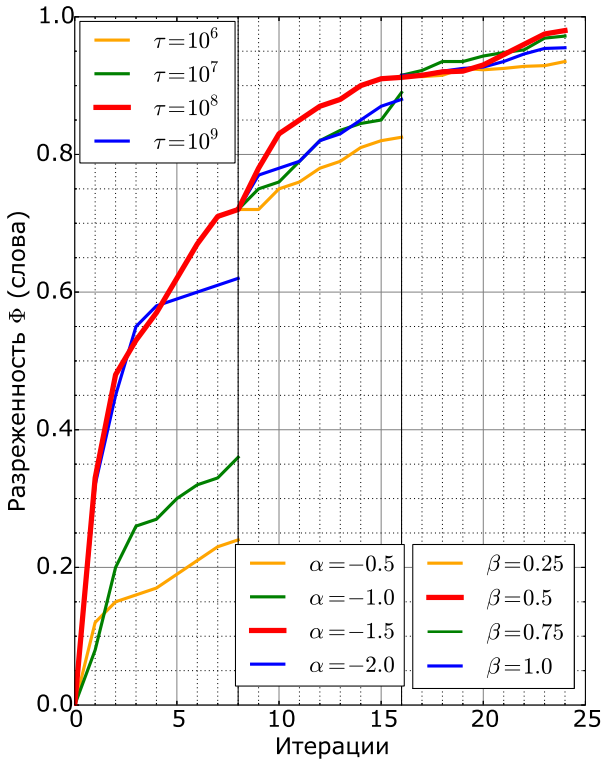
3. сглаживание распределений $p(w|t)$:

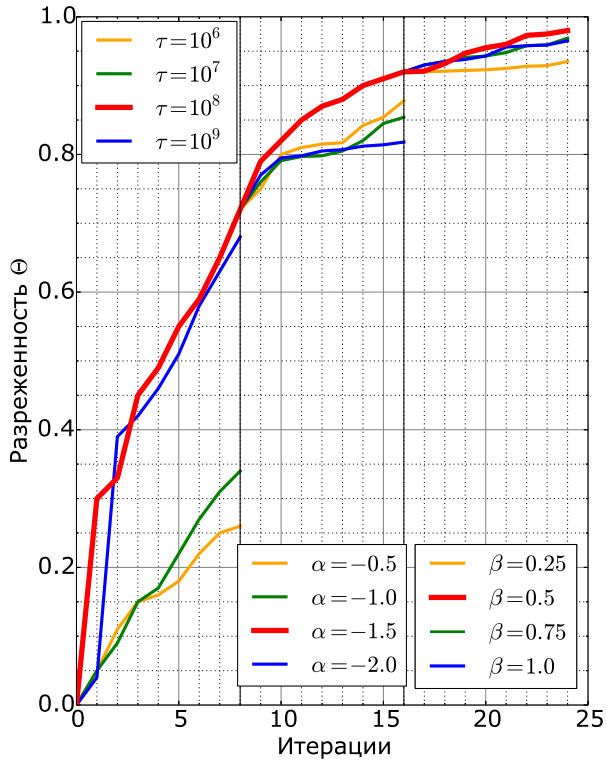
$$R(\Phi) = \beta \sum_{t,w} \ln \phi_{wt}$$

Последовательный подбор коэффициентов регуляризации

- декоррелирование распределений терминов в темах (τ),
- разреживание распределений тем в документах (α),
- сглаживание распределений терминов в темах (β).



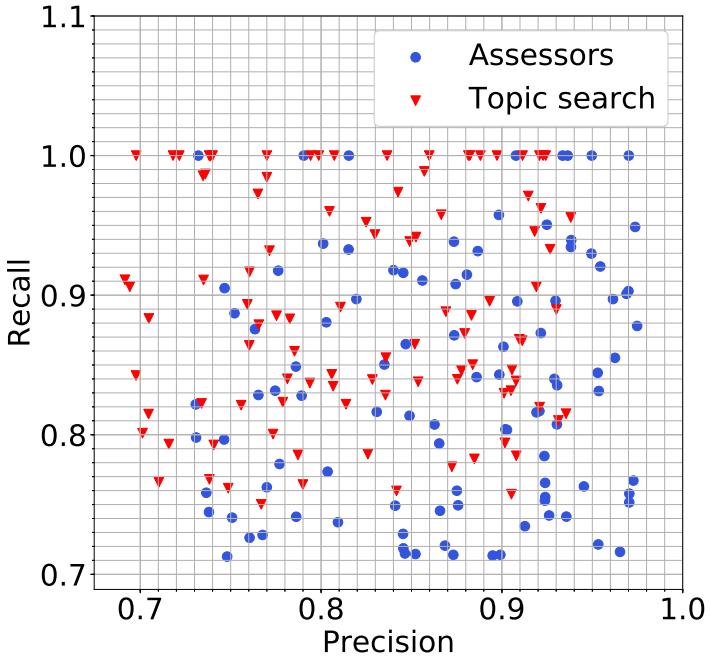




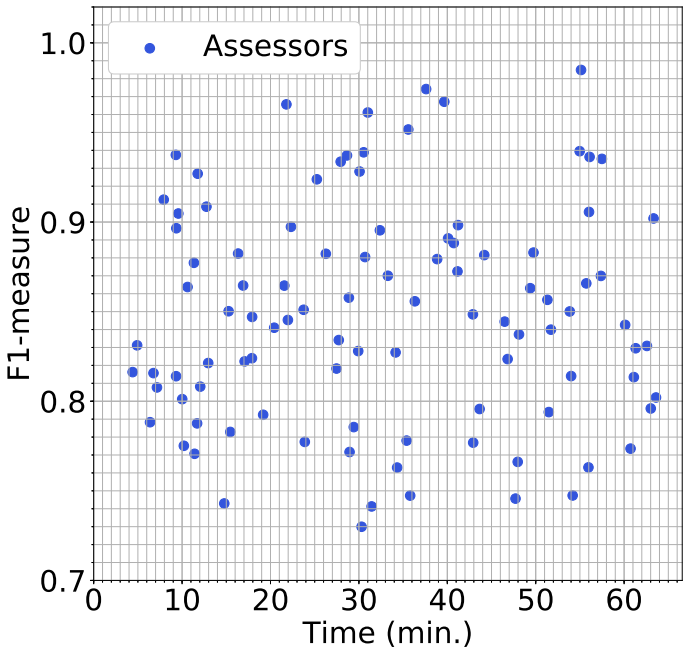
Результаты измерения точности и полноты по запросам

100 запросов, 3 асессора на запрос

точность и полнота поиска



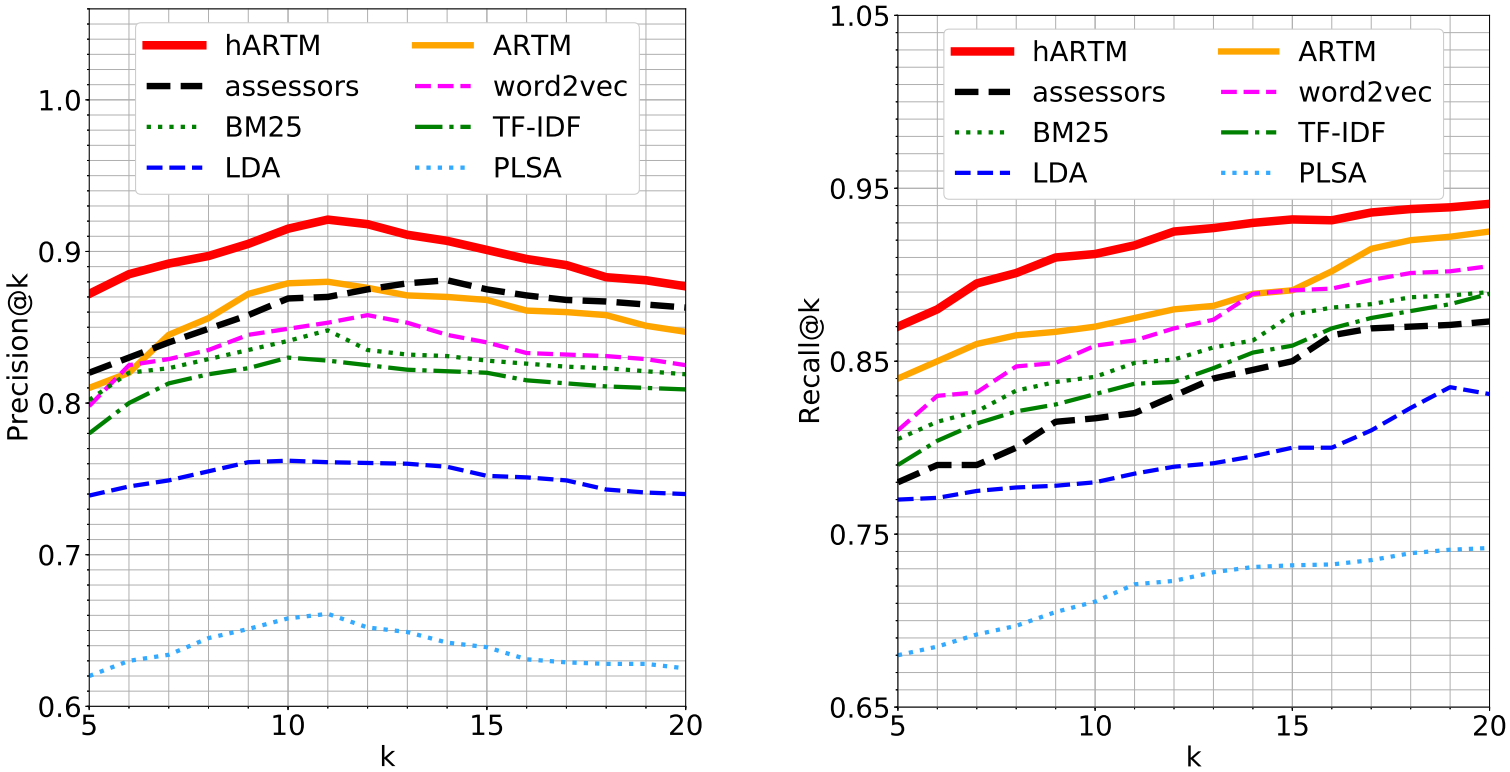
время и F_1-мера (асессоры)



- среднее время обработки запроса асессором — 30 минут
- точность выше у асессоров, полнота — у поисковика

Сравнение с асессорами по качеству поиска

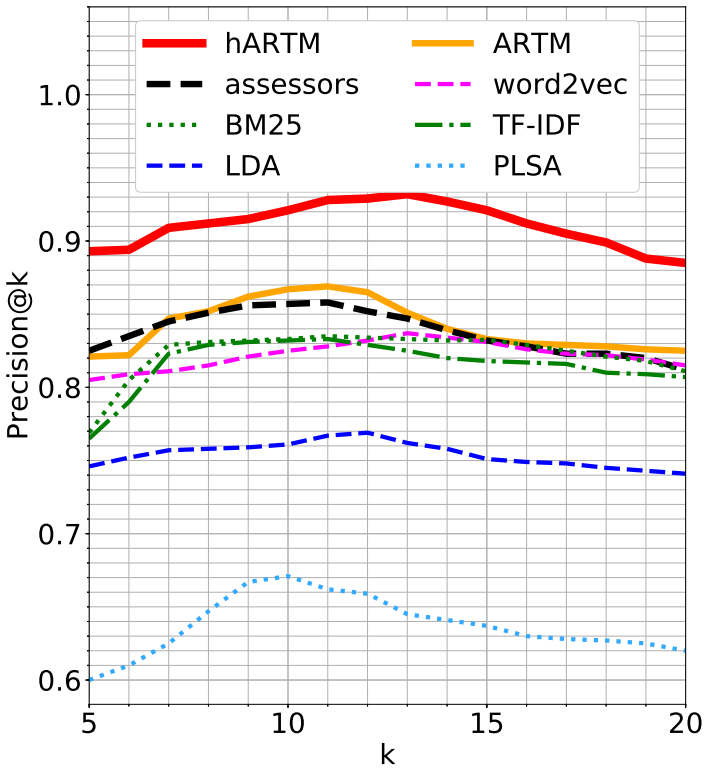
Точность и полнота по первым k позициям поисковой выдачи (коллекция Habrahabr.ru)

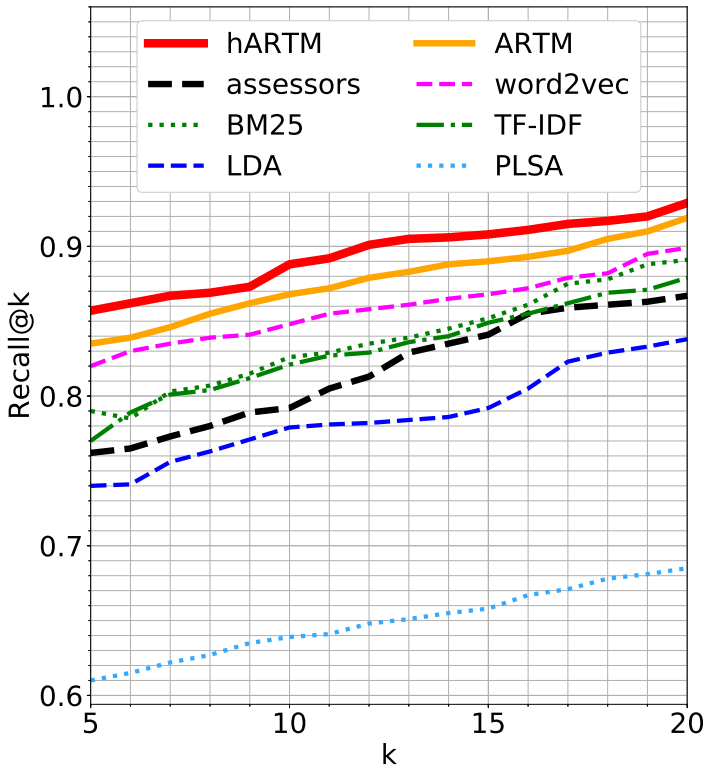


A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Сравнение с асессорами по качеству поиска

Точность и полнота по первым k позициям поисковой выдачи (коллекция TechCrunch.com)





A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Влияние меры близости документа и запроса на качество поиска

Меры близости распределений:
Euclidean, Cosine, Manhattan, Hellinger, Kullback–Leibler

	Коллекция Habrahabr.ru					Коллекция TechCrunch.com				
	Eu	**cos**	Ma	He	KL	Eu	**cos**	Ma	He	KL
Prec@5	0.612	**0.810**	0.682	0.709	0.721	0.635	**0.819**	0.673	0.732	0.715
Prec@10	0.657	**0.879**	0.697	0.735	0.749	0.665	**0.867**	0.683	0.752	0.732
Prec@15	0.627	**0.868**	0.635	0.727	0.711	0.643	**0.833**	0.642	0.742	0.724
Prec@20	0.619	**0.847**	0.627	0.728	0.707	0.638	**0.825**	0.638	0.729	0.708
Recall@5	0.672	**0.840**	0.692	0.721	0.803	0.658	**0.835**	0.669	0.733	0.775
Recall@10	0.682	**0.870**	0.707	0.775	0.856	0.671	**0.868**	0.682	0.753	0.787
Recall@15	0.705	**0.891**	0.725	0.791	0.878	0.715	**0.890**	0.708	0.785	0.809
Recall@20	0.703	**0.925**	0.732	0.812	0.888	0.712	**0.919**	0.715	0.808	0.812

- Наилучшее качество поиска — при косинусной мере
- Одни и те же асессорские оценки можно использовать для оценивания новых моделей и поисковых движков

Влияние комбинаций регуляризаторов на качество поиска

Декоррелирование, Θ-разреживание, Φ-сглаживание

	Коллекция Habrahabr.ru				Коллекция TechCrunch.com			
	$R = 0$	Д	ДΘ	ДΘΦ	$R = 0$	Д	ДΘ	ДΘΦ
Prec@5	0.628	0.748	0.771	**0.810**	0.652	0.775	0.779	**0.819**
Prec@10	0.653	0.776	0.812	**0.879**	0.679	0.787	0.819	**0.867**
Prec@15	0.642	0.765	0.792	**0.868**	0.669	0.773	0.798	**0.833**
Prec@20	0.643	0.759	0.783	**0.847**	0.673	0.777	0.792	**0.825**
Recall@5	0.692	0.784	0.805	**0.840**	0.673	0.812	0.812	**0.835**
Recall@10	0.714	0.814	0.834	**0.870**	0.685	0.821	0.845	**0.868**
Recall@15	0.725	0.835	0.867	**0.891**	0.712	0.859	0.869	**0.890**
Recall@20	0.735	0.862	0.891	**0.925**	0.723	0.882	0.895	**0.919**

- Комбинирование регуляризаторов улучшает качество поиска,
- хотя исходно все регуляризаторы нацелены на улучшение интерпретируемости тем и не оптимизируют поиск явно

Влияние сочетания модальностей на качество поиска

Коллекция Habrahabr.ru. Число тем $|T| = 200$. Модальности: Слова, Биграммы, Теги, Хабы, Комментаторы, Авторы.

	асессоры	С	К	СБ	СБТХ	все
Prec@5	0.821	0.612	0.549	0.654	0.737	**0.810**
Prec@10	0.869	0.635	0.568	0.701	0.752	**0.879**
Prec@15	0.875	0.625	0.532	0.685	0.682	**0.868**
Prec@20	0.863	0.616	0.533	0.682	0.687	**0.847**
Recall@5	0.780	0.722	0.636	0.797	0.827	**0.840**
Recall@10	0.817	0.744	0.648	0.812	0.875	**0.870**
Recall@15	0.850	0.778	0.677	0.842	0.893	**0.891**
Recall@20	0.873	0.803	0.685	0.852	0.898	**0.925**

- Наилучшее качество поиска — по всем модальностям
- Наиболее полезные модальности — слова и теги

Влияние сочетания модальностей на качество поиска

Коллекция TechCrunch.com. Число тем $|T| = 450$.
Модальности: Слова, Категории, Биграммы, Авторы.

	асессоры	С	К	СБ	СБК	все
Prec@5	0.822	0.711	0.557	0.767	0.808	**0.819**
Prec@10	0.851	0.721	0.581	0.783	0.818	**0.867**
Prec@15	0.835	0.733	0.594	0.793	0.833	**0.833**
Prec@20	0.813	0.727	0.566	0.772	0.822	**0.825**
Recall@5	0.762	0.752	0.657	0.775	0.825	**0.835**
Recall@10	0.792	0.776	0.669	0.808	0.855	**0.868**
Recall@15	0.835	0.782	0.684	0.825	0.877	**0.890**
Recall@20	0.867	0.825	0.702	0.837	0.901	**0.919**

- Наилучшее качество поиска — по всем модальностям
- Наиболее полезные модальности — слова и категории

Влияние числа тем на качество поиска

Коллекция Habrahabr.ru

Используем все 5 модальностей, меняем $|T|$

	асессоры	100	150	**200**	250	400
Prec@5	0.821	0.662	0.721	**0.810**	0.761	0.693
Prec@10	0.869	0.761	0.812	**0.879**	0.825	0.673
Prec@15	0.875	0.733	0.795	**0.868**	0.791	0.651
Prec@20	0.863	0.724	0.795	**0.847**	0.792	0.642
Recall@5	0.780	0.732	0.807	**0.840**	0.821	0.721
Recall@10	0.817	0.771	0.843	**0.870**	0.851	0.751
Recall@15	0.850	0.824	**0.895**	0.891	0.871	0.773
Recall@20	0.873	0.857	0.905	**0.925**	0.892	0.771

- Наилучшее качество поиска — при 200 темах
- Тематический поиск превосходит асессоров по полноте

Влияние числа тем на качество поиска

Коллекция TechCrunch.com

Используем все 4 модальности, меняем $|T|$

	асессоры	350	400	450	**475**	500
Prec@5	0.822	0.653	0.725	0.752	**0.819**	0.777
Prec@10	0.851	0.663	0.732	0.762	**0.867**	0.811
Prec@15	0.835	0.682	0.743	0.787	**0.833**	0.793
Prec@20	0.813	0.650	0.743	0.773	**0.825**	0.793
Recall@5	0.762	0.731	0.762	0.793	**0.835**	0.817
Recall@10	0.792	0.763	0.793	0.812	**0.868**	0.855
Recall@15	0.835	0.782	0.807	0.855	**0.890**	0.882
Recall@20	0.867	0.792	0.823	0.862	**0.919**	0.903

- Наилучшее качество поиска — при 475 темах
- Тематический поиск превосходит асессоров по полноте

Выводы по результатам экспериментов

- Регуляризаторы, улучшающие интерпретируемость модели, повышают также и качество поиска
- Двухуровневая иерархия улучшает качество поиска (в основном точность) благодаря сужению области поиска
- Подбор траектории регуляризации и оптимизация коэффициентов регуляризации влияет на качество поиска
- При тщательной оптимизации тематический поиск превосходит как асессоров, так и конкурирующие модели
- Асессорские данные относятся не к темам, а к коллекции; поэтому с их помощью можно оценивать новые модели
- Небольших асессорских данных хватает для оценивания тематических моделей, которые обучаются *без учителя*

Резюме

Разведочный информационный поиск (exploratory search):

- это поиск по смыслу, а не по ключевым словам
- может быть построен на тематическом моделировании
- требует многофункциональности от тематических моделей
- является одной из главных мотиваций для ARTM
- и, в частности, для иерархических моделей

Открытые проблемы

- построение разноуровневых иерархий в ARTM
- оценивание качества тематических иерархий
- оптимизация числа тем на каждом уровне иерархии

Датасеты для заданий по спецкурсу

- Научно-популярные статьи: ПостНаука, Элементы, Хабр
- Википедия
- Вики-227
- Новостной поток (RSS lenta.ru / нефильтрованый поток)
- Акты арбитражных судов РФ
- TechCrunch (английский)
- Открытые датасеты (английский): 20 newsgroups, NIPS, KOS
- Транзакции клиентов Sberbank DSD 2016

Проблемы для исследования

Задача-минимум: научиться решать задачи NLP и строить тематические модели в BigARTM

Задача-максимум: решить открытую проблему

- Несбалансированность и семантическая однородность тем
- Агрегирование гетерогенных коллекций
- Создание новых тем при расширении коллекции
- В том числе, создание новых тем в иерархиях
- Прослеживание тем в новостных потоках
- Тематическая сегментация и посегментный поиск
- Предобученные тематические векторные представления слов
- Визуализация «карт знаний» (D3.js)