Вероятностное тематическое моделирование несбалансированных текстовых коллекций

Панкратов Виктор Владимирович

Московский физико-технический институт
Кафедра интеллектуальных систем

Научный руководитель: д.ф.-м.н. Воронцов К.В.

21/06/2023

Постановка задачи: вероятностная модель

Заданы три множества:
D - множество документов,W - множество слов,T - множество тем
Задано n_{wd} - число вхождений слова w в документ d.

Предположение о порождении коллекции

Появление слова $w \in W$ в документе $d \in D$ описывается двумя матрицами: Φ, Θ.

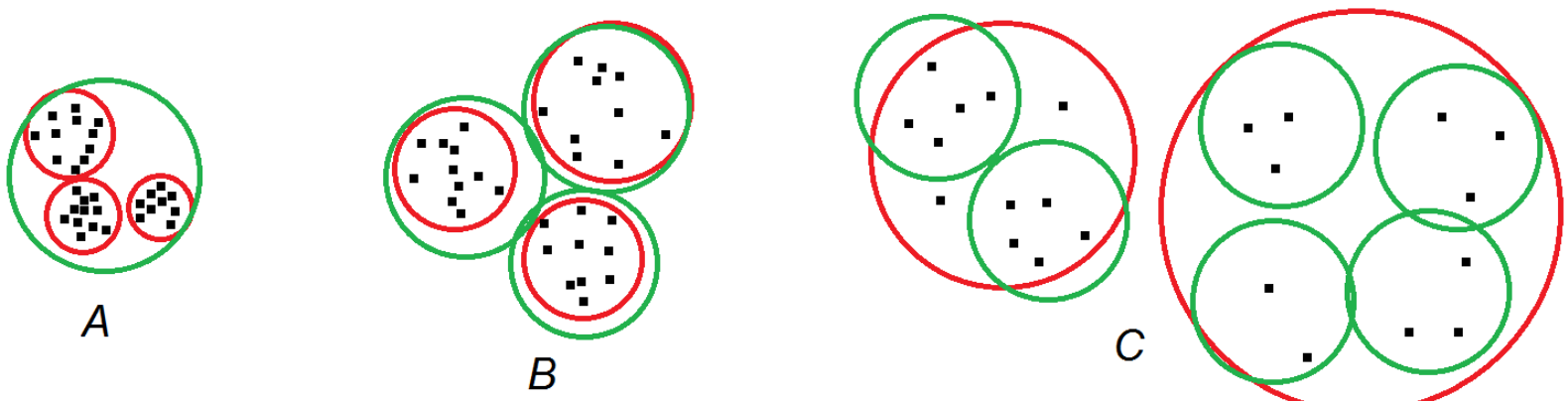
$$\phi_{wt} = p(w|t) \qquad \theta_{td} = p(t|d)$$

Задача: восстановить Φ, Θ.
Критерий качества:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_{i} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta} \qquad (1)$$

максимизация правдоподобия, используется EM-алгоритм

Проблема несбалансированности

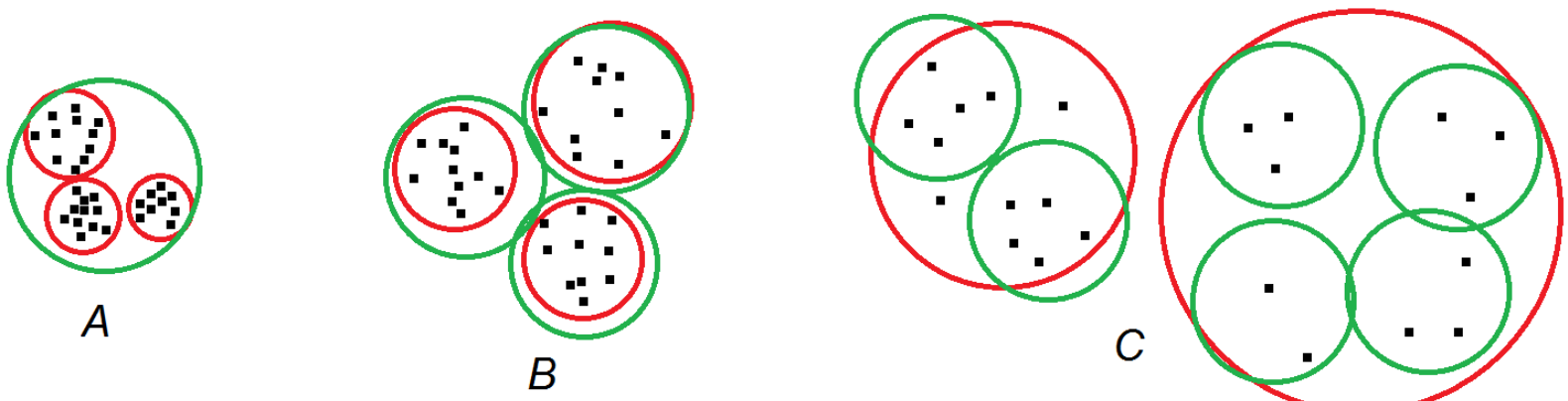


Проблема несбалансированности: максимизация правдоподобия приводит к дроблению крупных тем(A) и слиянию мелких(C).

Цель работы

Предложить и экспериментально проверить решение проблемы несбалансированности с помощью регуляризатора.

Семантическая неоднородность



Гипотеза условной независимости:

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

Проверка - статистика семантической неоднородности.

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

Тема - кластер размерности $|W|$, центр которого - $p(w|t)$.
S_t - удаленность $p(w|d, t)$ от центра кластера.

Регуляризатор семантической неоднородности

Статистика семантической неоднородности

$$\mathrm{S_t} = \mathrm{KL}(\hat{\mathrm{p}}(\mathrm{w},\mathrm{d}|\mathrm{t})||\mathrm{p}(\mathrm{w}|\mathrm{t})\mathrm{p}(\mathrm{d}|\mathrm{t})) = \sum_{\mathrm{d}\in\mathrm{D}}\sum_{\mathrm{w}\in\mathrm{d}} \hat{\mathrm{p}}(\mathrm{w},\mathrm{d}|\mathrm{t})\ln\frac{\hat{\mathrm{p}}(\mathrm{w},\mathrm{d}|\mathrm{t})}{\mathrm{p}(\mathrm{w}|\mathrm{t})\mathrm{p}(\mathrm{d}|\mathrm{t})}$$

Здесь $\hat{\mathrm{p}}$ - частотные оценки вероятности.
Преобразовывая и суммируя по всем темам:

$$\sum_{\mathrm{t}\in\mathrm{T}}\mathrm{S_t} = \sum_{\mathrm{d}\in\mathrm{D}}\sum_{\mathrm{w}\in\mathrm{d}}\left(\sum_{\mathrm{t}\in\mathrm{T}}\hat{\mathrm{p}}(\mathrm{w},\mathrm{d}|\mathrm{t})\right)\ln\frac{\hat{\mathrm{p}}(\mathrm{w}|\mathrm{d})}{\mathrm{p}(\mathrm{w}|\mathrm{d})}$$

Используется регуляризатор, полученный из статистики семантической неоднородности:

$$\mathrm{R} = \sum_{\mathrm{d}\in\mathrm{D}}\sum_{\mathrm{w}\in\mathrm{d}}\beta_{\mathrm{dw}}\mathrm{n_{dw}}\ln\sum_{\mathrm{t}\in\mathrm{T}}\phi_{\mathrm{wt}}\theta_{\mathrm{td}},\quad \beta_{\mathrm{dw}} = \sum_{\mathrm{t}\in\mathrm{T}}\frac{\mathrm{p}(\mathrm{t}|\mathrm{d},\mathrm{w})}{\mathrm{p}(\mathrm{t})} \qquad (2)$$

Сравнение двух моделей

Пусть на одной коллекции построены две тематические модели,

- Φ_1 - матрица вероятностей p(w|t), полученная первой моделью
- Φ_2 - матрица вероятностей p(w|t), полученная второй моделью

Для всех пар i, j проверяются равенства:

$$\arg\min_{k} \ (\mathrm{dist}(\Phi_1[i], \Phi_2[k])) = j \tag{3}$$

$$\arg\min_{k} \ (\mathrm{dist}(\Phi_1[k], \Phi_2[j])) = i \tag{4}$$

Здесь dist – косинусное расстояние.
Взаимно близкие темы: (3),(4) выполнены для некоторых i, j.

Подготовка данных

Для экспериментов использовалась коллекция 20newsgroups[1]. Она преобразовывалась согласно следующему алгоритму

- Составляется матрица n_{dw}
- Удаляются не монотематические документы. Для этого строится произвольная тематическая модель, для каждого документа d считается $t_d = \text{argmax } p(t|d)$ и проверяется $\frac{p(t_d|d)}{p(t_i|d} > 2 \; \forall t_i \neq t_d$
- Для каждого генерируемого документа выбирается его тема
- Документ генерируется как множество случайно выбранных сочетаний из k подряд идущих слов в исходных документах соответствующей темы

[1] http://qwone.com/ jason/20Newsgroups/

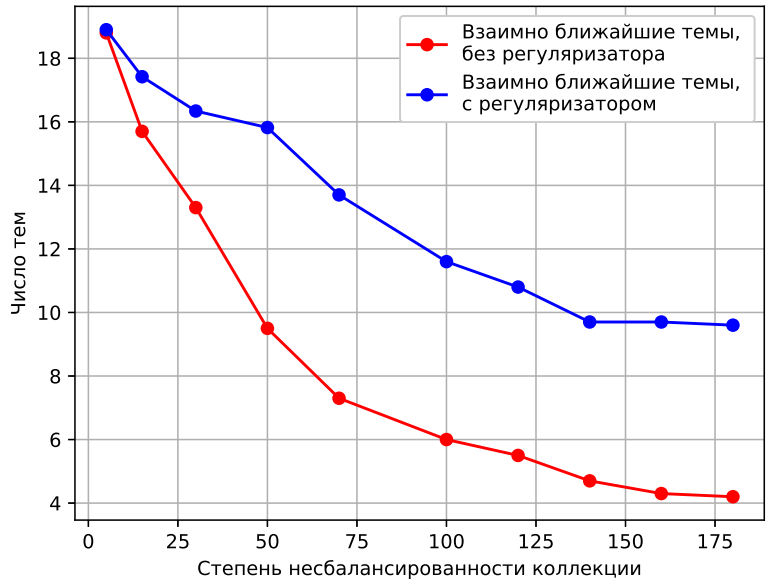
Параметры эксперимента

- Мощность темы - число документов, сгенерированных по этой теме.
- Степень несбалансированности коллекции - отношение максимальной и минимальной мощностей.
- Параметры эксперимента, если не указано иное:
  - Число тем $|\mathrm{T}| = 20$
  - Число генераций коллекции и запусков: 50
  - Число слов в каждом документе $|\mathrm{w} \in (\mathrm{d} \in \mathrm{D})| = 500$
  - Коэффициент регуляризатора декоррелирования: 1
  - Коэффициент регуляризатора семантической неоднородности: 0.6
  - Регуляризатор разреживания: каждые 10 итераций обнуляются 0.1 от всех элементов каждого столбца Φ и каждой строки Θ. Не обнуляются элементы больше 1/10000.

Параметры регуляризации указаны для экспериментов с соответствующими регуляризаторами.

Демонстрация проблемы несбалансированности

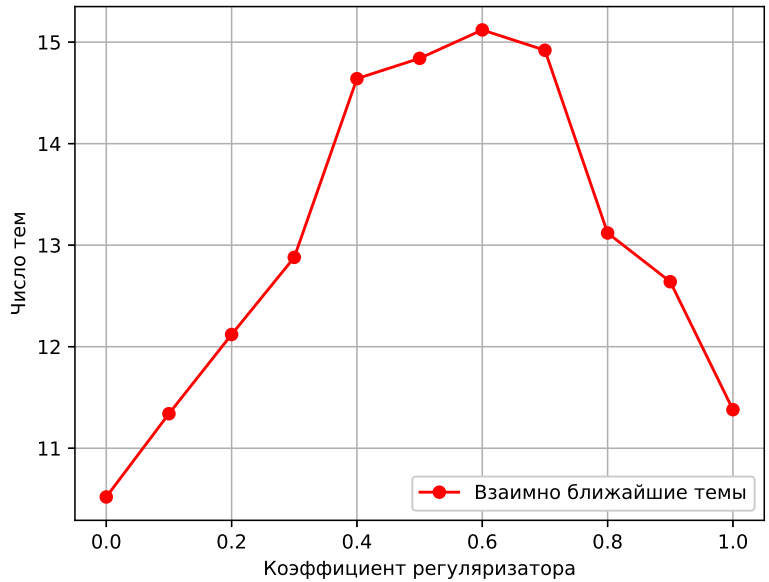
На основе 20newsgroups построены коллекции с разными ρ_{imb}.
Мощность одной темы 3000, остальных - $3000/\rho_{\text{imb}}$.



При возрастании степени несбалансированности уменьшается число взаимно близких тем. Уменьшение сильнее для модели без регуляризатора

Подбор коэффициента регуляризатора семантической неоднородности

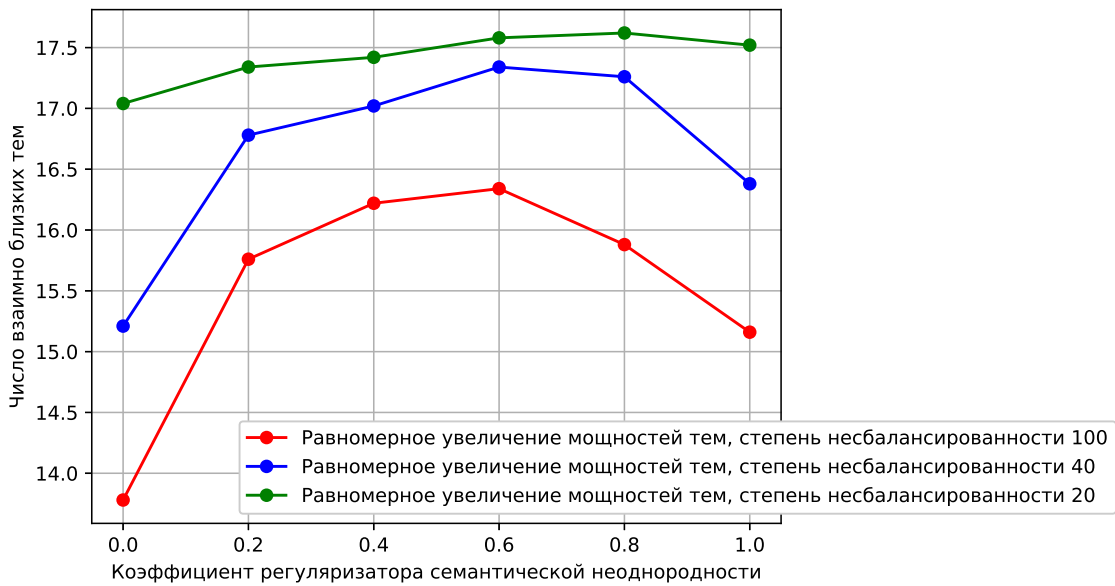
Построена зависимость для коллекции, сгенерированной на основе 20newsgroups. Мощность четырех тем - 500, остальных - 10.



Исходная модель выдала 10 тем взаимно близких к искомым. Подбор коэффициента регуляризации улучшил соответствие найденных и искомых тем.

Подбор коэффициента регуляризатора семантической неоднородности

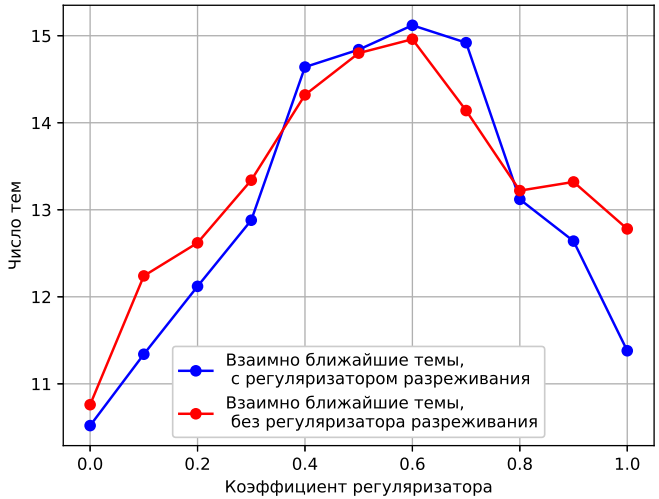
Построена зависимость для коллекций, сгенерированной на основе 20newsgroups. Мощность темы i равна $10 \cdot (\mathrm{ki} + 1)$, $\mathrm{k} \in \{1, 2, 5\}$.

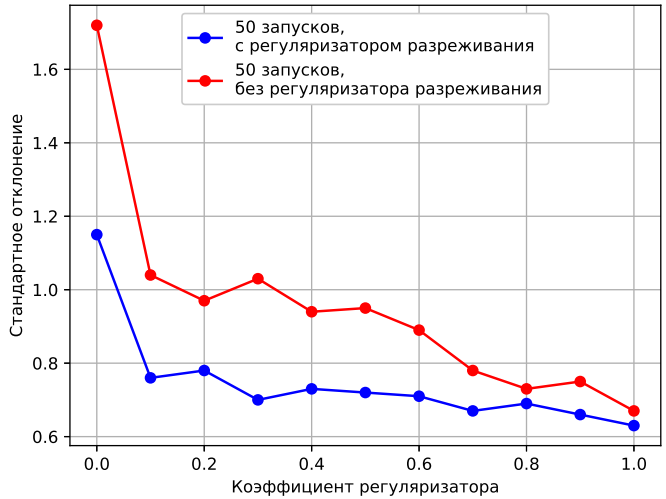


Наибольшее число взаимно близких тем в каждом эксперименте было для модели с регуляризатором с коэффициентом 0.6

Добавление регуляризатора разреживания

Добавим регуляризатор: каждые 10 итераций обнуляются 0.1 от всех элементов каждой строки Φ и каждого столбца Θ. Не обнуляются элементы больше 1/10000. Проводится 50 экспериментов для каждого значения регуляризатора семантической неоднородности:

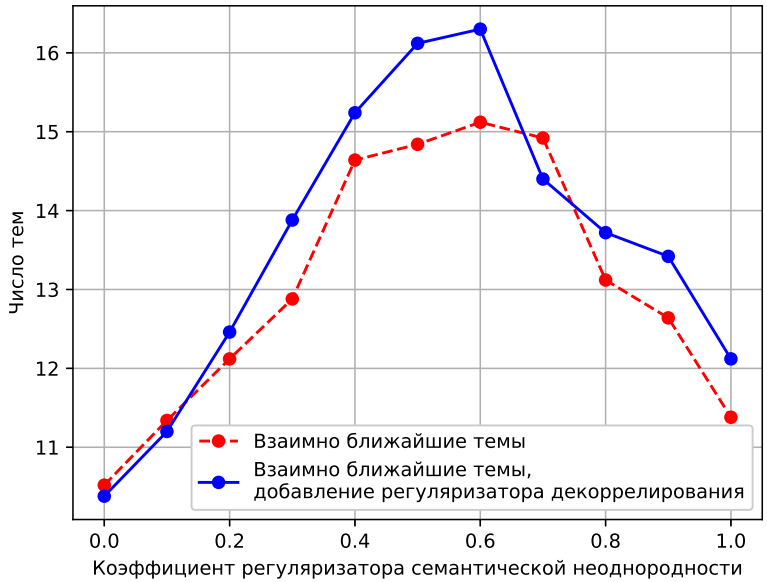




Стандартное отклонение результатов становится меньше, но результаты не улучшаются.

Добавление регуляризатора декоррелирования

В модель добавлен регуляризатор декоррелирования ($\tau = 1$)



Число взаимно близких тем при значении коэффициента 0.6 является одним из максимумов, как и в прошлых экспериментах. При таком значении результаты улучшились.

Слияние и расщепление тем

Построена тематическая модель на коллекции 20newsgroups для 40 тем. 20 тем:оставлены 10 документов. 20 тем: без изменений
Построено две модели - без регуляризатора семантической неоднородности и с добавлением регуляризатора($\tau = 0.6$)

Исходная модель	Модель без регуляризатора
void include null int code	application window use get problem code
menu mouse button application window	application user window server program
firearm legal citizen law crime	wrong society person life crime
wrong society orientation partner relationship	wrong society person life crime

Исходная модель	Модель с регуляризатором
void include null int code	function include void return null
menu mouse button application window	button menu user command server
firearm legal citizen law crime	murder person crime moral society
wrong society orientation partner relationship	orientation wife partner relationship marriage

Слияние тем - две исходные темы соответствуют одной полученной моделью при обработке без регуляризатора.
Добавление регуляризатора устранило проблему

Слияние и расщепление тем

Построена тематическая модель на коллекции 20newsgroups для 40 тем. 20 тем:оставлены 10 документов. 20 тем: без изменений
Построено две модели - без регуляризатора семантической неоднородности и с добавлением регуляризатора

Исходная модель	Модель без регуляризатора
size frame pixel quality image	user graphic support software animation
size frame pixel quality image	display image library animation screen
spirit soul eternal life heaven	death sin church doctrine god
spirit soul eternal life heaven	death faith heaven sin believe

Исходная модель	Модель с регуляризатором
size frame pixel quality image	animation interactive graphic draw image
spirit soul eternal life heaven	soul holy god divine revelation

Расщепление тем - исходная тема соответствует двум полученным моделью при обработке без регуляризатора. Добавление регуляризатора устранило проблему.

Результаты, выносимые на защиту

- Показано, что тематическая несбалансированность коллекции приводит к дроблению крупных и слиянию мелких тем
- Предложен алгоритм устранения проблемы несбалансированности, заключающийся в добавлении регуляризации на основе семантической однородности тем
- Проведены эксперименты, демонстрирующие возможные модификации модели и показывающие, что модель улучшает интерпретируемость результатов обработки несбалансированных коллекций