Вероятностные тематические модели коллекций текстовых документов

К. В. Воронцов
vokov@forecsys.ru

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

ноябрь 2011

Содержание

Определения и обозначения

Дано:

W — словарь, множество слов (терминов);

D — множество (коллекция, корпус) текстовых документов;

каждый $d \in D$ — это последовательность слов из W.

Найти:

T — множество скрытых (латентных) тем;

$p(w|t)$ — распределение на W, задающее тему $t \in T$;

$p(t|d)$ — *тематический профиль* документа, для всех $d \in D$.

Дополнительно:

$p(w|t,y)$, $p(t|y)$, $p(y|t)$ — изменения тематики по годам y;

$p(t|a)$, $p(t|a,y)$ — *тематический профиль* автора a;

$p(t|x)$, $p(t|x,y)$ — *тематический профиль* объекта x, связанного с документами (журнала, конференции, организации, страны);

Цели тематического моделирования (topic modeling)

- Поиск документов по теме, а не по фразе
- Поиск схожих документов по документу
- Визуализация тематической структуры коллекции
- Автоматизация построения рубрикатора
- Мониторинг новых документов по заданной теме
- Поиск экспертов
- Выявление трендов и фронта исследований
- Автоматическая аннотация документов
- Автоматическая суммаризация документов

Типичные приложения:

- Поиск научной информации по коллекции статей
- Анализ новостных потоков
- Рубрикация коллекций изображений, видео, музыки
- Аннотация генома и другие задачи биоинформатики

Стандартные гипотезы тематического моделирования

1. Порядок документов в коллекции не важен
2. Порядок слов в документе не важен (bag of words)
3. Слова, встречающиеся «почти во всех» документах, не важны
4. Слово в разных формах — это одно и то же слово

Предварительная обработка текстов:

- Приведение всех слов к нормальной форме (лемматизация или стемминг)
- Выделение терминов (term extraction) и выделение словосочетаний (key phrase extraction); (сводятся к задачам классификации или ранжирования)
- Удаление стоп-слов $w \in W$: $\#\{d : w \in d\} \geqslant \alpha|D|$, $\alpha \sim 0.05 \ldots 0.5$

Основная вероятностная гипотеза:
коллекция документов — i.i.d. выборка $\big\{(d, w) : d \in D,\ w \in d\big\}$.

Униграммная модель порождения текста

Вероятностная модель документа d — это вероятность реализации цепочки слов документа $W_d = \{w_1, \ldots, w_{n_d}\}$:

$$p(W_d|d) = \prod_{w \in d} p(w|d)^{n_{dw}},$$

$p(w|d)$ — неизвестное мультиномиальное распределение на W;
n_{dw} — число вхождений слова w в документ d;
n_d — длина документа d.

Принцип максимума правдоподобия:

$$\ln \prod_{d \in D} p(W_d|d) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) \to \max_{\{p(w|d)\}}$$

при ограничениях нормировки

$$\sum_{w \in W} p(w|d) = 1, \quad d \in D.$$

Униграммная модель порождения текста

Лагранжиан $\mathscr{L}$ — сумма независимых слагаемых по документам:

$$\mathscr{L} = \sum_{d \in D} \underbrace{\sum_{w \in d} n_{dw} \ln p(w|d) - \lambda_d \left(\sum_{w \in W} p(w|d) - 1 \right)}_{\mathscr{L}(d)}.$$

$$\frac{\partial \mathscr{L}}{\partial p(w|d)} = \frac{\partial \mathscr{L}(d)}{\partial p(w|d)} = n_{dw} \frac{1}{p(w|d)} - \lambda_d = 0.$$

Умножим обе части равенства на $p(w|d)$ и просуммируем по w:

$$\sum_{w \in W} \lambda_d p(w|d) = \sum_{w \in W} n_{dw} \;\Rightarrow\; \lambda_d = n_d,$$

где n_d — длина документа d.

Получим (тривиальную) оценку максимума правдоподобия:

$$\hat{p}(w|d) = \frac{n_{dw}}{n_d} \quad \left(\text{сравните: по определению } p(w|d) = \frac{p(d,w)}{p(d)} \right).$$

Униграммная модель порождения текста

Недостатки униграммной модели:

- тематика не выявляется
- число $|W| \cdot |D|$ оцениваемых параметров $p(w|d)$ линейно зависит от $|D|$ — числа документов в коллекции
- зависимости между документами не учитываются

Эти недостатки устраняются в модели смеси униграмм:

Nigam, McCallum, Thrun, Mitchell.
Text classification from labeled and unlabeled documents using EM.
Journal of Machine Learning, 2000, 39(2–3): 103–134

Модель смеси униграмм [Nigam и др., 2000]

1. Документ $W_d = \{w_1, \ldots, w_{n_d}\}$ — это смесь униграмм:
$$p(W_d|d) = \sum_{t \in T} p(t) p(W_d|d,t).$$
2. Документы, относящиеся к разным темам $t \in T$, описываются разными униграммными моделями:
$$p(W_d|d,t) = \prod_{w \in d} p(w|d,t)^{n_{dw}}.$$
3. **Гипотеза условной независимости:** $p(w|d,t) = p(w|t)$ (распределения слов связаны с темами, а не с документами)

Тематическая модель *смеси униграмм:*
$$p(W_d|d) = \sum_{t \in T} p(t) \prod_{w \in d} p(w|t)^{n_{dw}}.$$

Преимущества и недостатки модели смеси униграмм

Преимущества:

- Модель позволяет выявлять тематику
- Число параметров $|T| + |W| \cdot |T|$ не зависит от $|D|$

Недостатки:

- Слова в документе независимы и относятся к одной теме t:

$$p(W_d|d,t) \equiv p(w_1,\ldots,w_{n_d}|t) = p(w_1|t)\cdots p(w_{n_d}|t)$$

- Вместо *тематического профиля* $p(t|d)$ вычисляется другая (хотя и похожая) величина. По формуле Байеса:

$$H(t|d) \equiv p(t|W_d,d) = \frac{p(t)p(W_d|d,t)}{p(W_d|d)} = \frac{p(t)\prod\limits_{w\in d} p(w|t)^{n_{dw}}}{\sum\limits_{s\in T} p(s)\prod\limits_{w\in d} p(w|s)^{n_{dw}}}$$

Обучение модели смеси униграмм

Принцип максимума правдоподобия:

$$\ln \prod_{d\in D} p(W_d|d) = \ln \prod_{d\in D} \sum_{t\in T} p(t) \prod_{w\in d} p(w|t)^{n_{dw}} \to \max_{\{p(t),p(w|t)\}};$$

при ограничениях $\sum_{w\in W} p(w|t) = 1,\ t \in T; \quad \sum_{t\in T} p(t) = 1.$

Лагранжиан:

$$\mathscr{L} = \sum_{d\in D} \ln \underbrace{\sum_{t\in T} p(t) \overbrace{\prod_{w\in d} p(w|t)^{n_{dw}}}^{p(W_d|d,t)}}_{p(W_d|d)} -$$

$$- \sum_{t\in T} \lambda_t \left(\sum_{w\in W} p(w|t) - 1 \right) - \mu \left(\sum_{t\in T} p(t) - 1 \right).$$

Оценка максимального правдоподобия для $p(t)$

$$\mathscr{L} = \sum_d \ln \underbrace{\sum_t p(t) \overbrace{\prod_{w \in d} p(w|t)^{n_{dw}}}^{p(W_d|d,t)}}_{p(W_d|d)} - \sum_t \lambda_t \left(\sum_w p(w|t) - 1 \right) - \mu \left(\sum_t p(t) - 1 \right).$$

$$\frac{\partial \mathscr{L}}{\partial p(t)} = \sum_{d \in D} \frac{1}{p(W_d|d)} p(W_d|d,t) - \mu = 0.$$

Умножим обе части равенства на $p(t)$ и просуммируем по t:

$$\mu \sum_{t \in T} p(t) = \sum_{d \in D} \sum_{t \in T} \frac{p(W_d|d,t)p(t)}{p(W_d|d)} \quad \Rightarrow \quad \mu = |D|.$$

Если умножить, но не суммировать:

$$\mu p(t) = \sum_{d \in D} \frac{p(W_d|d,t)p(t)}{p(W_d|d)} = \sum_{d \in D} H(t|d) \quad \Rightarrow \quad p(t) = \frac{\sum\limits_{d \in D} H(t|d)}{|D|}.$$

Оценка максимального правдоподобия для $p(w|t)$

$$\mathscr{L} = \sum_d \ln \underbrace{\sum_t p(t) \overbrace{\prod_{w \in d} p(w|t)^{n_{dw}}}^{p(W_d|d,t)}}_{p(W_d|d)} - \sum_t \lambda_t \left(\sum_w p(w|t) - 1 \right) - \mu \left(\sum_t p(t) - 1 \right).$$

$$\frac{\partial \mathscr{L}}{\partial p(w|t)} = \sum_{d \in D} \frac{1}{p(W_d|d)} n_{dw} \frac{p(W_d|d,t)}{p(w|t)} - \lambda_t = 0.$$

Умножим обе части равенства на $p(w|t)$, просуммируем по w:

$$\lambda_t \sum_w p(w|t) = \sum_{d,w} n_{dw} \frac{p(W_d|d,t)}{p(W_d|d)} \quad \Rightarrow \quad \lambda_t = \sum_{d,w} n_{dw} \frac{H(t|d)}{p(t)}.$$

Если умножить, но не суммировать:

$$\lambda_t p(w|t) = \sum_{d \in D} n_{dw} \frac{H(t|d)}{p(t)} \quad \Rightarrow \quad p(w|t) = \frac{\sum_{d \in D} n_{dw} H(t|d)}{\sum_{d \in D} \sum_{u \in W} n_{du} H(t|d)}.$$

EM-алгоритм (оценки максимума правдоподобия)

- **Инициализировать** $H(t|d)$;
- **M-шаг:** оценить параметры модели

$$p(t) := \frac{\sum\limits_{d \in D} H(t|d)}{|D|} \quad \text{для всех } t \in T;$$

$$p(w|t) := \frac{\sum\limits_{d \in D} H(t|d) n_{dw}}{\sum\limits_{d \in D} H(t|d) \sum\limits_{u \in d} n_{du}} \quad \text{для всех } w \in W,\ d \in D;$$

- **E-шаг:** вычислить профили документов

$$H(t|d) := \frac{p(t) \prod\limits_{w \in d} p(w|t)^{n_{dw}}}{\sum\limits_{s \in T} p(s) \prod\limits_{w \in d} p(w|s)^{n_{dw}}} \quad \text{для всех } d \in D,\ t \in T;$$

EM-алгоритм (оценки максимума апостериорной вероятности)

- **Инициализировать** $H(t|d)$;
- **M-шаг:** оценить параметры модели

$$p(t) := \frac{1 + \sum_{d \in D} H(t|d)}{|T| + |D|} \quad \text{для всех } t \in T;$$

$$p(w|t) := \frac{1 + \sum_{d \in D} H(t|d) n_{dw}}{|W| + \sum_{d \in D} H(t|d) \sum_{u \in d} n_{du}} \quad \text{для всех } w \in W,\ d \in D;$$

- **E-шаг:** вычислить скрытые профили документов

$$H(t|d) := \frac{p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}{\sum_{s \in T} p(s) \prod_{w \in d} p(w|s)^{n_{dw}}} \quad \text{для всех } d \in D,\ t \in T;$$

Лирическое отступление: оценки ML и MAP

Оценки параметра θ распределения $p(x|\theta)$ по выборке $\{x_i\}_{i=1}^n$.

1. Метод *максимума правдоподобия* (ML, maximum likelihood):

$$\prod_{i=1}^{n} p(x_i|\theta) \to \max_\theta;$$

$$\sum_{i=1}^{n} \log p(x_i|\theta) \to \max_\theta.$$

2. Метод *максимума апостериорной вероятности* (MAP, maximum a posteriori probability):

$$\left(\prod_{i=1}^{n} p(x_i|\theta)\right) p(\theta) \to \max_\theta;$$

$$\sum_{i=1}^{n} \log p(x_i|\theta) + \log p(\theta) \to \max_\theta.$$

Инициализация $H(t|d)$ и частичное обучение

1. Использование априорной классификации документов $y_{dt} = \big[$ документ d относится к теме $t\big]$:

$$H(t|d) := \frac{y_{dt}}{\sum\limits_{s \in T} y_{ds}}.$$

2. Если размечена только часть коллекции $D^\ell \subset D$, то
— на первом M-шаге суммируем только по $d \in D^\ell$,
— на первом E-шаге вычисляем $H(t|d)$ для всех $d \in D$;
— далее EM-алгоритм выполняется как обычно,
на E-шаге вычисляем $H(t|d)$ для всех $d \in D$ либо $D \setminus D^\ell$.

3. Если $|D^\ell| \ll |D|$, то вес неразмеченных документов $\lambda < 1$:

$$\Lambda_d = \begin{cases} 1, & d \in D^\ell; \\ \lambda, & d \notin D^\ell. \end{cases}$$

EM-алгоритм (с частичным обучением)

- **Инициализировать** $H(t|d)$ для всех $d \in D^\ell$, $t \in T$;
- **M-шаг:** оценить параметры модели

$$p(t) := \frac{1 + \sum_{d \in D} \Lambda_d H(t|d)}{|T| + |D|} \quad \text{для всех } t \in T;$$

$$p(w|t) := \frac{1 + \sum_{d \in D} \Lambda_d H(t|d) n_{dw}}{|W| + \sum_{d \in D} \Lambda_d H(t|d) \sum_{u \in d} n_{du}} \quad \text{для всех } w \in W, \ d \in D;$$

- **E-шаг:** вычислить скрытые профили документов

$$H(t|d) := \frac{p(t) \prod_{w \in d} p(w|t)^{n_{dw}}}{\sum_{s \in T} p(s) \prod_{w \in d} p(w|s)^{n_{dw}}} \quad \text{для всех } d \in D, \ t \in T;$$

Основные недостатки униграммной модели

Предположение, что для каждого документа $d = \{w_1, \ldots, w_{n_d}\}$

$$p(W_d|d,t) \equiv p(w_1, \ldots, w_{n_d}|t) = p(w_1|t) \cdots p(w_{n_d}|t)$$

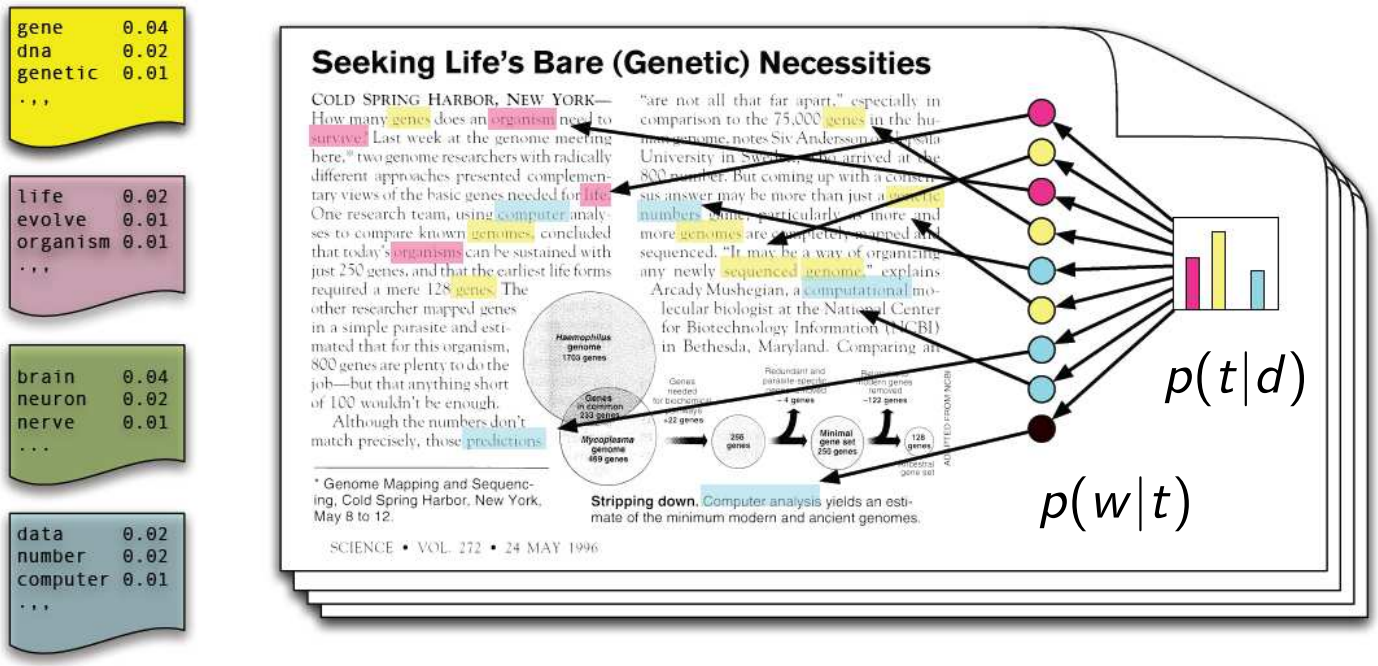
фактически означает, что

- все слова документа относятся к одной и той же теме t;
- внутри темы t слова появляются независимо.

Это довольно сильные ограничения. Они снимаются в последующих вероятностных тематических моделях.

Вероятностная тематическая модель порождения текстов

1: задать распределение $p(w|t)$ для каждой темы $t \in T$;
2: задать распределение $p(t|d)$ для каждого документа $d \in D$;
3: **для всех** документов $d \in D$
4: **для всех** слов w в документе d
5: выбрать тему t из $p(t|d)$;
6: выбрать слово w из $p(w|t)$;



Вероятностный латентный семантический анализ
Probabilistic Latent Semantic Analysis, PLSA [Hofmann, 1999]

- **Вероятностная модель коллекции документов:**

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

(теперь каждое слово связано со своей темой).

- **Гипотеза условной независимости:** $p(w|d,t) = p(w|t)$
(распределения слов связаны с темами, а не с документами)

Задача максимизации правдоподобия по $p(t)$, $p(d|t)$, $p(w|t)$:

$$\mathscr{L} = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln p(d,w) \to \max$$

при ограничениях нормировки

$$\sum_{t \in T} p(t) = 1; \quad \sum_{d \in D} p(d|t) = 1; \quad \sum_{w \in W} p(w|t) = 1.$$

Особенности, преимущества, недостатки модели PLSA

- Симметричность модели относительно $d \rightleftarrows w$:

$$\begin{aligned} p(d,w) &= \sum_{t\in T} p(t)\,p(d|t)\,p(w|t); \\ &= \sum_{t\in T} p(d)\,p(t|d)\,p(w|t); \\ &= \sum_{t\in T} p(w)\,p(t|w)\,p(d|t). \end{aligned}$$

- *Тематические профили* вычисляются по формуле Байеса:

$$p(t|d) = \frac{p(d|t)\,p(t)}{\sum_{s\in T} p(d|s)\,p(s)}; \qquad p(t|w) = \frac{p(w|t)\,p(t)}{\sum_{s\in T} p(w|s)\,p(s)}.$$

- Нет наивного байесовского предположения $p(w_1,\ldots,w_{n_d}|t) = p(w_1|t)\cdots p(w_{n_d}|t)$.
- Число параметров $|D||T| + |W||T|$, возможно переобучение!

Максимизация правдоподобия: EM-алгоритм

Сформировать начальные приближения $p(t)$, $p(d|t)$, $p(w|t)$;
Повторять итерации до сходимости:

- **E-шаг:** скрытые переменные H по формуле Байеса:

$$H(t|d,w) = \frac{p(t)p(d|t)p(w|t)}{p(d,w)};$$

- **M-шаг:** аналитическое решение задачи $\mathscr{L} \to \max$:

$$p(t) = \frac{S(t)}{S}; \quad S(t) = \sum_{d,w} n_{dw} H(t|d,w); \quad S = \sum_{d,w} n_{dw};$$

$$p(d|t) = \frac{1}{S(t)} \sum_{w} n_{dw} H(t|d,w);$$

$$p(w|t) = \frac{1}{S(t)} \sum_{d} n_{dw} H(t|d,w).$$

Вывод формул M-шага

Распишем Лагранжиан:

$$\mathscr{L} = \sum_{d,w} n_{dw} \ln p(d,w) - \nu\left(\sum_{t\in T} p(t) - 1\right) -$$
$$- \sum_{t\in T} \lambda_t \left(\sum_{d\in D} p(d|t) - 1\right) - \sum_{t\in T} \mu_t \left(\sum_{w\in W} p(w|t) - 1\right).$$

$$\frac{\partial \mathscr{L}}{\partial p(t)} = \sum_{d,w} n_{dw} \frac{p(d|t)p(w|t)}{p(d,w)} - \nu = 0;$$

$$\sum_{d,w} n_{dw} \frac{p(d|t)p(w|t)p(t)}{p(d,w)} = \nu p(t) \;\Rightarrow\; \nu = \sum_{d,w} n_{dw} = S;$$

$$p(t) = \frac{1}{S} \sum_{d,w} n_{dw} H(t|d,w) = \frac{S(t)}{S};$$

Вывод формул M-шага (продолжение)

$$\frac{\partial \mathscr{L}}{\partial p(d|t)} = \sum_w n_{dw} \frac{p(t)p(w|t)}{p(d,w)} - \lambda_t = 0;$$

$$\sum_w n_{dw} \frac{p(t)p(w|t)p(d|t)}{p(d,w)} = \lambda_t p(d|t) \;\Rightarrow\; \lambda_t = \sum_{d,w} n_{dw} H(t|d,w);$$

$$p(d|t) = \sum_w n_{dw} H(t|d,w) \Big/ \sum_{d,w} n_{dw} H(t|d,w).$$

$$\frac{\partial \mathscr{L}}{\partial p(w|t)} = \sum_d n_{dw} \frac{p(t)p(d|t)}{p(d,w)} - \mu_t = 0;$$

$$\sum_d n_{dw} \frac{p(t)p(d|t)p(w|t)}{p(d,w)} = \mu_t p(w|t) \;\Rightarrow\; \mu_t = \sum_{d,w} n_{dw} H(t|d,w);$$

$$p(w|t) = \sum_d n_{dw} H(t|d,w) \Big/ \sum_{d,w} n_{dw} H(t|d,w).$$

Замечания о методе PLSA

1. **Оценка профиля нового документа** (folding-in):

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

где $p(w|d) = n_{dw}/n_d$ — оценка униграммной модели.

2. **Меры по уменьшению переобучения:**
   - обнуление незначимых компонент профилей $p(t|d)$, $p(t|w)$ (сокращение числа параметров модели);
   - обнуление незначимых скрытых переменных $H(t|d, w)$;
   - аккуратное формирование начального приближения;
   - ранний останов (оптимизация числа итераций);
   - симметризованный EM-алгоритм, в котором профили $p(t|w)$ и $p(t|d)$ уточняются по очереди.

Латентное размещение Дирихле
Latent Dirichlet Allocation [David Blei, 2003]

Вероятностная тематическая модель:

$$p(d,w) = \sum_{t\in T} p(w|t)\underbrace{p(t|d)}_{\theta_{td}} p(d).$$

Гипотеза: $\theta_d = (\theta_{td})_{t\in T} \in \mathbb{R}^{|T|}$ — случайные векторы из априорного распределения Дирихле с параметром $\alpha \in \mathbb{R}^{|T|}$:

$$p(\theta|\alpha) = \mathrm{Dir}(\theta;\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_t^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \quad \sum_t \theta_t = 1.$$

Тематическая модель документа d как цепочки слов $W_d = \{w_1, \ldots, w_{n_d}\}$ — непрерывная смесь распределений $p(d|\theta_d)$:

$$p(W_d) = \int p(\theta_d|\alpha) \prod_{w\in d} \left(\sum_{t\in T} p(w|t)\theta_{td} \right)^{n_{dw}} d\theta_d$$

Почему именно Дирихле?

Пусть профили документов $\theta_d = (p(t|d))_{t \in T}$ выбираются из Dir:

$$p(\theta_d|\alpha) = \text{Dir}(\theta_d; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad d \in D,$$

затем темы слов в документах $d \in D$ выбираются из θ_d:

$$X_d = \{t_1, \ldots, t_{n_d}\} \sim \theta_d.$$

Тогда вероятность встретить каждую из тем t ровно n_{td} раз подчиняется мультиномиальному распределению:

$$p(X_d|\theta_d) = \text{Mult}(n_{1d}, \ldots, n_{Td}|\theta_d) = \frac{n_d!}{\prod_t n_{td}!} \prod_t \theta_{td}^{n_{td}},$$

при этом апостериорное распределение — снова Дирихле!, т. е. распределение Дирихле — *сопряжённое* к мультиномиальному:

$$p(\theta_d|X_d, \alpha) = \frac{p(X_d|\theta_d)p(\theta_d|\alpha)}{\int p(X_d|\theta)p(\theta|\alpha)\,d\theta} = \text{Dir}(\theta_d; \alpha'), \quad \alpha'_t = \alpha_t + n_{td}.$$

Свойство сопряжённости упрощает оценивание параметров

Априорный профиль документа d задаётся параметрами α:

$$p(t|d,\alpha) = \int p(t|d)p(\theta_d|\alpha)\,d\theta_d = \int \theta_{td}\mathrm{Dir}(\theta_d,\alpha)\,d\theta_d = \frac{\alpha_t}{\alpha_0}.$$

Апостериорный профиль документа d
при известной выборке тем $X_d = \{t_1,\ldots,t_{n_d}\} \sim \theta_d$:

$$p(t|d,X_d,\alpha) = \int p(t|d)p(\theta_d|X_d,\alpha)\,d\theta_d = \int \theta_{td}\mathrm{Dir}(\theta_d,\alpha')\,d\theta_d =$$

$$= \frac{n_{td}+\alpha_t}{\sum_{t'}(n_{t'd}+\alpha_{t'})} = \frac{n_{td}+\alpha_t}{n_d+\alpha_0},$$

n_{td} — сколько раз слово документа d было отнесено к теме t,
n_d — длина документа в словах.

Это MAP-оценка профиля $p(t|d)$ и MP-оценка при $\alpha_t \equiv 0$.

Аналогично, для распределений слов $p(w|t)$ в темах

Вероятностная тематическая модель:

$$p(d,w) = \sum_{t\in T} \underbrace{p(w|t)}_{\phi_{wt}} \underbrace{p(t|d)}_{\theta_{td}} p(d).$$

Гипотеза: $\phi_t = (\phi_{wt})_{w\in W} \in \mathbb{R}^{|W|} \sim \mathrm{Dir}(\phi;\beta)$.

Документ порождается двумя распределениями $p(t|d)$, $p(w|t)$

$$X_d = \{(t_i, w_i)\}_{i=1}^{n_d}, \quad t_i \sim \theta_d, \quad w_i \sim \phi_{t_i}.$$

Сопряжённое апостериорное распределение — снова Дирихле:

$$p(\phi_t|X_d,\beta) = \frac{p(X_d|\phi_t)p(\phi_t|\beta)}{\int p(X_d|\phi)p(\phi|\beta)\,d\phi} = \mathrm{Dir}(\phi_t;\beta'), \quad \beta'_w = \beta_w + n_{wt}.$$

Апостериорное распределение слов в теме t:

$$p(w|t,X_d,\beta) = \int \phi_{tw}\mathrm{Dir}(\phi_t,\beta')\,d\phi_t = \frac{n_{wt}+\beta_w}{n_t+\beta_0}.$$

Сэмплирование Гиббса [Griffiths, Steyvers, 2004]

Тематическая модель вхождения слова w в документ d:

$$p(t|d,w) = \frac{p(d,w,t)}{p(d,w)} = \frac{p(d)}{p(d,w)} p(w|t)p(t|d) \;\propto\; \frac{p(w,t)}{p(t)} \frac{p(d,t)}{p(d)}.$$

Апостериорные (ненормированные) оценки $p(w|t)$ и $p(t|d)$:

$$\tilde{p}(t|d,w) = \frac{n_{wt} + \beta_w}{n_t + \beta_0} \cdot \frac{n_{td} + \alpha_t}{n_d + \alpha_0},$$

n_{wt} — сколько раз слово w было отнесено к теме t;
n_{td} — сколько раз слово документа d было отнесено к теме t;
n_t — сколько раз слово было отнесено к теме t;
n_d — длина документа в словах.

Идея сэмплирования: для всех слов в коллекции (d,w) выберем тему $t \sim \tilde{p}(t|d,w)$, обновим счётчики n_{wt}, n_{td}, n_t, n_d, уточним распределение $\tilde{p}(t|d,w)$, и так много раз в цикле.

Алгоритм сэмплирования Гиббса

Вход: коллекция $X = \{(d, w)\colon d \in D,\ w \in d\}$; параметры α, β;
Выход: оценки $\hat{p}(w|t)$, $\hat{p}(t|d)$;

$n_{wt} := \beta$; $n_{td} := \alpha$ для всех $d \in D$, $w \in W$, $t \in T$;
$n_t := \beta|W|$; $n_d := \alpha|T|$ для всех $d \in D$, $t \in T$;
для $i := 1, \ldots, M$ — генерация M сэмплов
  для всех документов $d \in D$ и всех слов $w \in d$
    $\tilde{p}(t|d, w) := (n_{wt}/n_t) \cdot (n_{td}/n_d)$ для всех $t \in T$;
    если $i \geqslant 2$ **то** $t := t_{dw}$; $--n_{wt}$; $--n_{td}$; $--n_t$; $--n_d$;
    выбрать t из ненормированного распределения $\tilde{p}(t|d, w)$;
    $t_{dw} := t$; $++n_{wt}$; $++n_{td}$; $++n_t$; $++n_d$;
$\hat{p}(w|t) := n_{wt}/n_t$ для всех $w \in W$, $t \in T$;
$\hat{p}(t|d) := n_{td}/n_d$ для всех $d \in D$, $t \in T$;

Сэмплирование Гиббса: особенности, преимущества, недостатки

- Алгоритм прост и эффективен, время $O(LM)$, где L — длина коллекции (в словах), M — число сэмплов;
- Добавление нового документа (возможно в онлайне!) — тот же алгоритм, но без инициализации счётчиков.
- Сэмплов иногда достаточно десятка, иногда нужны сотни...

Griffiths, Steyvers. Finding scientific topics // Proceedings of the National Academy of Sciences. USA, 2004. — Vol. 101. — Pp. 5228–5235.

- Более современные вариационные методы ещё быстрее (CVB — Collapsed Variational Bayesian);

Teh, Newman, Wellingm. A collapsed variational Bayesian inference algorithm for Latent Dirichlet Allocation // Advances in Neural Information Processing Systems (NIPS). Cambridge, MA, MIT Press, 2006

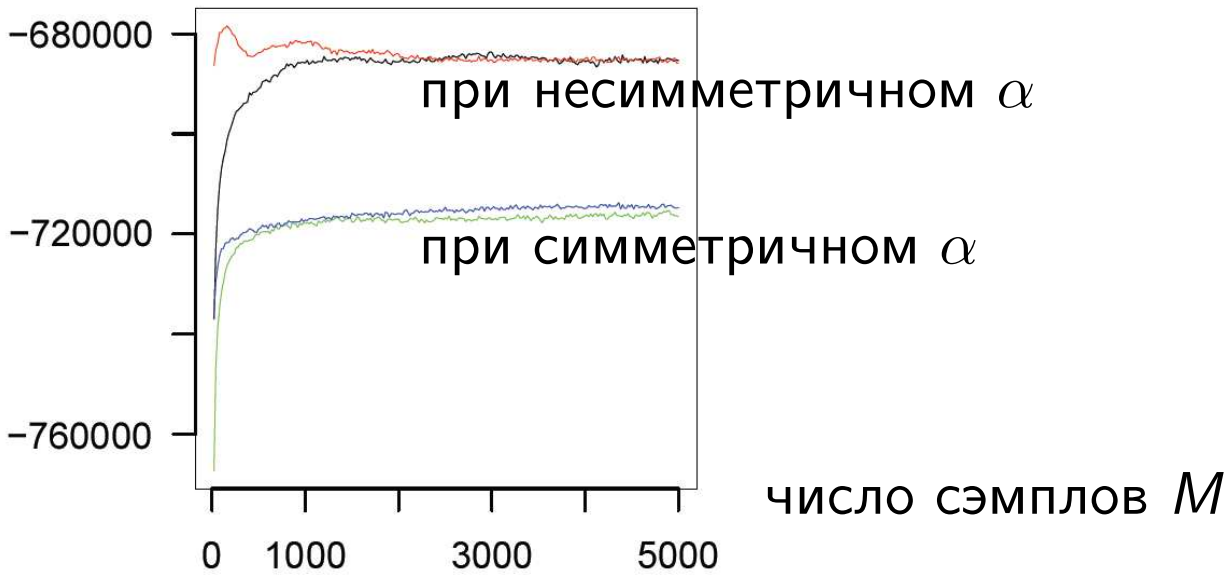
Проблема выбора гиперпараметров α и β

Стандартная рекомендация [2004]: $\alpha_t = 50/|T|, \ \beta_t = 0.01$.

Выводы по результатам более тонкого исследования [2009]:

- $p(t|d) \sim \mathrm{Dir}(\theta; \alpha)$, оптимизировать $\alpha = (\alpha_1, \ldots, \alpha_T)$.
- $p(w|t) \sim \mathrm{Dir}(\phi; \beta)$, взять симметричное $\beta_1 = \cdots = \beta_T \ll 1$.

правдоподобие



Hanna Wallach, David Mimno, Andrew McCallum.
Rethinking LDA: why priors matter. Neural Information Processing Systems, 2009.

Оптимизация гиперпараметра α

Обоснованность (evidence) модели на коллекции $X = (X_d)_{d \in D}$:

$$P(X|\alpha) = \int P(X|\theta) p(\theta|\alpha)\, d\theta =$$

$$= \prod_{d \in D} \frac{\Gamma(\alpha_0)}{\Gamma(n_d + \alpha_0)} \prod_{t \in T} \frac{\Gamma(n_{td} + \alpha_t)}{\Gamma(\alpha_t)} \to \max_{\alpha}$$

Метод неподвижной точки [Minka, 2003] — итерационный процесс, встраиваемый между последовательными сэмплами:

$$\alpha_t := \alpha_t \frac{\sum_d \psi(n_{td} + \alpha_t) - \psi(\alpha_t)}{\sum_d \psi(n_d + \alpha_0) - \psi(\alpha_0)},$$

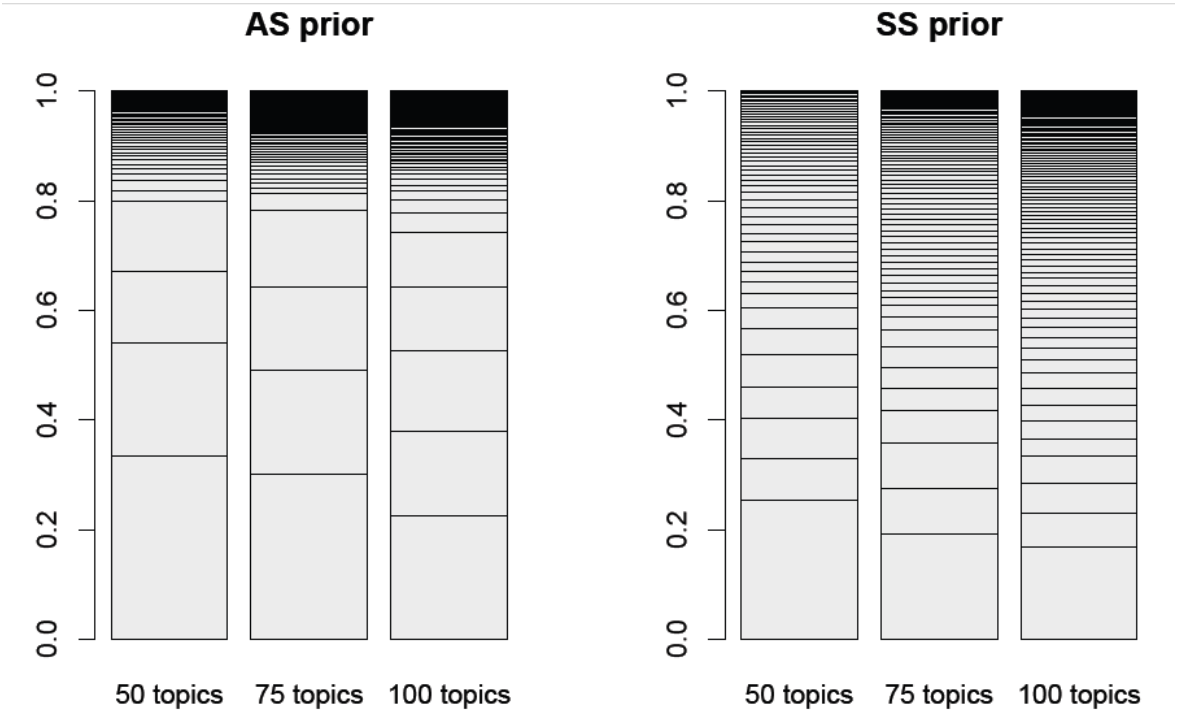
где $\psi(z) = \big(\ln \Gamma(z)\big)' = \Gamma'(z)/\Gamma(z)$ — дигамма-функция.

Более быстрые и точные методы оптимизации α:

Hanna Wallach. Structured Topic Models for Language.
PhD thesis, University of Cambridge, 2008.

Преимущество оптимизации гиперпараметра α

- Правдоподобие существенно выше.
- Сходимость быстрее, сэмплов нужно намного меньше.
- Меньшая чувствительность к избыточному $|T|$.
- Меньшее дробление тематики (это хорошо или плохо?):



Преимущества LDA

- Модель порождения документа d общая для всей коллекции D, а не отдельная для каждого $d \in D$.
- Профиль нового документа $p(t|d)$ оценивается по той же модели, что и для всех документов обучающей коллекции.
- Число параметров $|T| + |T| \cdot |W|$ не зависит от $|D|$.
- При $p(w|t) \sim \mathrm{Dir}(\beta)$ число параметров $|T|+|W|$ или $|T|+1$.
- Существуют эффективные методы оценивания профилей $p(t|d)$, распределений $p(w|t)$ и гиперпараметров α, β.

David Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

Эффективная реализация — в проекте Vowpal Wabbit:
`http://hunch.net/~vw/`

Методы сравнения тематических профилей

При классификации и поиске документов необходимо оценивать сходство документов по профилям $p(t|d)$, $p(t|d')$

- Среднеквадратичное отклонение не подходит!
- Расстояние Кульбака-Лейблера:

$$\mathrm{KL}(d, d') = \sum_{t \in T} p(t|d) \log \frac{p(t|d)}{p(t|d')}$$

(не симметрично, поэтому также не рекомендуется).

- Симметризованное расстояние Кульбака-Лейблера:

$$\mathrm{SymKL}(d, d') = \tfrac{1}{2}\mathrm{KL}(d, d') + \tfrac{1}{2}\mathrm{KL}(d', d).$$

- Симметризованное расстояние Йенсена-Шеннона:

$$\mathrm{SymJS}(d, d') = \tfrac{1}{2}\mathrm{KL}(d, \bar{d}) + \tfrac{1}{2}\mathrm{KL}(d', \bar{d}), \quad \bar{d} = \tfrac{1}{2}(d + d').$$

Методы оценивания качества тематических моделей

- На размеченной тестовой коллекции D':
  — число ошибок классификации (чем меньше, тем лучше).
- На неразмеченной тестовой коллекции D':
  — perplexity, неопределённость (чем меньше, тем лучше):

$$\text{perplexity} = \exp\left(-\frac{\sum\limits_{d\in D'}\sum\limits_{w\in d} n_{dw}\ln p(d,w)}{\sum\limits_{d\in D'}\sum\limits_{w\in d} n_{dw}}\right)$$

- На обучающей коллекции D: насколько нарушается гипотеза условной независимости $p(w|d,t) = p(w|t)$

$$\text{KL}\Big(\hat{p}(d,w|t),\ \hat{p}(d|t)\cdot\hat{p}(w|t)\Big) = \sum_{d,w}\frac{n_{dwt}}{n_t}\log\frac{n_{dwt}\cdot n_t}{n_{td}\cdot n_{wt}}$$

David Mimno, David Blei. Bayesian Checking for Topic Models // Empirical Methods in Natural Language Processing, 2011.

Обобщения и модификации тематических моделей

- Иерархические модели, в том числе с адаптивной иерархией
- Темпоральные модели, учитывающие годы публикаций
- Author-topic models — пытаются приписать распределение авторов $p(a|w)$ каждому слову документа
- Entity-topic models — оценивают тематику связанных сущностей: журналов, конференций, организаций
- Модели, учитывающие связь слов внутри документа
- Модели связей между документами (ссылки, цитирование)

Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad.
Knowledge discovery through directed probabilistic topic models: a survey.
Frontiers of Computer Science in China, Vol. 4, No. 2., 2010, Pp. 280–301.
(имеется русский перевод)

Topic Modeling Bibliography:
`http://www.cs.princeton.edu/~mimno/topics.html`

Открытая проблема: построить иерархическую модель, удовлетворяющую следующей совокупности требований

1. тематическое дерево — направленный ациклический граф: вершины — темы, рёбра — связи «подтема→тема»;
2. тематическое дерево должно быть сбалансированным: (в каждой теме $\sim 10^1$ подтем, $\sim 10^1$ документов);
3. граф должен строиться по корпусу из $\sim 10^7$ текстов;
4. корпус растёт динамически $\sim 10^3$ в день;
5. граф учитывает стандартные рубрикаторы (например, УДК);
6. эксперты могут добавлять и удалять вершины и рёбра графа;
7. эксперты могут создавать и запрещать связи «документ→тема», «термин→тема»;
8. исходные данные корпуса текстов, рубрикаторов и экспертов могут противоречить друг другу;
9. тематические профили должны быть разреженными;
10. термины и темы должны быть многоязычными.

Проект RUCONT

Официальное название проекта:
Проект web-сервиса для интеллектуального поиска, классификации и агрегации научной информации в пополняемых мультидисциплинарных коллекциях текстовых документов.

Сроки проекта:
октябрь 2011 — май 2013

Четыре основные составные части проекта:

1. Хранение документов и защита авторских прав.
2. Поиск: обычный, тематический, антиплагиат.
3. Наукометрический анализ (индексы цитирования по темам).
4. Поддержка процессов подготовки электронных изданий (приём, рецензирование, обсуждение, оценивание).

12 исследовательских задач

1. Обзор функционалов качества тематических моделей
2. Сравнение и симбиоз PLSA и сэмплирования Гиббса GS
3. Сравнение регуляризации и topic selection
4. Выбор эффективной реализации для больших коллекций
5. Построения тематической сети произвольной глубины
6. Формирование предложений по реструктуризации сети
7. Учёт имеющихся рубрикаторов УДК, ББК и др.
8. Учёт обучающей информации от пользователей (тематические классификации слов, документов, подтем)
9. Выявление тематики в двуязычной коллекции
10. Извлечение терминов-словосочетаний для PLSA, GS
11. Выявление и устранение опечаток по контексту
12. Учёт слов в заголовках и аннотациях с бóльшим весом