Теория, алгоритмы и приложения вероятностного тематического моделирования

Воронцов Константин Вячеславович
(ВЦ РАН ФИЦ ИУ РАН • МФТИ)

Математический кружок школы ПМИ МФТИ
11 ноября 2016

Аннотация

Мы рассмотрим численные методы решения некорректно поставленных задач матричного разложения, используемые для тематического моделирования текстовых коллекций. Разберём примеры формализации и комбинирования моделей.

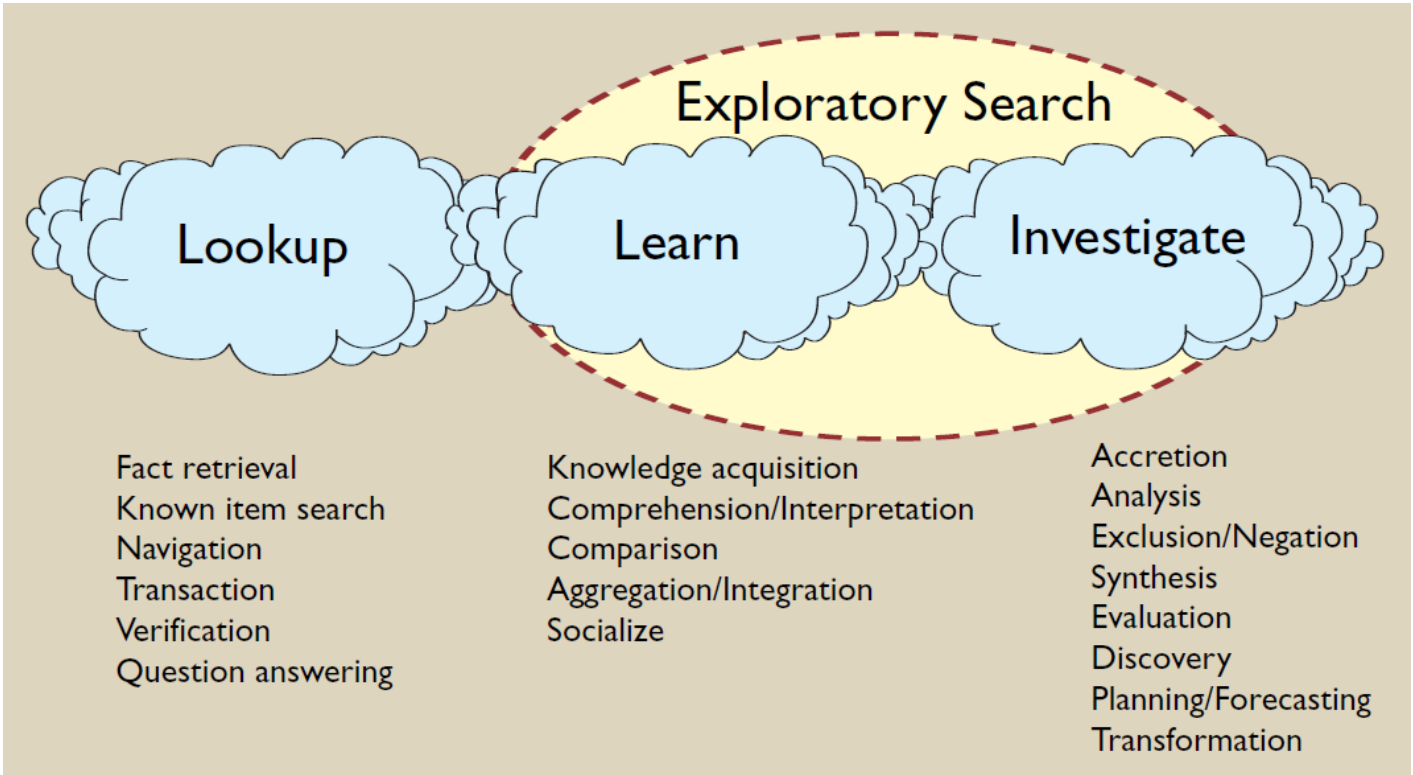
Я расскажу о проекте с открытым кодом BigARTM и прикладных задачах, которые мы решаем с его помощью.

В заключение затронем несколько открытых проблем, связанных с вопросами единственности, устойчивости, полноты, сходимости и вычислительной сложности алгоритмов тематического моделирования больших текстовых коллекций.

Содержание

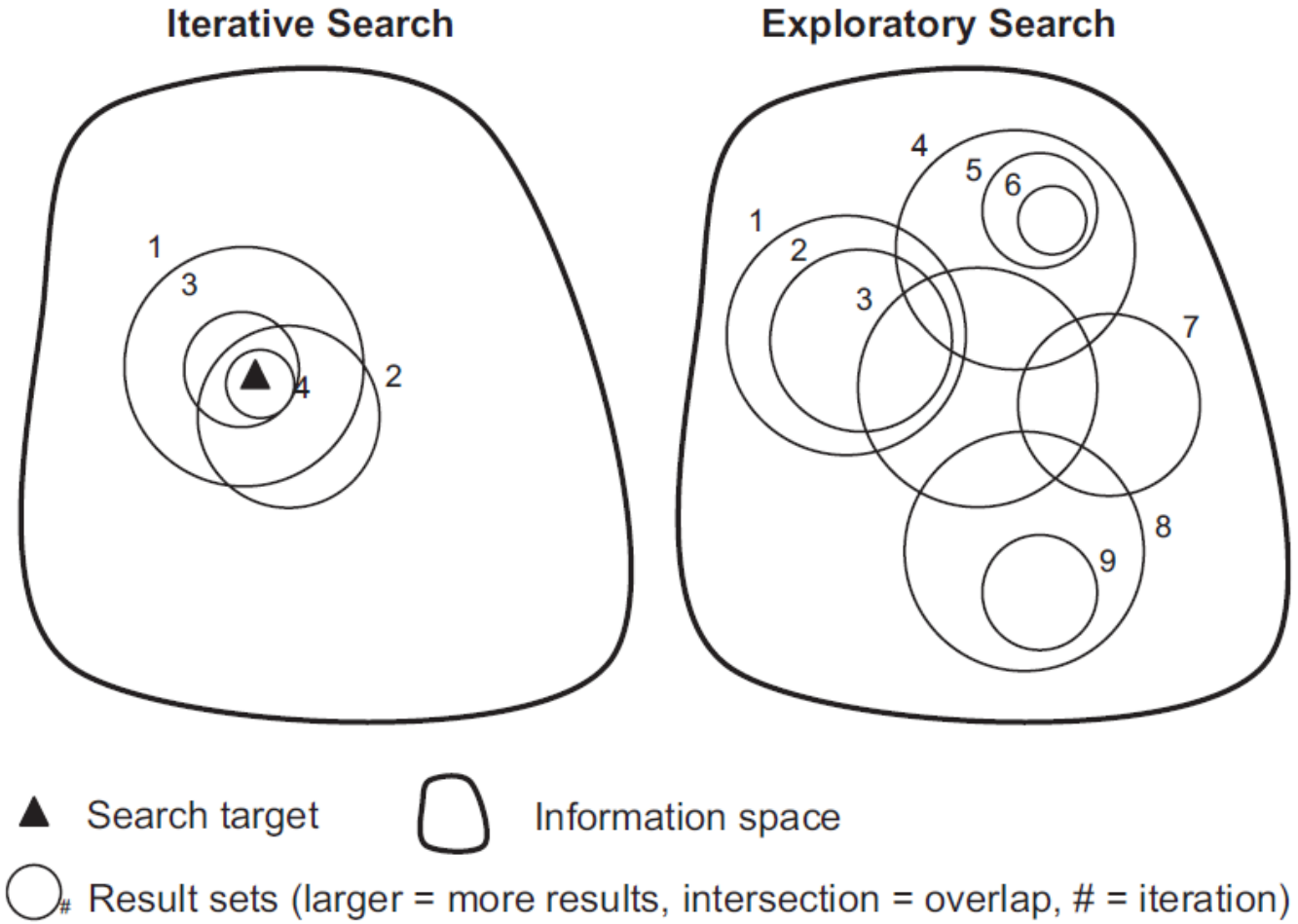
Концепция разведочного поиска (exploratory search)

- пользователь может не знать ключевых терминов
- пользователя может интересовать множество ответов



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)

Information Seeking Mantra [B.Shneiderman, 1996]

«Overview first, zoom and filter, details on demand»

Понятие *дальнего чтения* [Franco Moretti, 2005]

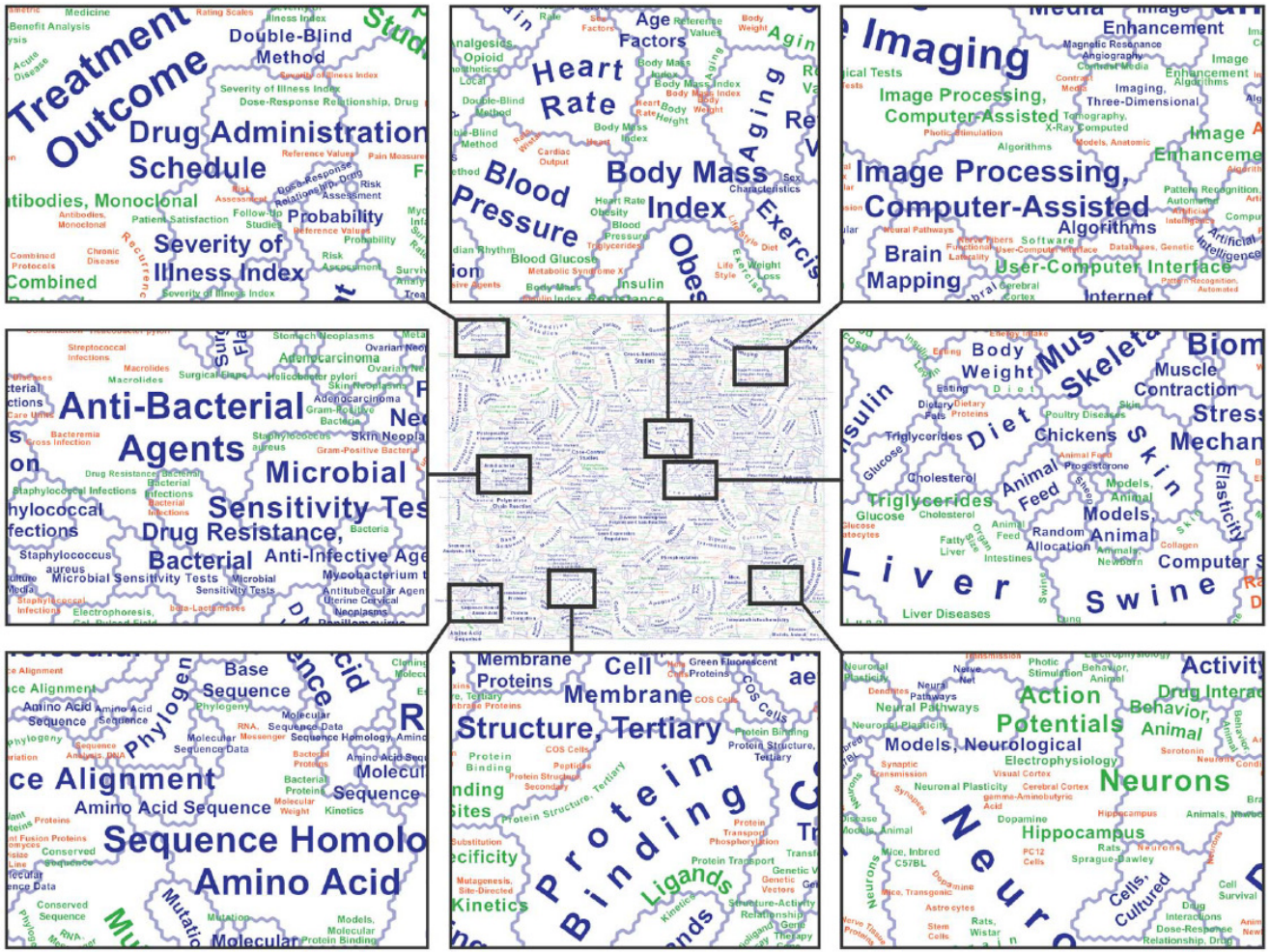
«*Distant reading* is not an obstacle but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models.»

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.

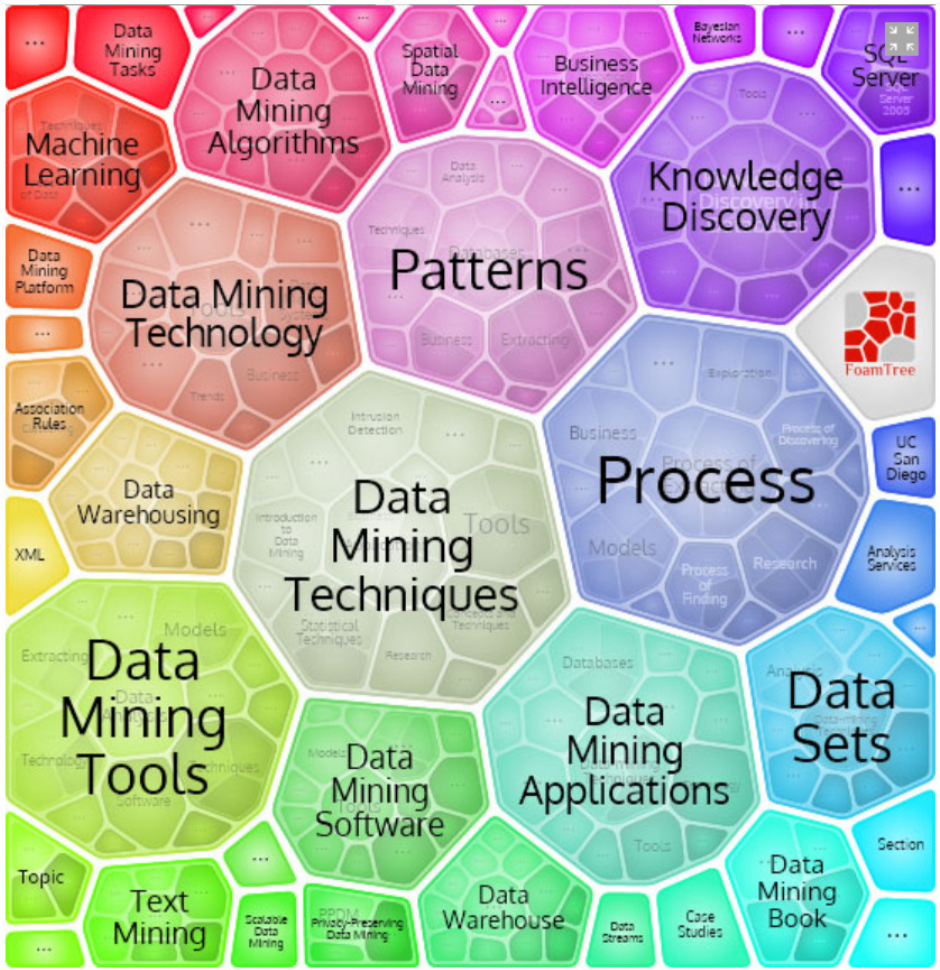
S.Janicke, G.Franzini, M.F.Cheema, G.Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis, 2015.

Пример карты медицинских знаний



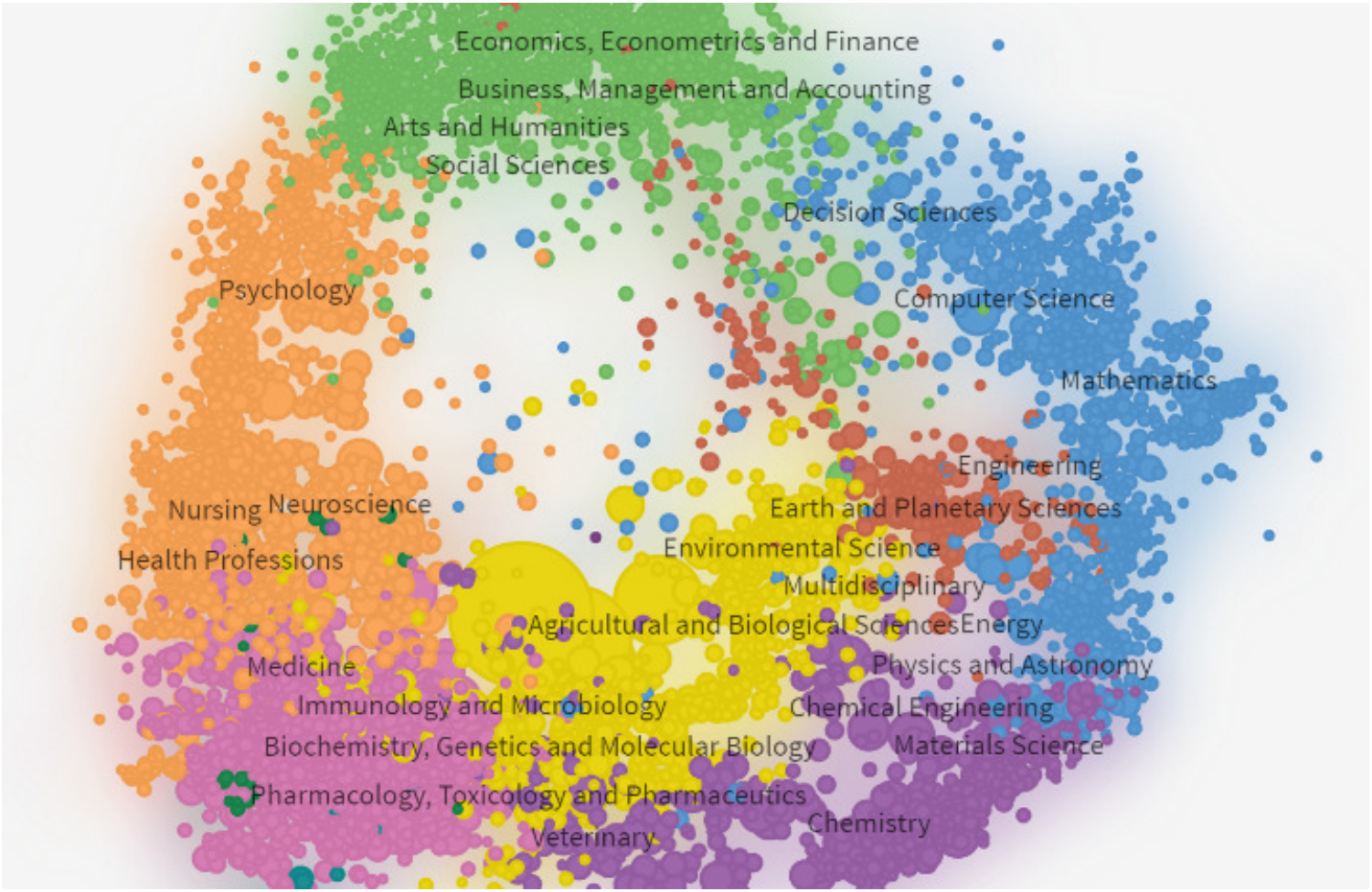
Skupin, Biberstine, Borner. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS ONE, 2013.

Пример иерархической карты области *Data Mining*



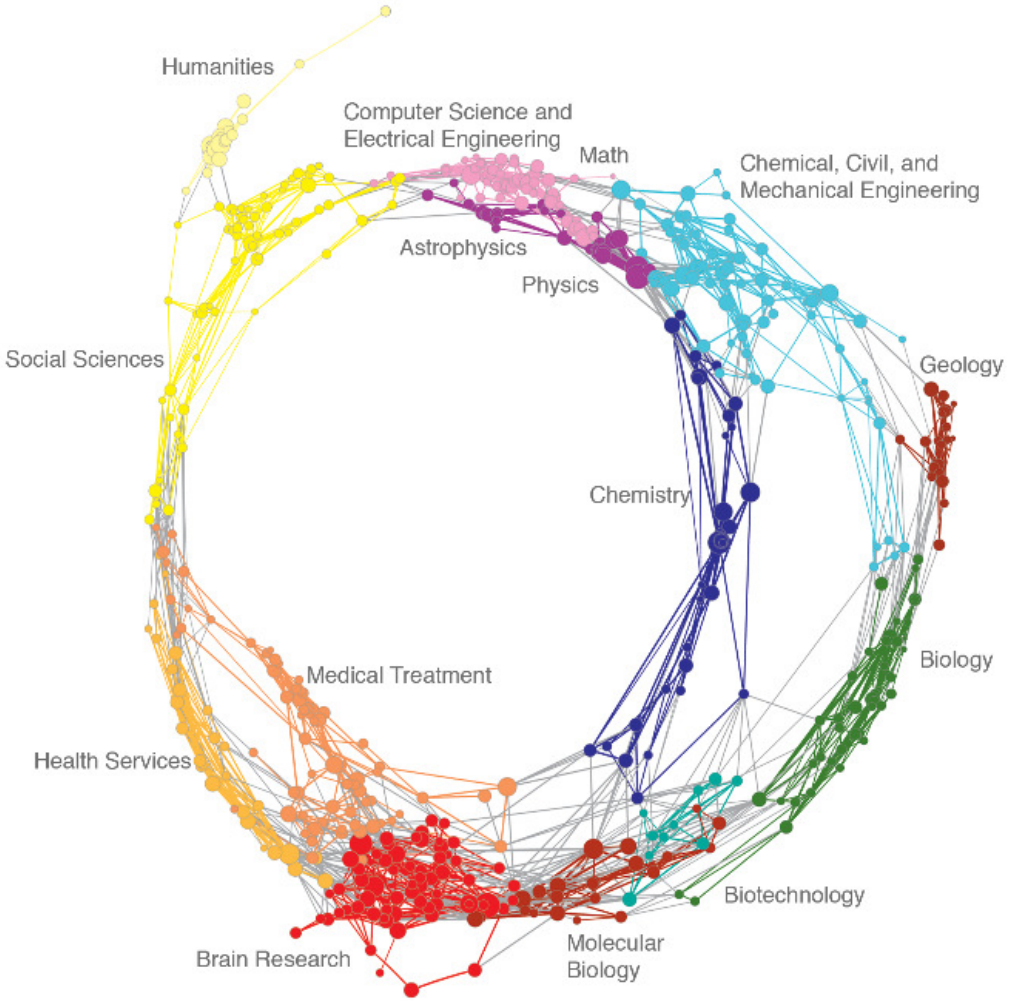
FoamTree: https://carrotsearch.com/foamtree

Пример карты науки



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

Ещё один пример карты науки



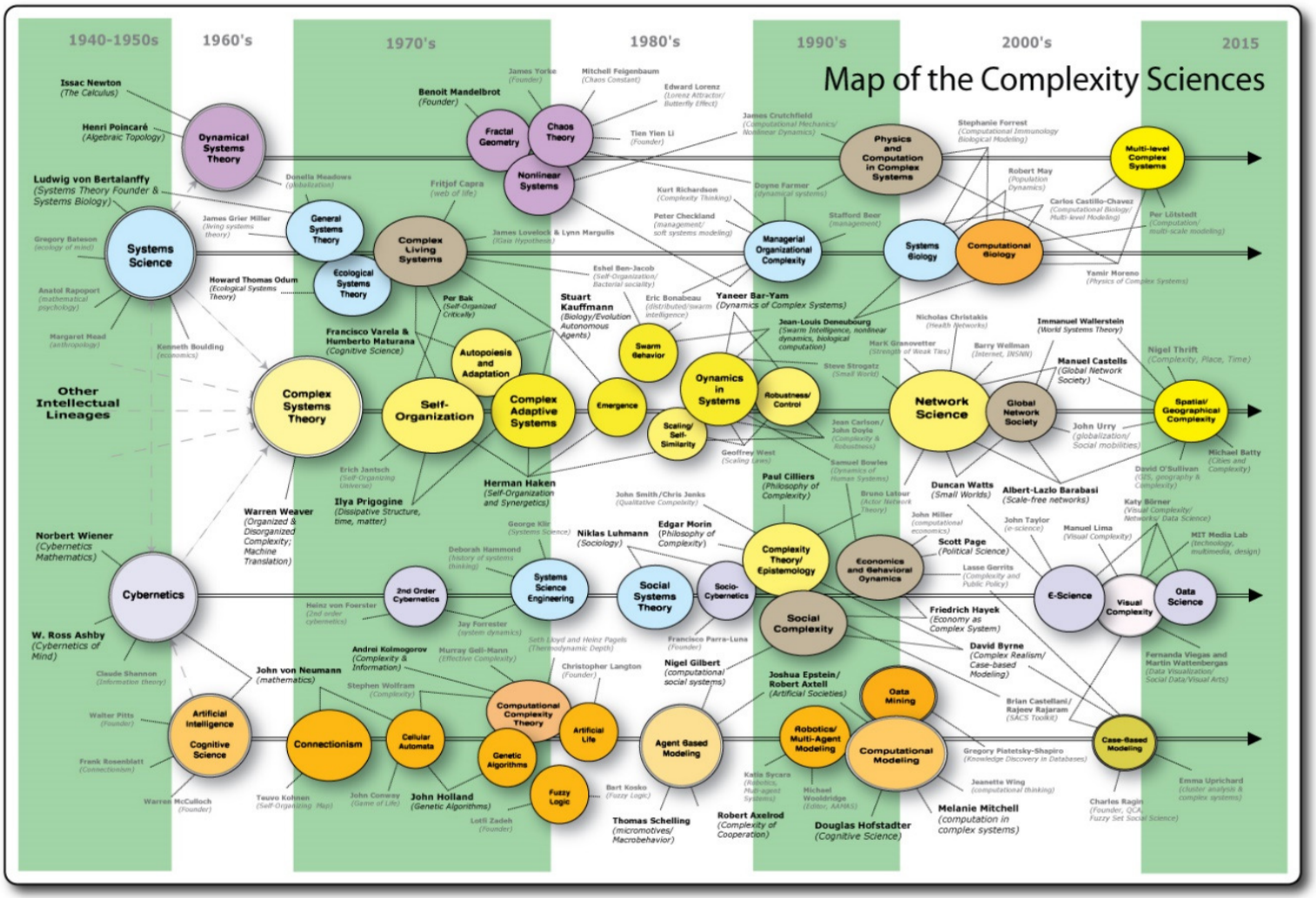
http://scimaps.org

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

Недостатки:

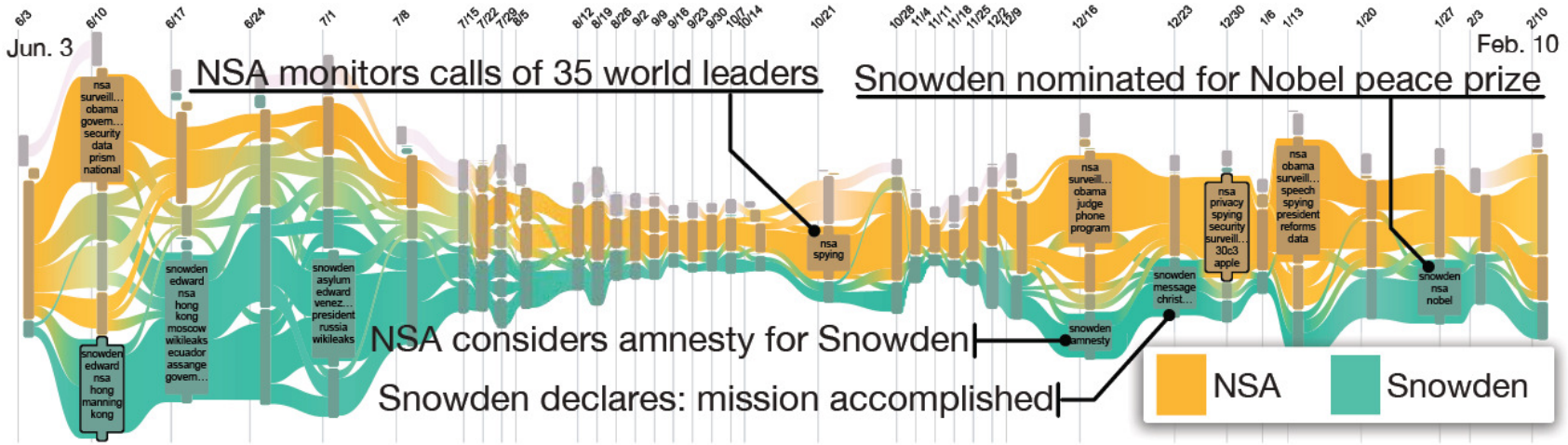
- оси не имеют интерпретации
- искажение сходства при двумерном проецировании

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)

- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- генерирует отчёт.

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

Возможный сценарий разведочного поиска

Поисковый запрос:

- документ любой длины или даже коллекция документов

Цели поиска:

- к каким темам относится мой запрос?
- что ещё известно по этим темам?
- какова тематическая структура этой предметной области?
- какие области являются смежными?
- что ещё есть понятного, обзорного, важного, свежего?

Сценарий поиска:

1. имея любой текст под рукой, в любом приложении,
2. получаем картину содержащихся в нём тем-подтем
3. и «дорожную карту» предметной области в целом

Тематическая модель для разведочного поиска должна быть...

1. **Интерпретируемая**: каждая тема понятна людям
2. **Иерархическая**: систематизация областей знания
3. **Темпоральная**: обнаружение и прослеживание тем
4. **Мультиграммная**: выделение тематичных словосочетаний
5. **Мультимодальная**: авторы, связи, тэги, пользователи,...
6. **Мультиязычная**: кросс- и много-языковой поиск
7. **Разреженная**: сокращение поискового индекса
8. **Сегментирующая**: выделение тем внутри документа
9. **Обучаемая**: учёт обратной связи с пользователями
10. **Создающая и именующая темы** автоматически
11. **Онлайновая**: обрабатывающая коллекцию за 1 проход
12. **Параллельная, распределённая**: для больших коллекций

Коллекция текстовых документов

D — множество документов
W — множество терминов (слов или словосочетаний)
T — множество тем, $|T| \ll |D|$, $|T| \ll |W|$

$(d_i, w_i, t_i)_{i=1}^n \subset D \times W \times T$ — коллекция текстовых документов

Когда автор документа d_i писал термин w_i, он думал о теме t_i, и мы хотели бы выявить, о какой именно.

Основные предположения:

- d_i, w_i — наблюдаемые, темы t_i — скрытые
- порядок терминов в документе не важен (bag of words)
- порядок документов в коллекции не важен (bag of docs)
- слова приведены к нормальным формам (лемматизированы)

Обозначения частот терминов

Ненаблюдаемые частоты:

n_{dwt} — число троек (d, w, t) во всей коллекции

$n_{wt} = \sum_d n_{dwt}$ — число употреблений термина w в теме t

$n_{td} = \sum_w n_{dwt}$ — число терминов темы t в документе d

$n_t = \sum_{d,w} n_{dwt}$ — число терминов темы t в коллекции

Наблюдаемые частоты:

$n_{dw} = \sum_t n_{dwt}$ — частота термина w в документе d

$n_w = \sum_{d,t} n_{dwt}$ — число употреблений термина w в коллекции

$n_d = \sum_{w,t} n_{dwt}$ — длина документа d

$n = \sum_{d,w,t} n_{dwt}$ — длина коллекции

Основные вероятностные предположения

- $(d_i, w_i, t_i)_{i=1}^n$ — равновероятные элементарные события.

  Тогда можно записать условные вероятности:

  $p(w|d) = \dfrac{n_{dw}}{n_d}$ — распределение слов в документе d,

  $p(t|d) = \dfrac{n_{td}}{n_d}$ — распределение тем в документе d,

  $p(w|t) = \dfrac{n_{wt}}{n_t}$ — распределение слов в теме t.

- Гипотеза условной независимости: «вероятность слова в теме не зависит от документа»,

$$p(w|d,t) = p(w|t)$$

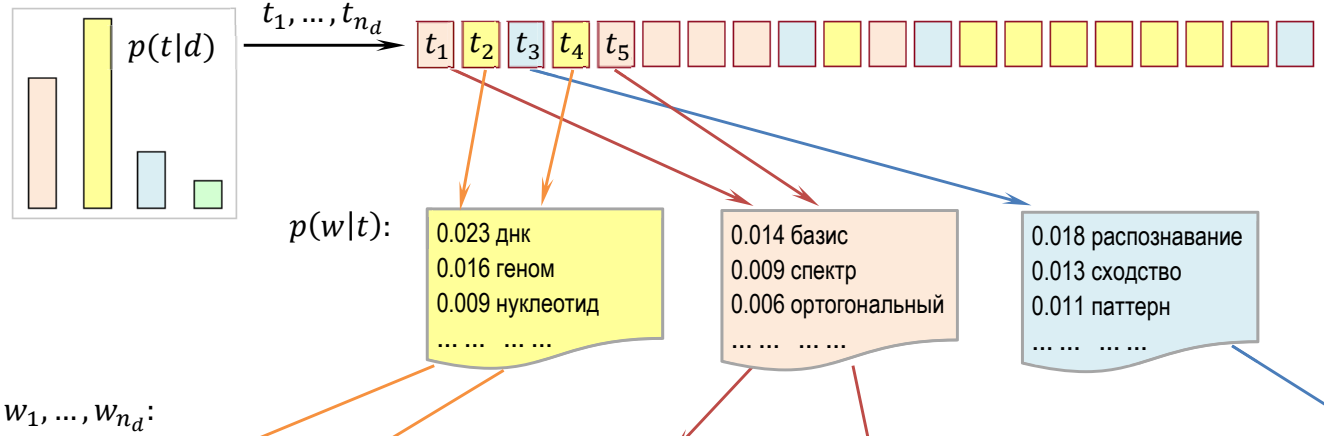
$$\frac{n_{dwt}}{n_{td}} = \frac{n_{wt}}{n_t}$$

Вероятностная модель языка (модель порождения текста)

Формула полной вероятности:

$$p(w|d) = \sum_{t \in T} p(w \mid \cancel{d}, t)\, p(t|d) = \sum_{t \in T} p(w|t) p(t|d)$$

$$\frac{n_{dw}}{n_d} = \sum_{t \in T} \frac{n_{dwt}}{n_{td}} \frac{n_{td}}{n_d} \qquad = \sum_{t \in T} \frac{n_{wt}}{n_t} \frac{n_{td}}{n_d}$$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов

Задача тематического моделирования

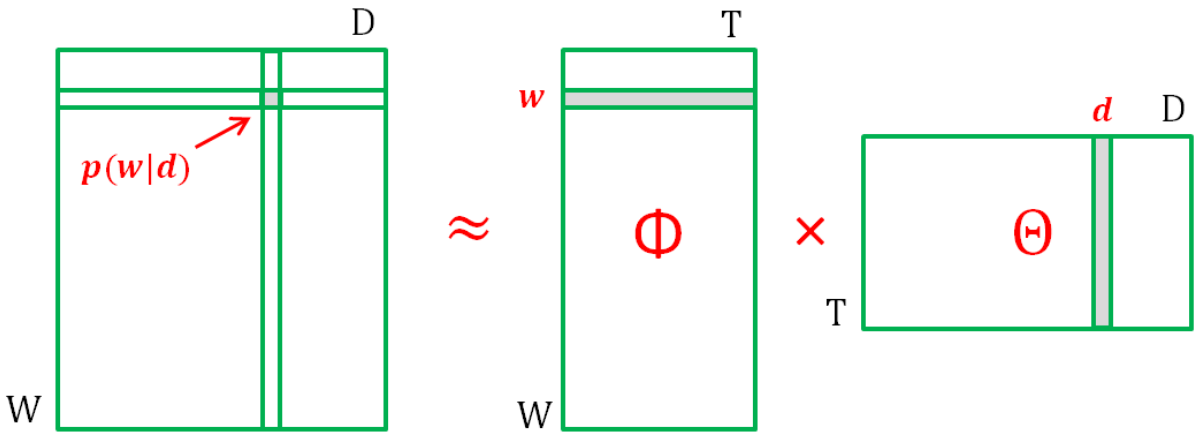
Дано: коллекция текстовых документов

- n_{dw} — частоты терминов в документах, $p(w|d) = \frac{n_{dw}}{n_d}$

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

- $\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:



Элементарный подход к решению задачи

Выразим n_{dwt} согласно гипотезе условной независимости:

$$n_{dwt} = \frac{n_{wt} n_{td}}{n_t} = \frac{n_{wt}}{n_t} \cdot \frac{n_{td}}{n_d} \cdot \frac{n_d}{n_{dw}} n_{dw} = n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} = n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$$

Получим систему уравнений относительно параметров модели ϕ_{wt}, θ_{td} и вспомогательных переменных n_{dwt}:

$$\begin{cases} n_{dwt} = n_{dw} \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}, & d \in D,\ w \in W,\ t \in T; \\ \phi_{wt} \equiv \dfrac{n_{wt}}{n_t} = \dfrac{\sum_d n_{dwt}}{\sum_{d,w} n_{dwt}}, & w \in W,\ t \in T; \\ \theta_{td} \equiv \dfrac{n_{td}}{n_d} = \dfrac{\sum_w n_{dwt}}{\sum_{t,w} n_{dwt}}, & d \in D,\ t \in T. \end{cases}$$

Численное решение — методом простых итераций

Принцип максимума правдоподобия

Правдоподобие — вероятность реализации коллекции $(d_i, w_i)_{i=1}^n$:

$$\prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}$$

Максимизация логарифма правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) p(d) \to \max_{\Phi, \Theta}$$

приводит к задаче максимизации функционала

$$\mathscr{L}(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

при ограничениях неотрицательности и нормировки

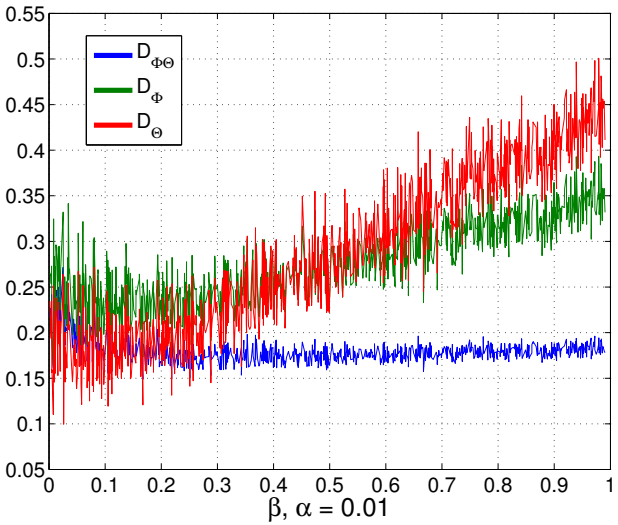
$$\phi_{wt} \geqslant 0; \quad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t \in T} \theta_{td} = 1.$$

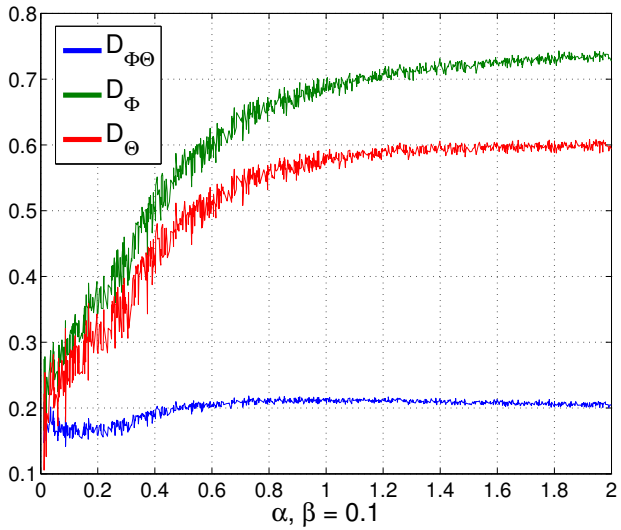
Задача тематического моделирования некорректно поставлена

Неединственность стохастического матричного разложения:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

Эксперимент. Произведение $\Phi\Theta$ восстанавливается устойчиво, матрица Φ и матрица Θ — только когда сильно разрежены:





Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning. 2015.

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

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

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

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

Условия вырожденности модели для тем и документов

Решение может быть вырожденным для некоторых тем (столбцов матриц Φ) и документов (столбцов матрицы Θ).

Тема t вырождена, если для всех терминов $w \in W$

$$n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \leqslant 0.$$

Если тема t вырождена, то $p(w|t) = \phi_{wt} \equiv 0$; это означает, что тема исключается из модели (происходит отбор тем).

Документ d вырожден, если для всех тем $t \in T$

$$n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \leqslant 0.$$

Если документ d вырожден, то $p(t|d) = \theta_{td} \equiv 0$; это означает, что модель не в состоянии описать данный документ.

Напоминания. Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_x; \\ g_i(x) \leqslant 0, \quad i = 1, \ldots, m; \\ h_j(x) = 0, \quad j = 1, \ldots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители μ_i, $i = 1, \ldots, m$, λ_j, $j = 1, \ldots, k$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0; \ h_j(x) = 0; \ \text{(исходные ограничения)} \\ \mu_i \geqslant 0; \ \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; \ \text{(условие дополняющей нежёсткости)} \end{cases}$$

Вывод системы уравнений из условий Каруша–Куна–Таккера

1. Условия ККТ для ϕ_{wt} (для θ_{td} всё аналогично):

$$\sum_d n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \mu_{wt}; \quad \mu_{wt} \geqslant 0; \quad \mu_{wt}\phi_{wt} = 0.$$

2. Умножим обе части равенства на ϕ_{wt} и выделим p_{tdw}:

$$\phi_{wt}\lambda_t = \sum_d n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}.$$

3. Если $\lambda_t \leqslant 0$, то тема t вырождена, $\phi_{wt} \equiv 0$ для всех w.
4. Если $\lambda_t > 0$, то либо $\phi_{wt} = 0$, либо $n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} > 0$:

$$\phi_{wt}\lambda_t = \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

5. Суммируем обе части равенства по $w \in W$:

$$\lambda_t = \sum_{w \in W} \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

6. Подставим λ_t из (5) в (4), получим требуемое. ■

Комбинирование регуляризованных тематических моделей

Максимизация log правдоподобия с n регуляризаторами R_i:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta},$$

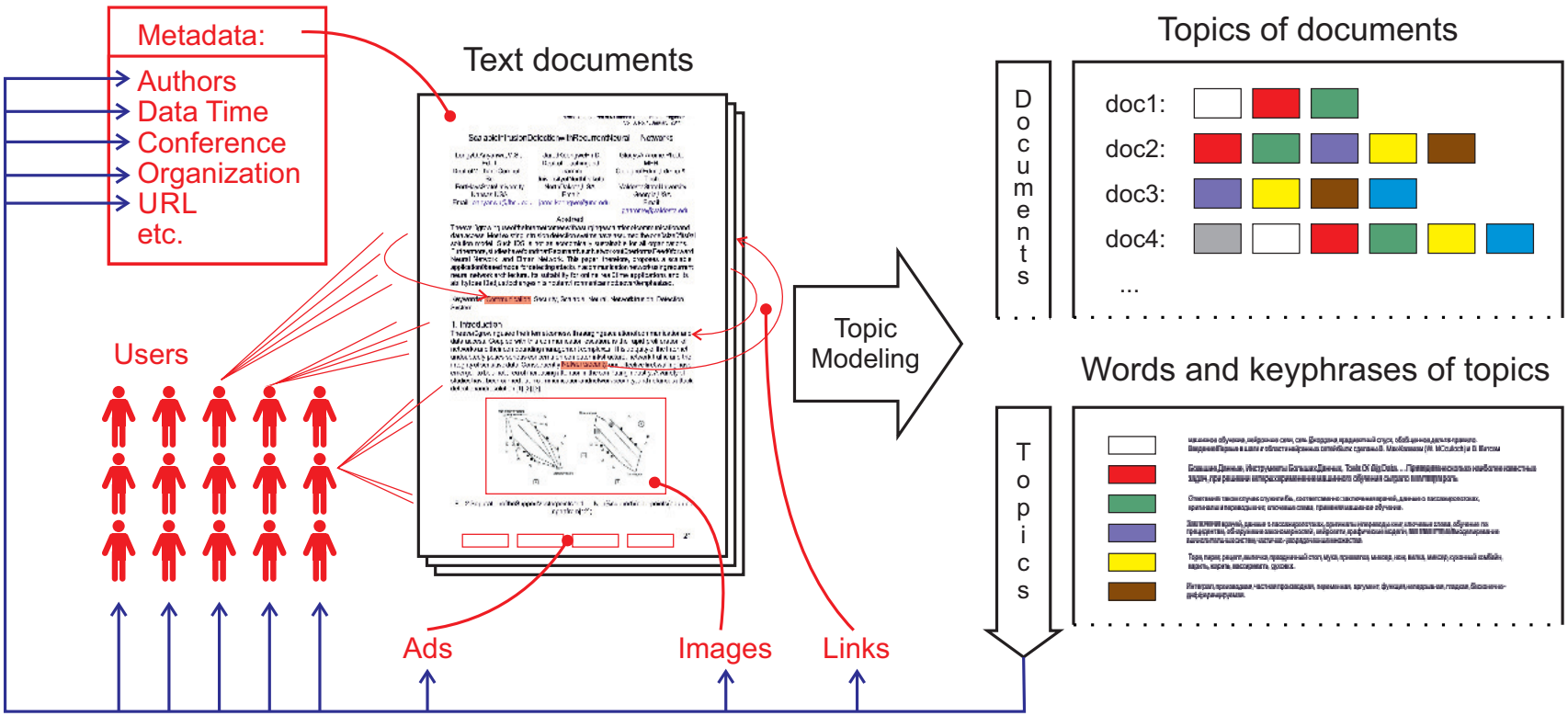
где τ_i — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum\limits_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\sum\limits_{i=1}^{n}\tau_i\frac{\partial R_i}{\partial \phi_{wt}}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum\limits_{w\in d} n_{dw}p_{tdw} + \theta_{td}\sum\limits_{i=1}^{n}\tau_i\frac{\partial R_i}{\partial \theta_{td}}\right) \end{cases}$$

Задачи мультимодального тематического моделирования

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



Мультимодальная ARTM

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

$W = W^1 \sqcup \cdots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Рациональный EM-алгоритм

Идея: E-шаг встраивается внутрь M-шага,
чтобы не хранить трёхмерный массив значений n_{dwt}.

Вход: коллекция D, число тем $|T|$, число итераций $i_{\max}$;
Выход: матрицы терминов тем Θ и тем документов Φ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D$, $w \in W$, $t \in T$;
для всех итераций $i = 1, \ldots, i_{\max}$

$n_{wt}, n_{td} := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
для всех документов $d \in D$ и всех слов $w \in d$

$n_{tdw} := n_{dw} \operatorname*{norm}_{t \in T}(\phi_{wt}\theta_{td})$ для всех $t \in T$;

$n_{wt} \mathrel{+}= n_{tdw}$; $n_{td} \mathrel{+}= n_{tdw}$ для всех $t \in T$;

$$\phi_{wt} := \operatorname*{norm}_{w \in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \text{ для всех } w \in W,\ t \in T;$$

$$\theta_{td} := \operatorname*{norm}_{t \in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \text{ для всех } d \in D,\ t \in T;$$

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый параллельный мультимодальный ARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов

Сообщество:

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

Лицензия и среда разработки:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их вероятности $p(w|t)$ в %:

Тема 88				Тема 251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Дударенко М. А. Регуляризация многоязычных тематических моделей. Вычислительные методы и программирование. 2015. Т. 16. С. 26–36.

Биграммы радикально улучшают интерпретируемость тем

Коллекция 850 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

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

Данные коллективного блога Хабрахабр.ру

Данные

- 132 157 статей
- Модальности:
  - 52 354 терминов (слов)
  - 524 авторов статей
  - 10 000 комментаторов (авторов комментариев к статьям)
  - 2546 тегов
  - 123 хаба (категории)

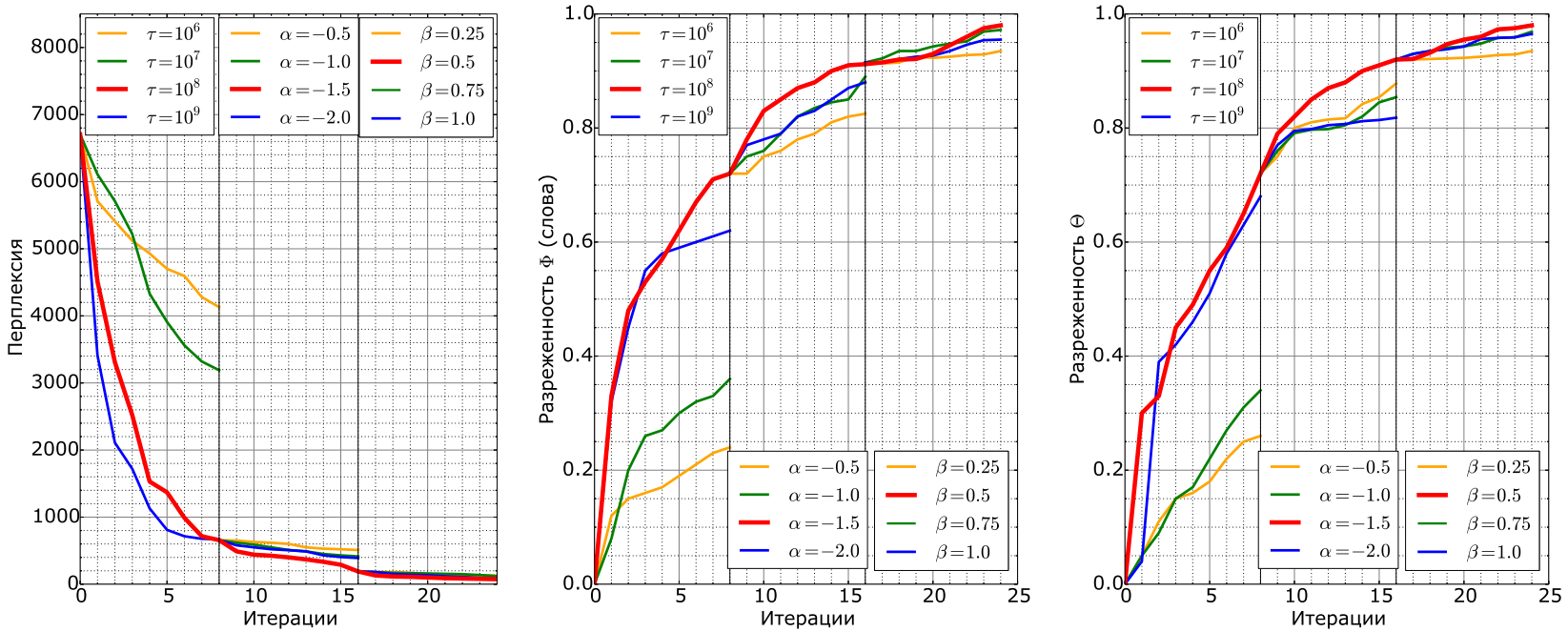
Предобработка текстов

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

Подбор коэффициентов регуляризации

Последовательное добавление регуляризаторов:

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



Разведочный поиск

$q = (w_1, \ldots, w_{n_q})$ — текст запроса произвольной длины n_q

$\theta_{tq} = p(t|q)$ — тематический профиль запроса q
$\theta_{td} = p(t|d)$ — тематические профили 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

Методика оценивания качества разведочного поиска

Поисковый запрос

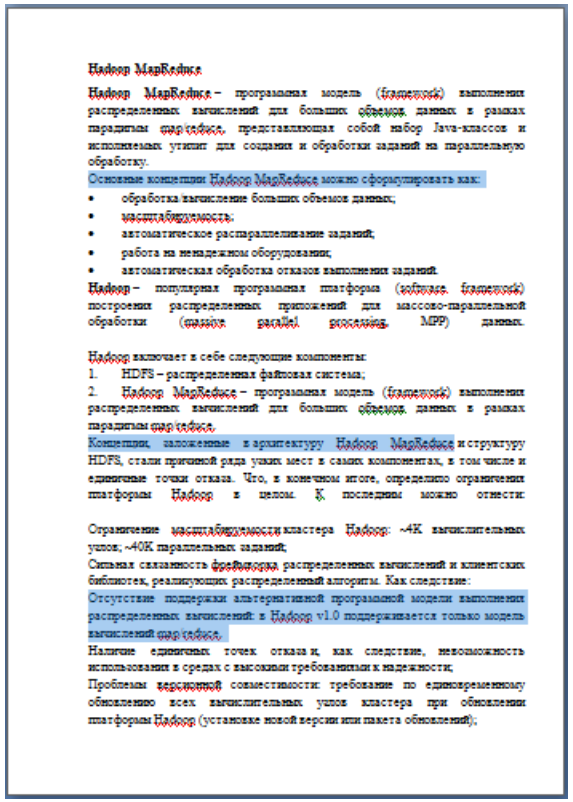
набор ключевых слов или фрагментов текста, около одной страницы A4

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 исполнимых 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

Оценки качества поиска

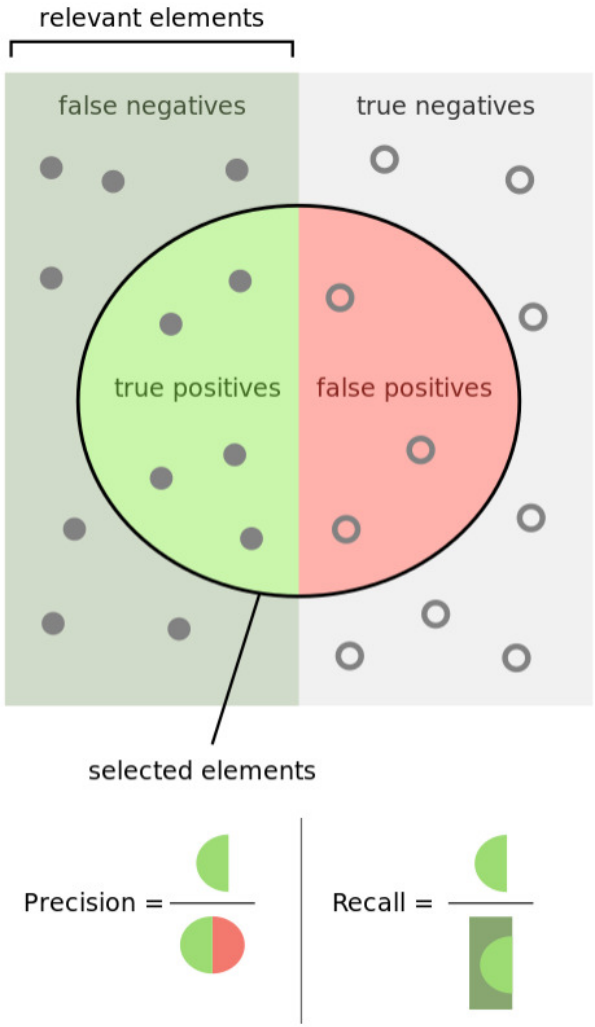
Precision — доля релевантных среди найденных
Recall — доля найденных среди релевантных

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

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

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

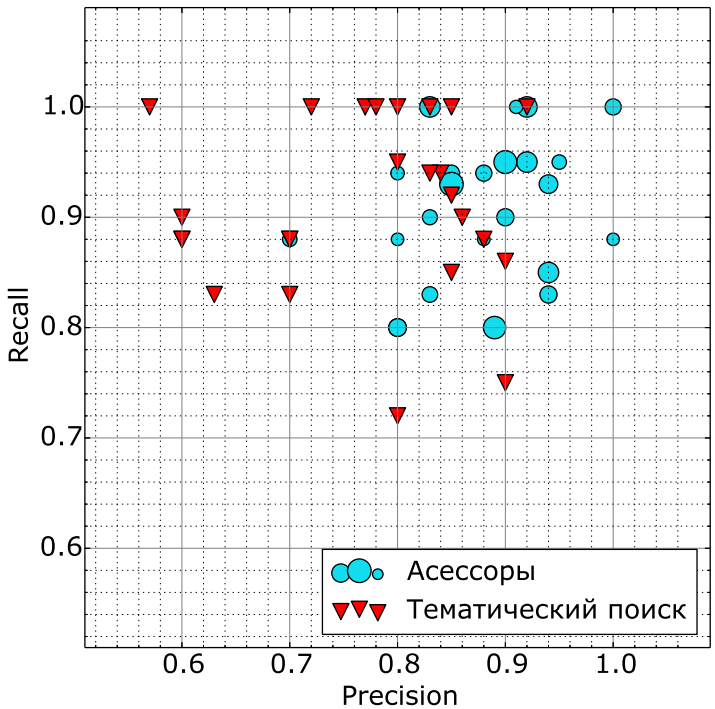
TP (true positive) — найденные релевантные
FP (false positive) — найденные нерелевантные
FN (false negative) — ненайденные релевантные



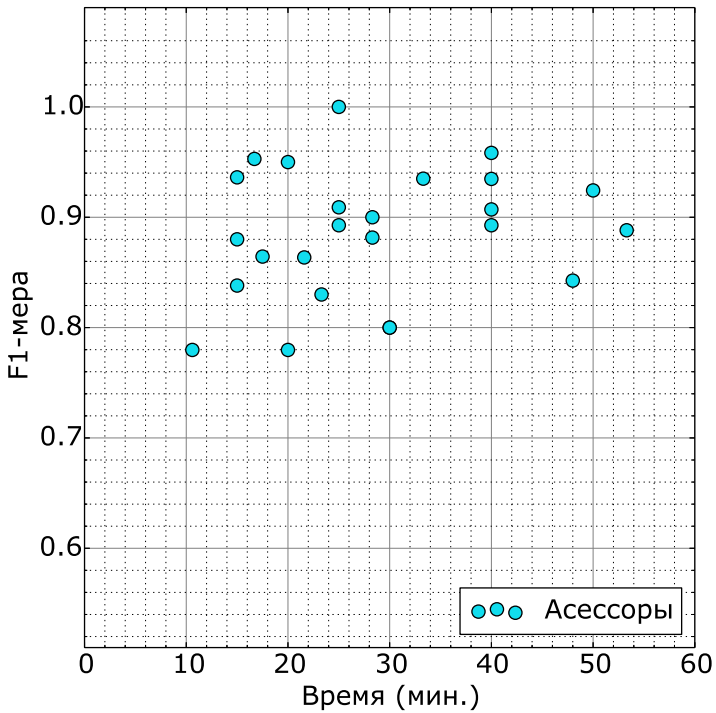
Результаты измерения точности и полноты по запросам

25 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

Выбор модальностей по критериям точности и полноты

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

	асессоры	C	K	TX	CT	CX	CTX	все
Precision@5	0.82	0.63	0.54	0.59	0.74	0.73	0.73	**0.74**
Precision@10	0.87	0.67	0.56	0.58	0.77	0.74	0.75	**0.77**
Precision@15	0.86	0.65	0.53	0.55	0.67	0.67	0.68	**0.68**
Precision@20	0.85	0.64	0.53	0.54	0.66	0.67	0.68	**0.68**
Recall@5	0.78	0.77	0.63	0.69	0.82	0.81	0.82	**0.82**
Recall@10	0.84	0.79	0.64	0.71	0.88	0.82	0.87	**0.88**
Recall@15	0.88	0.82	0.67	0.73	0.90	0.84	0.89	**0.90**
Recall@20	0.88	0.85	0.68	0.74	0.91	0.85	0.89	**0.91**

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

Выбор числа тем по критериям точности и полноты

Теперь используем все 5 модальностей, меняем число тем $|T|$

	асессоры	100	200	300	400	500
Precision@5	0.82	0.61	**0.74**	0.71	0.69	0.59
Precision@10	0.87	0.65	**0.77**	0.72	0.67	0.61
Precision@15	0.86	0.67	**0.68**	0.67	0.65	0.62
Precision@20	0.85	0.64	**0.68**	0.67	0.64	0.60
Recall@5	0.78	0.62	**0.82**	0.80	0.72	0.63
Recall@10	0.84	0.63	**0.88**	0.81	0.75	0.64
Recall@15	0.88	0.67	**0.90**	0.82	0.77	0.67
Recall@20	0.88	0.69	**0.91**	0.85	0.77	0.68

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

Янина А. О., Воронцов К. В. Мультимодальные тематические модели для разведочного поиска в коллективном блоге. JMLDA, 2016.

Открытые проблемы

- **Единственность.** Найти наименьшее число тем, такое, что при заданной разреженности матриц Φ и Θ разложение $\Phi\Theta$ единственно
- **Устойчивость.** Найти регуляризатор, обеспечивающий воспроизводимость тематического моделирования.
- **Интерпретируемость.** Найти регуляризатор, обеспечивающий интерпретируемость тем.
- **Полнота.** Найти регуляризатор, обеспечивающий построение всех тем за один запуск моделирования.
- **Траектория регуляризации.** Разработать метод адаптивного подбора коэффициентов регуляризации
- **Вычислительная сложность.** Сложность EM-алгоритма $O(|D| \cdot |W| \cdot |T|)$. Можно ли ещё быстрее?

Литература

- *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
- *K.Vorontsov, A.Potapenko.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. AIST 2014.
- *K.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.
- *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Yanina.* Non-bayesian additive regularization for multimodal topic modeling of large collections. 2015.
- *K.Vorontsov, A.Potapenko, A.Plavin.* Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.
- *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016. (в печати)
- *M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov.* Mining Ethnic Content Online with Additively Regularized Topic Models. Computación y Sistemas, 2016.
- *А.О.Янина, К.В.Воронцов.* Мультимодальные тематические модели для разведочного поиска в коллективном блоге. LMLDA, 2016.