Тематический спектр

Федоряка Дмитрий

Семинар BigARTM

3 мая 2017

Постановка задачи

- $\Phi = \{p(w|t)\}_{w \in W, t \in T}$ — матрица тематической модели;
- Требуется переставить темы так, чтобы рядом стояли семантически близкие темы:

$$\sum_{i=1}^{|T|-1} R[\pi_i, \pi_{i+1}] \to \min_{\pi}$$

- Как определять семантическую близость?
- Как искать перестановку?
- Как оценить качество?

Функции расстояния

$$R[i,j] = \rho(t_i, t_j)$$

- Евклидово расстояние

$$\rho(t,t') = \sqrt{\sum_w \left(\phi_{wt} - \phi_{wt'}\right)^2}$$

- Косинусное расстояние

$$\rho(t,t') = 1 - \frac{\sum_w \phi_{wt} \cdot \phi_{wt'}}{\|\Phi_t\| \|\Phi'_t\|}; \quad \|\Phi_t\| = \sqrt{\sum_w \phi_{wt}^2}$$

Функции расстояния

- Расстояние Хеллингера

$$\rho(t,t') = \sqrt{\sum_w \left(\sqrt{\phi_{wt}} - \sqrt{\phi_{wt'}}\right)^2}$$

- Дивергенция Йенсена-Шеннона

$$\rho(t,t') = \frac{D_{KL}(\Phi_t \| \frac{\Phi_t + \Phi_{t'}}{2}) + D_{KL}(\Phi_{t'} \| \frac{\Phi_t + \Phi_{t'}}{2})}{2}$$

$$D_{KL}(u\|v) = \sum_i u_i \log_2 \frac{u_i}{v_i}$$

- Расстояние Жаккара

$$\rho(t,t') = 1 - \frac{\left|\left\{w | \phi_{wt} < \frac{1}{|W|} \wedge \phi_{wt'} < \frac{1}{|W|}\right\}\right|}{\left|\left\{w | \phi_{wt} < \frac{1}{|W|} \vee \phi_{wt'} < \frac{1}{|W|}\right\}\right|}$$

Решение оптимизационной задачи

$$\sum_{i=1}^{|T|-1} R[\pi_i, \pi_{i+1}] \to \min_{\pi}$$

- Непосредственное решение оптимизационной задачи (легко сводится к задаче комивояжёра, есть очень быстрый и точный эвристический алгоритм [1]);
- Одномерная иерархическая агломеративная кластеризация.
- Многомерное шкалировние (MDS, t-SNE) — плохо.

[1] Helsgaun, K. An effective implementation of the Lin-Kernighan traveling salesman heuristic. // European Journal of Operational Research. 2000

Метрики качества

- Сумма расстояний между соседями (оптимизируемый функционал)

$$NDS(\pi) = \sum_{i=1}^{N-1} D[\pi_i, \pi_{i+1}]$$

- Средний ранг соседа

$$rank(v|u) = \left|\left\{w \in \overline{1,N} \middle| D[w,u] < D[v,u]\right\}\right|$$

$$ANR(\pi) = \frac{1}{2N-2} \sum_{i=1}^{N-1} \left(rank(\pi_{i-1}|\pi_i) + rank(\pi_i|\pi_{i-1})\right)$$

- Сохраняемость порядка в тройках

$$TOC(\pi) = \frac{6}{N(N-1)(N-2)} \sum_{1 \le x < y < z \le N} \left[\max(D[\pi_x, \pi_y], D[\pi_y, \pi_z]) < D[\pi_x, \pi_z]\right]$$

Пользовательские оценки качества

- Сбор оценок
  - Показать тему (каждую K раз)
  - Попросить выбрать из остальных тем несколько близких по смыслу
  - $C_{ij} = \frac{\nu_{ij}+\nu_{ji}}{2K}$, ν_{ij} — сколько раз тема i была указана, как близкая к j.
- Пользовательский штраф

$$UP(\pi) = \sum_{i<j} C_{ij}(|\pi_i^{-1} - \pi_j^{-1}| - 1)$$

- Корелляция

$$UMC(\pi) = -\frac{\sum_{i<j}(D_{ij} - \overline{D})(C_{ij} - \overline{C})}{\sqrt{\sum_{i,j}(D_{ij} - \overline{D})^2}\sqrt{\sum_{i,j}(C_{ij} - \overline{C})^2}};$$

Спектры

учёный, клетка, исследование, исследователь
земля, животное, учёный, животный
ракета, путин, запуск, глава_государство
россия, сирия, исламский_государство, сша
россия, страна, турция, ес
партия, кандидат, журналист, праймериза
россия, украина, крым, решение
закон, законопроект, документ, реклама
россия, страна, российский, ввоз
статья, убийство, задержать, суд
полицейский, полиция, мужчина, автомобиль
самолёт, километр, машина, борт
ребёнок, женщина, мужчина, летний
видео, youtube, ролик, фото
facebook, пользователь, интернет, страница
устройство, смартфон, компания, игра
бренд, модель, компания, обувь
россия, москва, турист, процент
процент, доллар, рубль, нефть
компания, миллиард_рубль, миллиард_доллар, россия
фильм, сериал, актёр, игра_престол
пройти, мероприятие, россия, москва
евро, евровидение, страна, россия
команда, матч, счёт, клуб
спортсмен, допинг, олимпиада, рио

остров, земля, период, территория
растение, япония, раса, более
вид, эволюция, животное, мозг
мозг, нейрон, заболевание, пациент
клетка, музей, стволовой, ткань
клетка, ген, днк, организм
система, материал, задача, дать
квантовый, свет, волна, информация
звезда, галактика, земля, планета
частица, энергия, кварк, взаимодействие
теория, пространство, вселенная, закон
память, задача, например, объект
язык, слово, русский, например
наука, учёный, научный, лекция
книга, фильм, автор, кино
искусство, литература, говорить, мир
век, история, русский, имя
право, власть, век, закон
политический, философия, свобода, идея
социальный, социология, мир, объект
исследование, социальный, группа, наука
город, пространство, социальный, городской
ребёнок, женщина, мужчина, жизнь
война, страна, государство, советский
экономический, экономика, страна, более

Иерархический спектр: постановка задачи

$$NDS(\pi_1) = \sum_{i=1}^{|T_1|-1} D_1[\pi_1[i], \pi_1[j]]$$

$$NDS(\pi_2) = \sum_{i=1}^{|T_2|-1} D_2[\pi_2[i], \pi_2[j]]$$

$$SCC(\pi_1, \pi_2) = \sum_{i_1<j_1} \sum_{i_2<j_2} H[i_1, j_1]H[i_2, j_2]\Big[(\pi_1[i_1] < \pi_1[j_1]) \veebar (\pi_2[i_2] < \pi_1[j_2])\Big]$$

$$\Big(NDS(\pi_1), NDS(\pi_2), SCC(\pi_1, \pi_2)\Big) \to \min_{\pi_1, \pi_2}$$

Иерархический спектр: решение

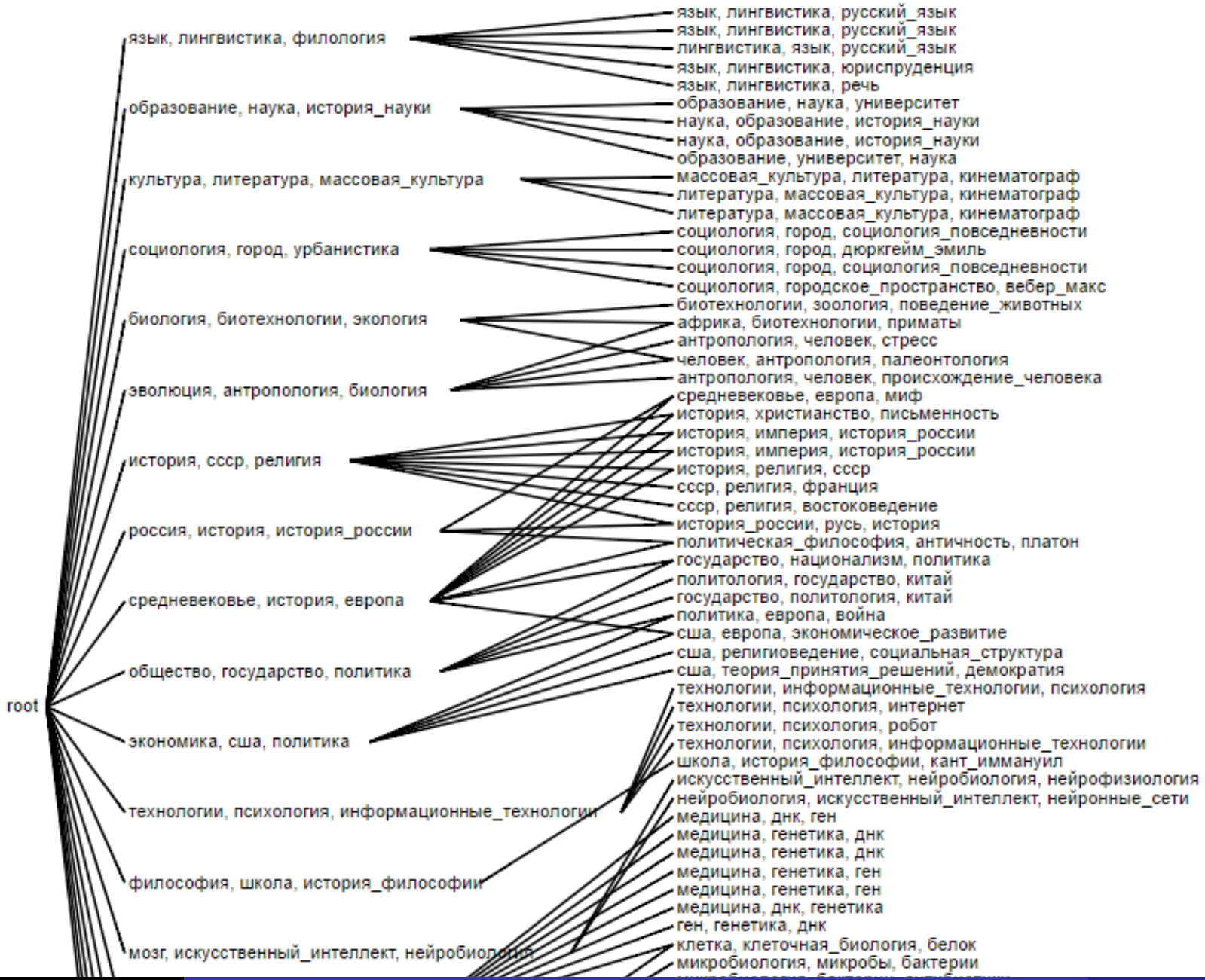
- Домножить $R[i,j]$ на нижнем уровне число $\beta < 1$ для всех тем i, j, имеющих общего родителя.
- Построить спектр для нижнего уровня.
- Зафиксировать перестановку тем на нижнем уровне.
- Переставить темы на верхнем уровне так, чтобы минимизировать число пересечений рёбер (эта задача сводится [2] к задаче целочисленного ЛП, $O(N^2)$ переменных, $O(N^3)$ ограничений.)

[2] Christof T. et al. A branch-and-cut approach to physical mapping of chromosomes by unique end-probes //Journal of Computational Biology. 1997.

Иерархический спектр (Постнаука, 10-30)

Иерархический спектр (lenta.ru, 10-30)

Иерархический спектр (модель Марии Селезневой)



visartm.vdi.mipt.ru

Мне нужна ваша помощь!