Математические методы анализа текстов

Тематическое моделирование (часть 1)

К. В. Воронцов
vokov@forecsys.ru

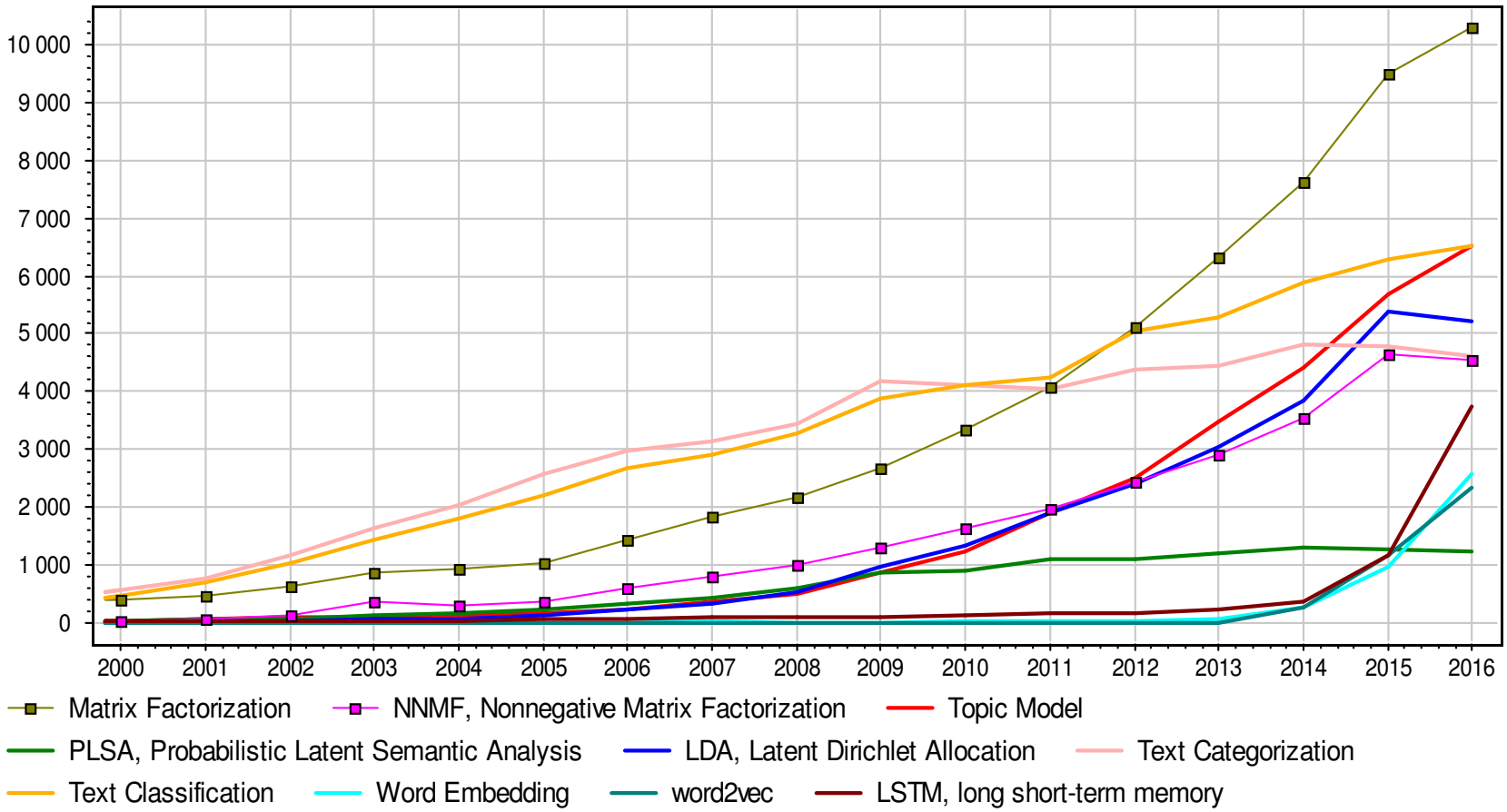
Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Математические методы анализа текстов (ВМиК МГУ) / 2017»

кафедра ММП • 31 марта 2017

Содержание

Тематическое моделирование и смежные области исследований

Динамика цитирования, по данным Google Scholar:



Что такое «тема» в коллекции текстовых документов?

- *тема* — семантически однородный кластер текстов
- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся терминов

Более формально,

- *тема* — условное распределение на множестве терминов, $p(w|t)$ — вероятность (частота) термина w в теме t;
- *тематика* документа — условное распределение $p(t|d)$ — вероятность (частота) темы t в документе d.

Когда автор писал термин w в документе d, он думал о теме t, и мы хотели бы выявить, о какой именно.

Тематическая модель выявляет латентные темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Цели и приложения тематического моделирования

- Выявить тематическую структуру коллекции текстов
- Найти сжатое семантическое описание каждого документа

Приложения:

- Категоризация, классификация, аннотирование, суммаризация, сегментация текстовых документов
- Разведочный информационный поиск (exploratory search)
- Анализ и агрегирование новостных потоков
- Поиск трендов, фронта исследований (research front)
- Поиск экспертов, рецензентов, подрядчиков (expert search)
- Рекомендательные системы
- Аннотирование изображений, видео, музыки
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов

Основные предположения

- Порядок терминов в документе не важен (bag of words)
- Порядок документов в коллекции не важен (bag of docs)
- Каждый термин в документе связан с некоторой темой $t \in T$
- $D \times W \times T$ — дискретное вероятностное пространство
- Коллекция — это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$

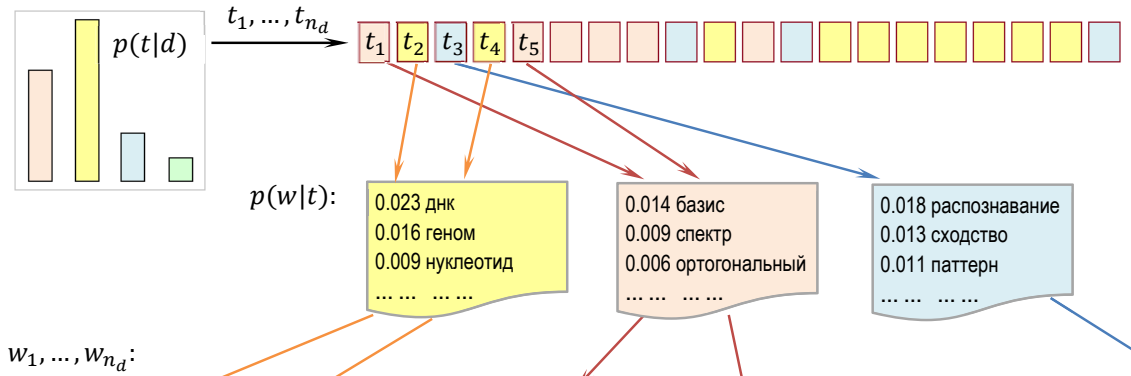
Предварительная обработка текста:

- Лемматизация (русский) или стемминг (английский)
- Выделение терминов (term extraction)
- Выделение именованных сущностей (named entities)
- Удаление стоп-слов и слишком редких слов

Прямая задача — порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D описывает появление терминов w в документах d темами t:

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



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 различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

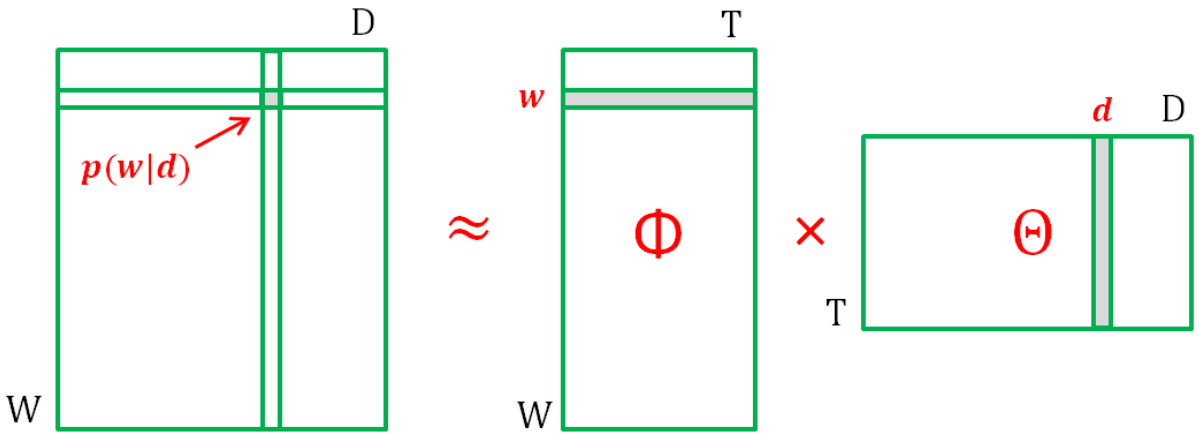
Дано: коллекция текстовых документов

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

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:



Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

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

Максимизация логарифма правдоподобия

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

эквивалентна 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

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

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар
(1865–1963)

Задача стохастического матричного разложения *некорректно поставлена*, так как имеет бесконечно много решений:

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

для невырожденных $S_{T\times T}$ таких, что Φ', Θ' — стохастические.

Регуляризация — стандартный приём, введение новых ограничений или критериев, доопределяющих решение.

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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.

Элементарная интерпретация EM-алгоритма

EM-алгоритм — это чередование E и M шагов до сходимости.

E-шаг: условные вероятности тем $p(t|d,w)$ для всех t,d,w вычисляются через ϕ_{wt}, θ_{td} по формуле Байеса:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}.$$

M-шаг: при $R=0$ частотные оценки условных вероятностей вычисляются суммированием счётчика $n_{tdw} = n_{dw}p(t|d,w)$:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d\in D} n_{tdw}, \qquad n_t = \sum_{w\in W} n_{wt};$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \qquad n_{td} = \sum_{w\in d} n_{tdw}, \qquad n_d = \sum_{t\in T} n_{td}.$$

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. ■

Онлайновый EM-алгоритм (реализован в BigARTM)

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

инициализировать $n_{wt} := 0$ и ϕ_{wt};
для всех $i = 1, \ldots, i_{\max}$ (для больших коллекций $i_{\max} = 1$)
  для всех документов $d \in D$
    инициализировать $\theta_{td} := \frac{1}{|T|}$;
    для всех $j = 1, \ldots, j_{\max}$ (итерации по документу)
      $n_{tdw} := n_{dw} \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$ для всех $w \in d$;
      $\theta_{td} := \underset{t \in T}{\text{norm}}\left(\sum_w n_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$;
    $n_{wt} := \gamma n_{wt} + n_{tdw}$;
    если пора обновить матрицу Φ **то**
      $\phi_{wt} := \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$;

Классические модели PLSA и LDA

PLSA: probabilistic latent semantic analysis [Hofmann, 1999] (вероятностный латентный семантический анализ):

$$R(\Phi, \Theta) = 0.$$

M-шаг — частотные оценки условных вероятностей:

$$\phi_{wt} = \underset{w}{\text{norm}}\big(n_{wt}\big), \qquad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td}\big).$$

LDA: latent Dirichlet allocation (латентное размещение Дирихле):

$$R(\Phi, \Theta) = \sum_{t,w}(\beta_w - 1)\ln \phi_{wt} + \sum_{d,t}(\alpha_t - 1)\ln \theta_{td}.$$

M-шаг — сглаженные частотные оценки с параметрами β_w, α_t:

$$\phi_{wt} = \underset{w}{\text{norm}}\big(n_{wt} + \beta_w - 1\big), \qquad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td} + \alpha_t - 1\big).$$

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

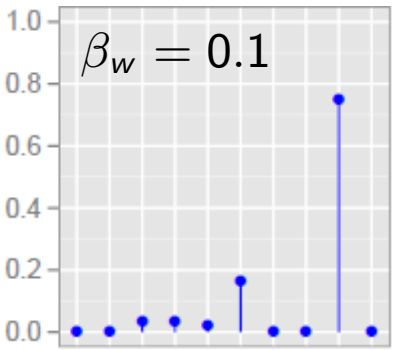
Вероятностная байесовская интерпретация LDA [Blei, 2003]

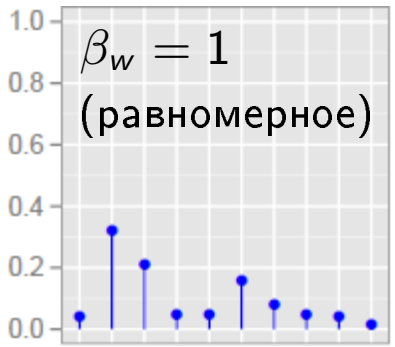
Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

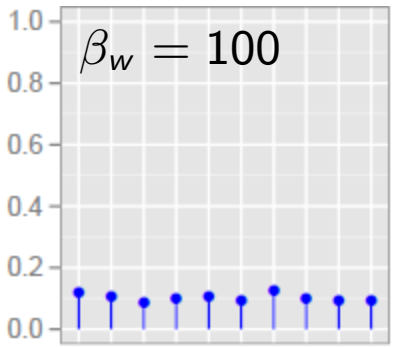
$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \ \beta_t > 0;$$

$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t > 0;$$

Пример. Распределение $\phi \sim \mathrm{Dir}(\beta)$ при $|W| = 10$, $\phi, \beta \in \mathbb{R}^{10}$:







Максимизация апостериорной вероятности для модели LDA

Регуляризатор — логарифм априорного распределения:

$$R(\Phi,\Theta) = \ln \prod_{t\in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d|\alpha) =$$
$$= \sum_{t,w}(\beta_w - 1)\ln \phi_{wt} + \sum_{d,t}(\alpha_t - 1)\ln \theta_{td} + \text{const}$$

M-шаг — сглаженные или слабо разреженные оценки:

$$\phi_{wt} = \underset{w}{\text{norm}}\big(n_{wt} + \beta_w - 1\big), \qquad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td} + \alpha_t - 1\big).$$

при $\beta_w > 1$, $\alpha_t > 1$ — сглаживание,
при $0 < \beta_w < 1$, $0 < \alpha_t < 0$ — слабое разреживание,
при $\beta_w = 1$, $\alpha_t = 1$ априорное распределение равномерно, PLSA.

Почему именно распределение Дирихле?

Плюсы:

- удобно для байесовского вывода, т. к. является сопряжённым к мультиномиальному распределению
- описывает широкий класс распределений на симплексе
- позволяет управлять разреженностью ϕ_{wt} и θ_{td}
- при малых n_{wt}, n_{td} уменьшает переобучение

Минусы:

- не имеет лингвистических обоснований
- не даёт выигрыша против PLSA на больших коллекциях
- слабый разреживатель: запрещены $\beta_w \leqslant 0$, $\alpha_t \leqslant 0$
- слабый регуляризатор: проблема неустойчивости остаётся

Обобщённая не-байесовская интерпретация LDA

Сглаживание распределений по KL-дивергенции:
приблизить $\phi_{wt} \equiv p(w|t)$ к заданным распределениям $\beta_t(w)$,
приблизить $\theta_{td} \equiv p(t|d)$ к заданным распределениям $\alpha_d(t)$:

$$\sum_{t\in T} \tau_t \mathrm{KL}\big(\beta_t(w)\|\phi_{wt}\big) \to \min_{\Phi}; \qquad \sum_{d\in D} \tau_d \mathrm{KL}\big(\alpha_d(t)\|\theta_{td}\big) \to \min_{\Theta}.$$

Взвешенная сумма регуляризаторов:

$$R(\Phi,\Theta) = \sum_{t\in T} \tau_t \sum_{w\in W} \beta_t(w) \ln \phi_{wt} + \sum_{d\in D} \tau_d \sum_{t\in T} \alpha_d(t) \ln \theta_{td}.$$

Формулы M-шага:

$$\phi_{wt} = \underset{w}{\mathrm{norm}}\big(n_{wt} + \underbrace{\tau_t\beta_t(w)}_{\beta_{wt}}\big), \qquad \theta_{td} = \underset{t}{\mathrm{norm}}\big(n_{td} + \underbrace{\tau_d\alpha_d(t)}_{\alpha_{td}}\big).$$

Сглаживание, разреживание и частичное обучение тем

Формулы M-шага:

$$\phi_{wt} = \underset{w}{\text{norm}}\big(n_{wt} + \beta_{wt}\big), \qquad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td} + \alpha_{td}\big).$$

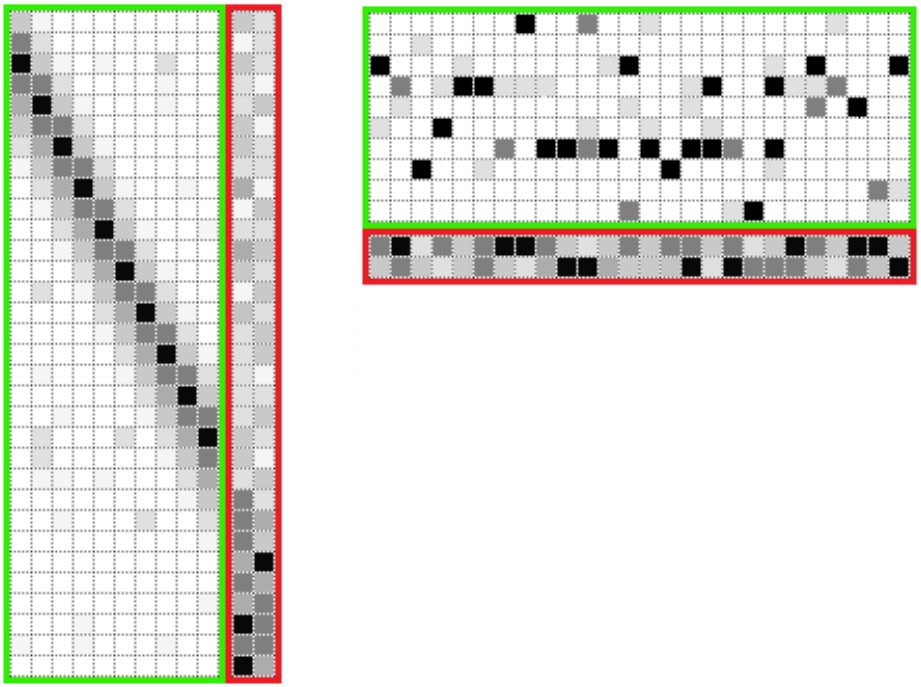
- максимизация KL ведёт к $\beta_{wt} < 0$, $\alpha_{td} < 0$ и разреживанию
- разреживание и сглаживание описывается общей формулой
- можно собирать предметные темы S, сглаживая их по словарю терминов W_0: $\beta_{wt} = p(w)[t \in S]$, $w \in W_0$
- можно использовать *частичное обучение*:
  $\beta_{wt} > 0$ — сглаживание, термин w в «белом списке» темы t
  $\beta_{wt} < 0$ — разреживание, термин w в «чёрном списке» темы t
  $\alpha_{td} > 0$ — сглаживание, тема t в «белом списке» документа d
  $\alpha_{td} < 0$ — разреживание, тема t в «чёрном списке» документа d
- можно разбивать темы на два подмножества, $T = S \sqcup B$:
  S — разреженные *предметные* темы со специальной лексикой
  B — сглаженные *фоновые* темы с общей лексикой языка

Разделение тем на предметные и фоновые

Предметные темы S разреженные, существенно различные

Фоновые темы B содержат слова общей лексики

$\Phi_{W \times T}$ $\Theta_{T \times D}$



Регуляризатор декоррелирования тем

Цель — выделить *лексическое ядро* каждой темы, набор терминов, отличающий её от других тем.

Минимизируем ковариации между вектор-столбцами ϕ_t, ϕ_s:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max.$$

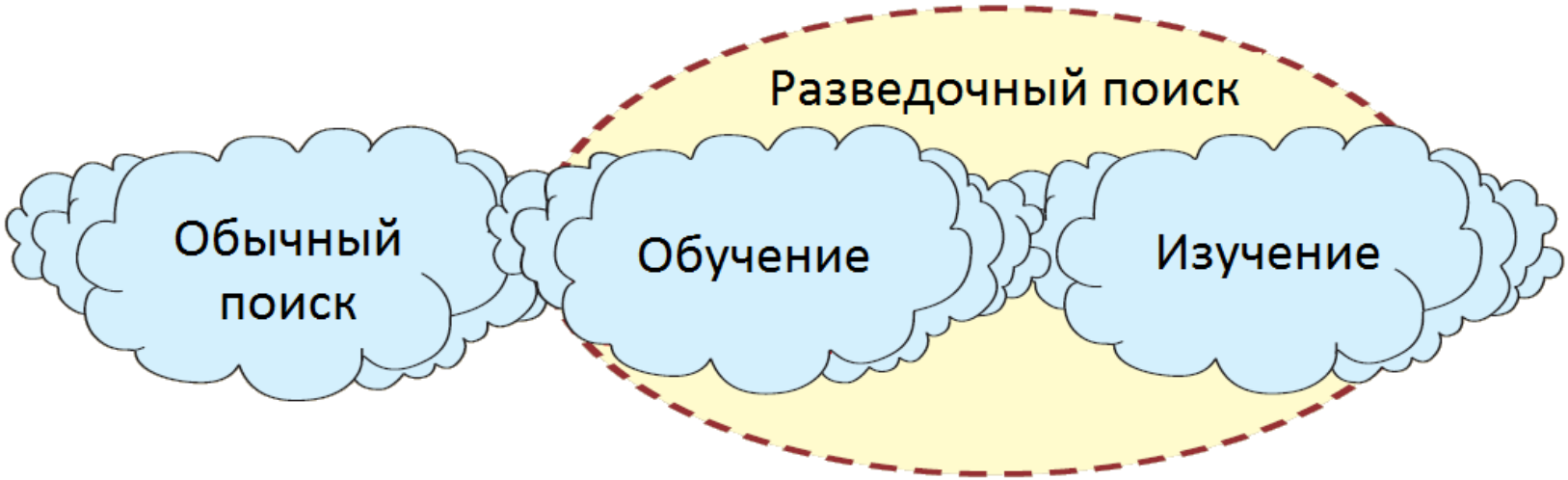
Подставляем, получаем ещё один вариант разреживания — постепенное контрастирование строк матрицы Φ:

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}} \left(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right).$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Концепция разведочного поиска (exploratory search)

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



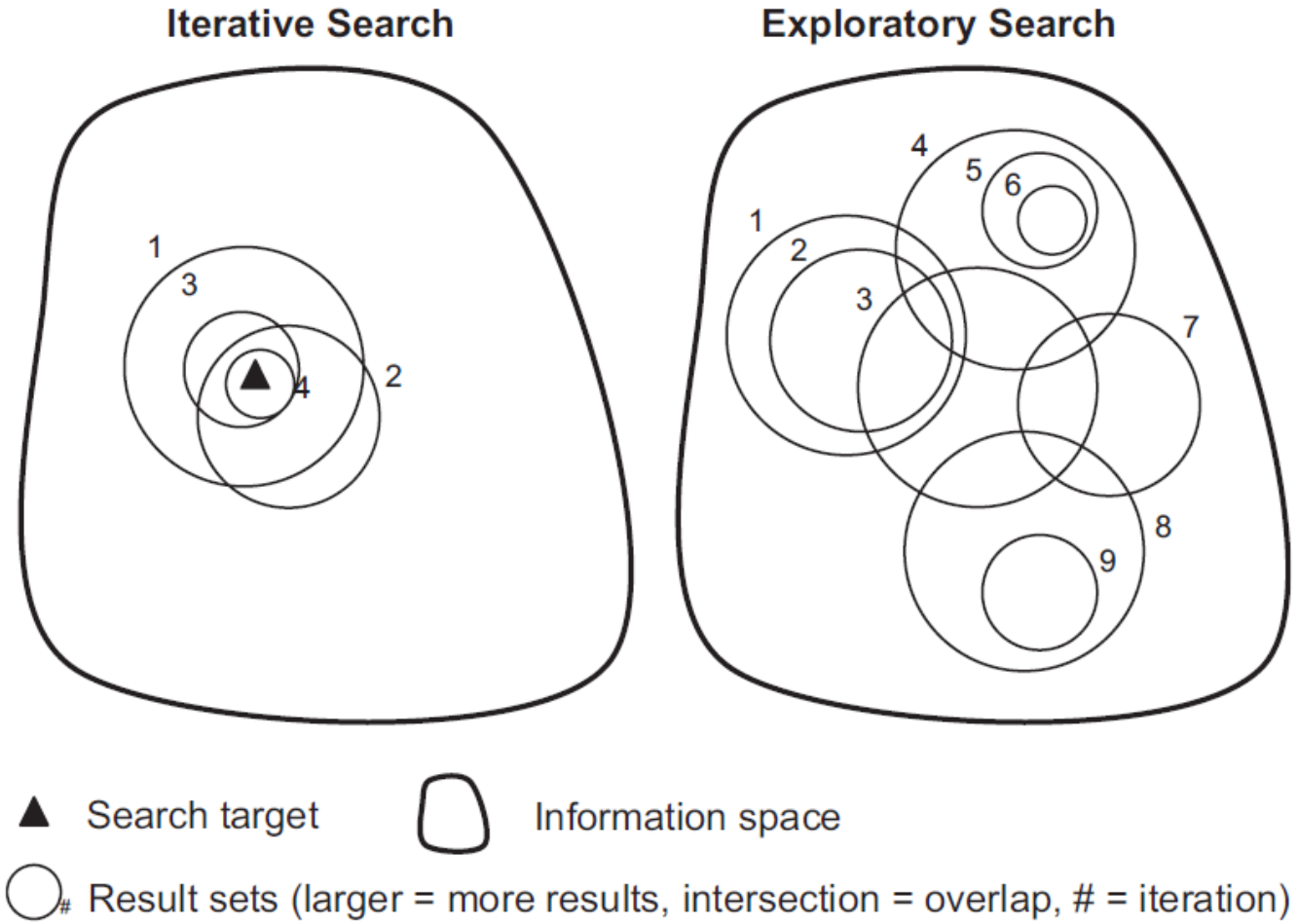
навигация в сети, поиск фактов, упоминаний, конкретных ответов

самообразование, тематический поиск систематизация знаний

исследование, экспертиза, реферирование, 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

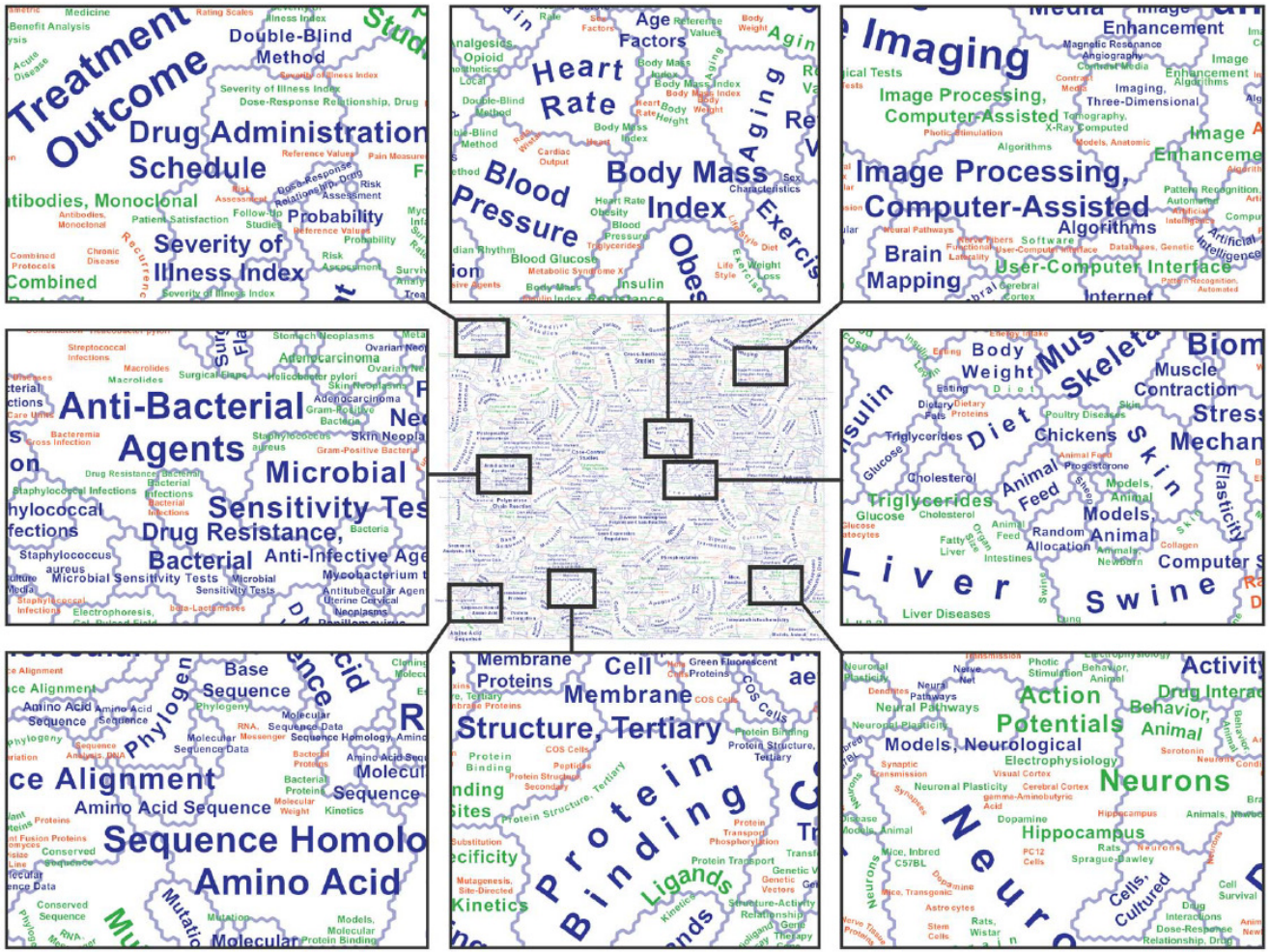
«Сначала крупный план, затем масшабирование и фильтрация, детали по требованию»

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.

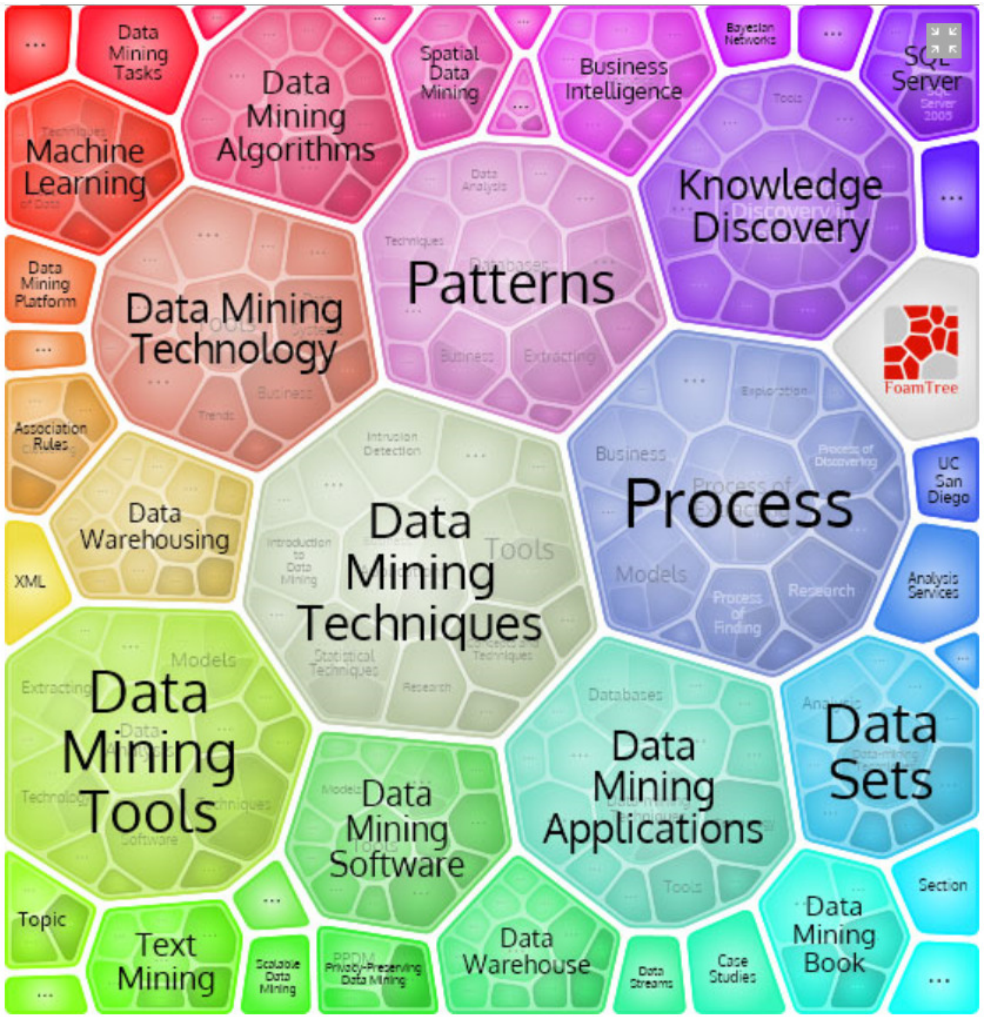
S.Janicke, G.Franzini, M.F.Cheema, G.Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis, 2015.

Пример карты медицинских знаний



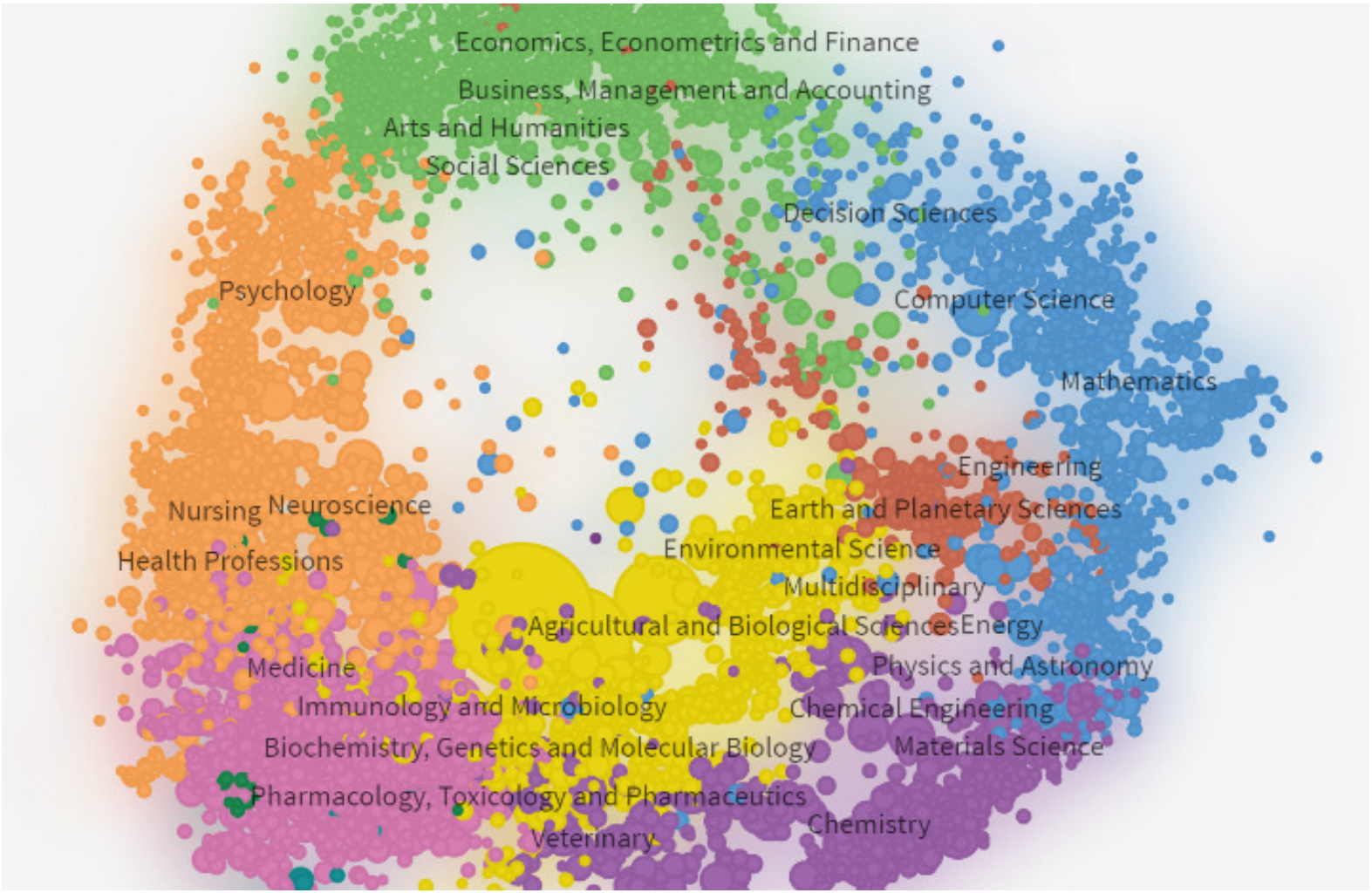
Skupin, Biberstine, Borner. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS ONE, 2013.

Пример иерархической карты области *Data Mining*



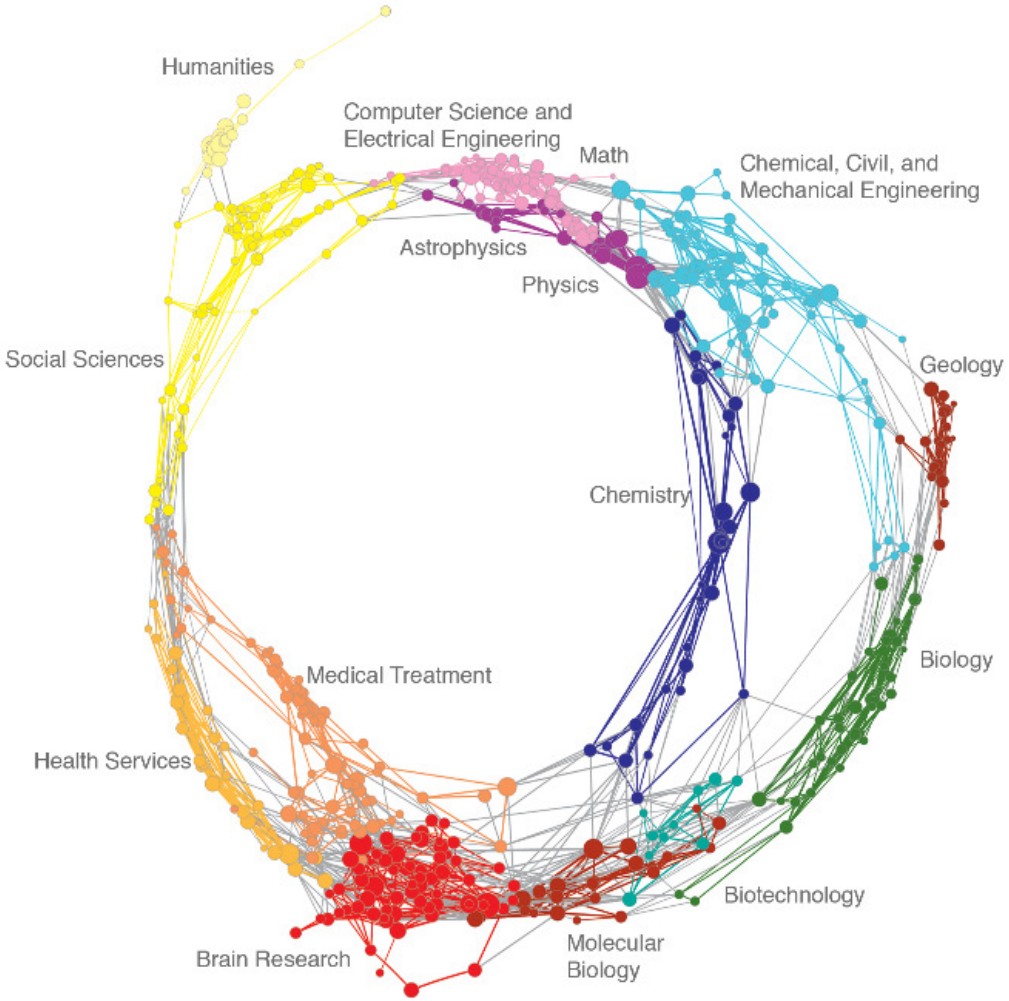
FoamTree: https://carrotsearch.com/foamtree

Пример карты науки



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

Ещё один пример карты науки



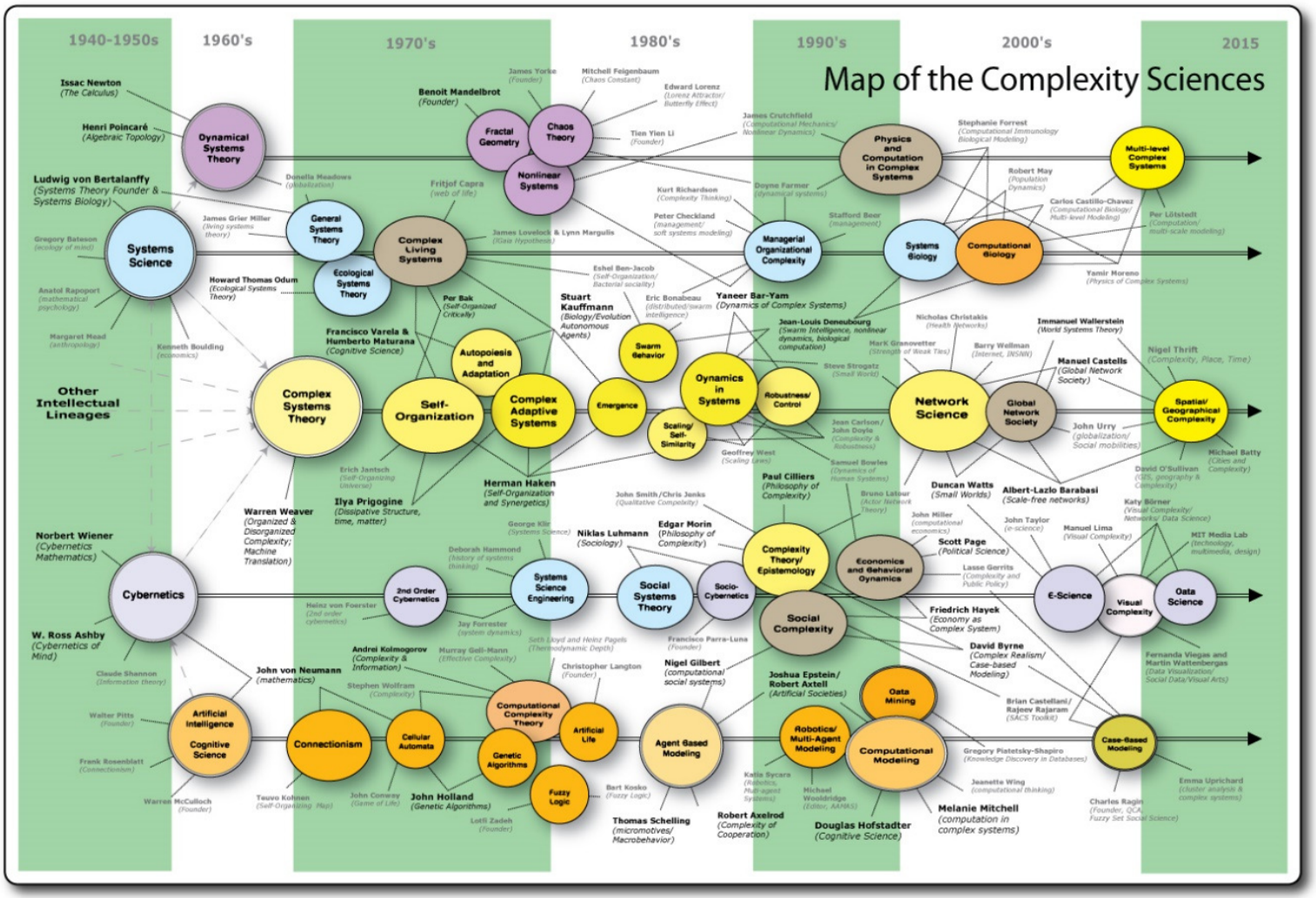
http://scimaps.org

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

Недостатки:

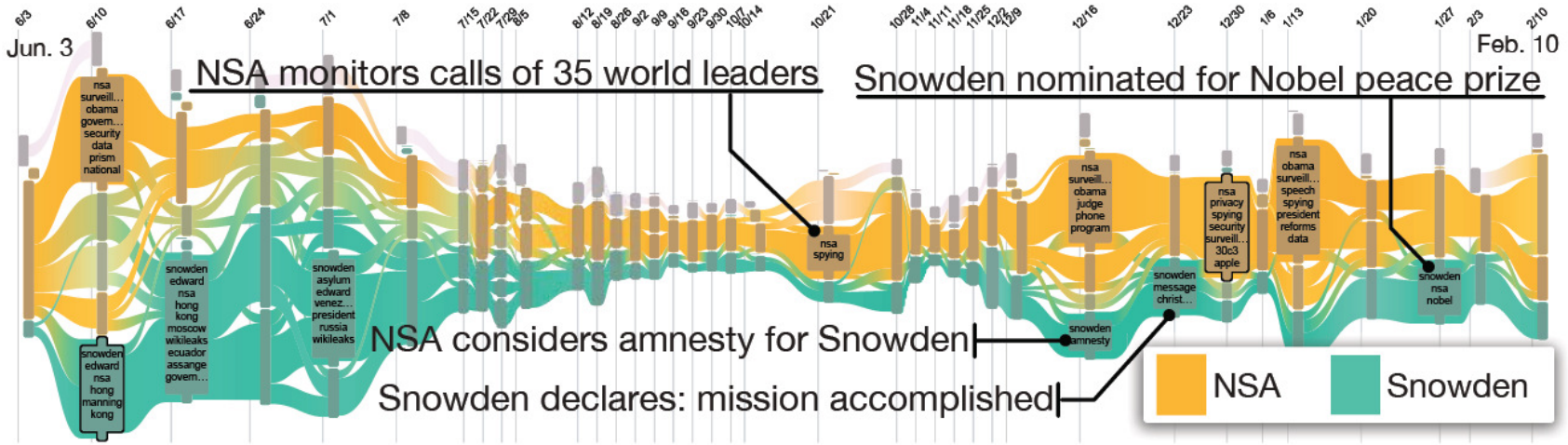
- оси не имеют интерпретации
- искажение сходства при двумерном проецировании

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,
- 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)

- Интерпретируемые оси: время–темы или сложность–темы
- Спектр тем: гуманитарные $\rightarrow$ естественные $\rightarrow$ точные
- Темы делятся на подтемы иерархически
- Интерактивность: реализация мантры Шнейдермана
- При любом масштабе на карте достаточно много текста

http://textvis.lnu.se

Интерактивный обзор 365 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?
- что ещё известно по этим темам?
- какова тематическая структура этой предметной области?
- какие области являются смежными?
- что ещё есть понятного, обзорного, важного, свежего?

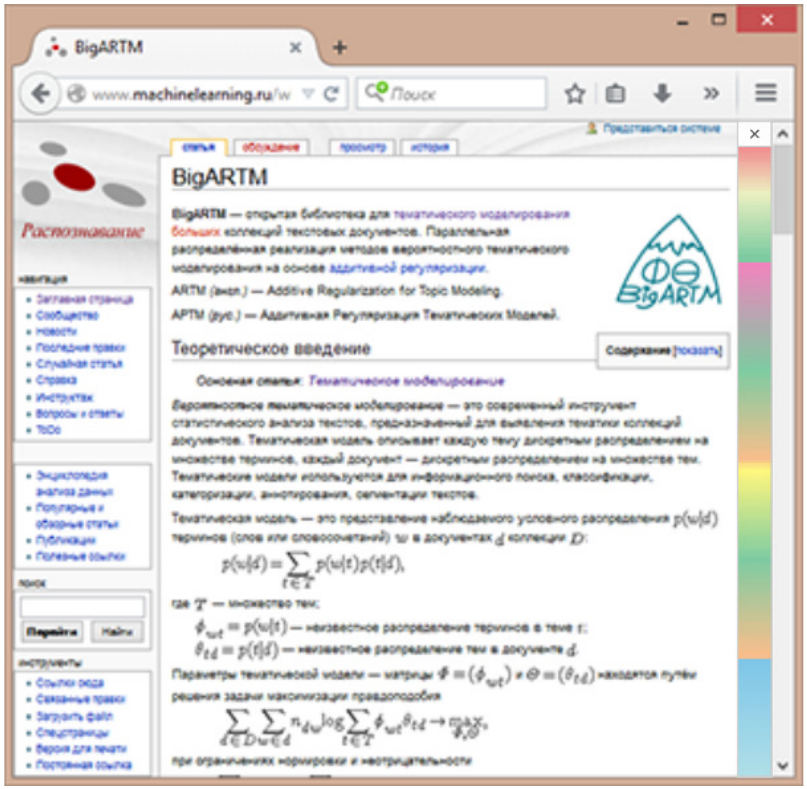
Сценарий поиска:

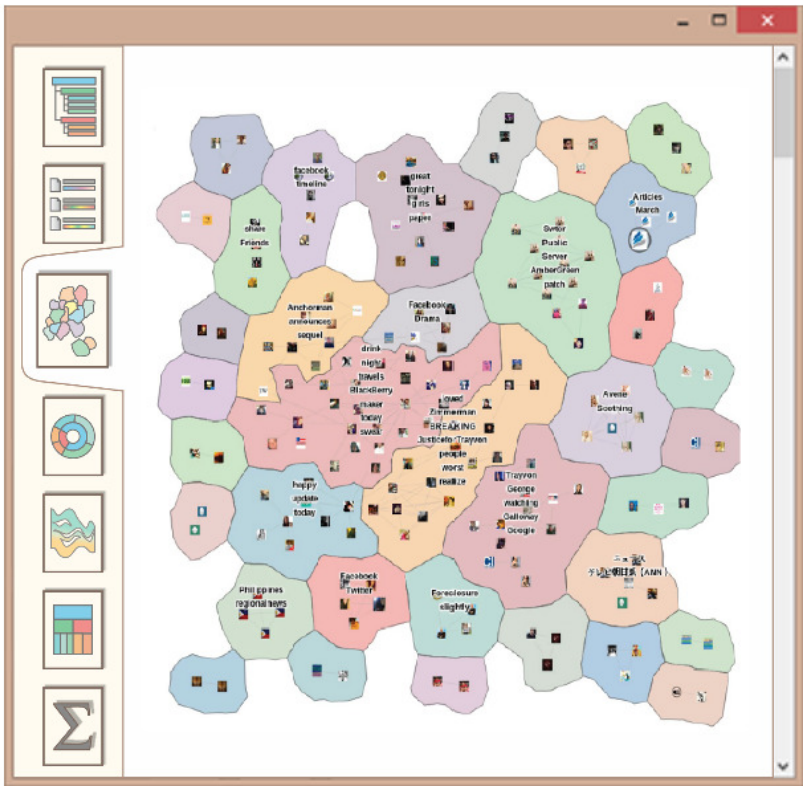
1. имея любой текст под рукой, в любом приложении,
2. получаем картину содержащихся в нём тем-подтем
3. и «дорожную карту» предметной области в целом

Документ-запрос и результат тематического поиска (концепт)

Тематическая сегментация: структура документа-запроса

Дорожная 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.

Теория ARTM позволяет комбинировать тематические модели и строить композитные модели с требуемыми свойствами.

Тематическая модель для разведочного поиска должна быть...

1. **Интерпретируемая**: каждая тема понятна людям
2. **Мультиграммная**: термины-словосочетания неразрывны
3. **Мультимодальная**: авторы, связи, тэги, пользователи,...
4. **Мультиязычная**: для кросс- и много-языкового поиска
5. **Иерархическая**: выявление иерархических связей тем
6. **Динамическая**: выявление истории развития тем
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

Резюме

- Тематическое моделирование — это восстановление латентных тем по коллекции текстовых документов
- Задача сводится к стохастическому матричному разложению
- Задача является некорректно поставленной, так как множество её решений в общем случае бесконечно
- Стандартные методы PLSA и LDA не решают эту проблему
- Аддитивная регуляризация (ARTM) доопределяет задачу и позволяет строить модели с заданными свойствами
- Онлайновый EM-алгоритм хорошо распараллеливается и тематизирует большие коллекции за один проход
- Разведочный информационный поиск — одно из основных перспективных приложений тематического моделирования
- В следующей лекции: регуляризаторы, визуализаторы, иерархии, метрики качества, счастье пользователя