Вероятностное тематическое моделирование

К. В. Воронцов

16 октября 2013 г.

Содержание

1 Задачи тематического моделирования

Тематическое моделирование (topic modeling) — одно из современных приложений машинного обучения к анализу текстов, активно развивающееся с конца 90-х годов. *Тематическая модель* (topic model) коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему.

Вероятностная тематическая модель (BTM) описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем. Предполагается, что коллекция документов — это последовательность терминов, выбранных случайно и независимо из смеси таких распределений, и ставится задача восстановления компонент смеси по выборке.

Поскольку документ или термин может относиться одновременно ко многим темам с различными вероятностями, говорят, что BTM осуществляет «мягкую» кластеризацию документов и терминов по кластерам-темам. Тем самым решаются проблемы синонимии и омонимии терминов, возникающие при обычной «жёсткой» кластеризации. Синонимы, часто употребляющиеся в схожих контекстах, с большой вероятностью попадают в одну тему. Омонимы, употребляющиеся в разных контекстах, распределяются между несколькими темами соответственно частоте употребления.

Тематические модели применяются для выявления трендов в научных публикациях или новостных потоках [68, 55], для классификации и категоризации документов [48] изображений и видеопотоков [30, 18, 56], для информационного поиска [65], в том числе многоязычного [57], для тегирования веб-страниц [26], для обнаружения текстового спама [6], для рекомендательных систем [64] и других приложений.

Одним из основных приложений является информационный поиск. Поисковые системы представляют документы векторами частот слов. Поиск документов по коротким запросам реализуется путём поиска векторов, в которых часто встречаются слова запроса [5]. Тематическая модель позволяет использовать тот же механизм для поиска документов схожей тематики по целому документу или по длинному фрагменту текста. При этом документы представляются векторами частот тем, а не отдельных слов. Векторами частот тем представляются также объекты, упоминаемые в документах или связанные с документами: термины, авторы, годы публикации, институты, конференции, журналы, сайты и т. д., что позволяет задавать в качестве запроса любой объект или совокупность объектов и искать по ним объекты того же или другого типа, имеющие схожую тематику.

Тематические модели могут учитывать различные особенности языка и текстовых коллекций. Существуют модели, выявляющие ключевые фразы (термины предметной области), учитывающие морфологию слов и синтаксическую структуру предложений, отслеживающие изменения тематики во времени или внутри отдельных документов, строящие иерархические отношения между темами, учитывающие связи между документами через авторство или ссылки, и т. д. Многочисленные разновидности вероятностных тематических моделей описаны в обзоре [14].

Большинство моделей разрабатываются на основе *латентного размещения Дирихле* LDA [9], математического аппарата графических моделей и байесовского вывода. Это современный активно развивающийся вероятностный инструментарий, находящий применения повсеместно в задачах анализа данных. Однако в тематическом

моделировании он порождает две открытые проблемы, которые на удивление мало освещаются в литературе по ВТМ.

Первая проблема заключается в том, что априорные распределения Дирихле и их обобщения — процессы Дирихле и Питмана-Йора — имеют крайне слабые лингвистические обоснования. Они не моделируют какие-либо явления языка. Их применение продиктовано исключительно математическим удобством, так как они позволяют аналитически выполнять интегрирование в байесовском выводе.

Второй проблемой является сложность совмещения большого числа функциональных требований в одной модели [14]. В частности, для обработки больших коллекций научных публикаций нужна модель, удовлетворяющая десятку требований одновременно (иерархическая, динамическая, n-грамная, мультиязычная, разреженная, робастная, и т. д.). Многие исследователи признают, что общепринятый математический аппарат слишком сложен для совмещения более 2–3 требований.

В данной работе развивается теория *аддитивной регуляризации тематических моделей* (АРТМ), которая решает обе эти проблемы. За основу берётся классическая модель *вероятностного латентного семантического анализа* PLSA [22]. Для оценивания параметров модели PLSA применяется EM-алгоритм, который ищет максимум правдоподобия. Максимум достигается на бесконечном множестве моделей, то есть задача построения ВТМ является некорректно поставленной. Это служит обоснованием для введения регуляризации [7]. К функционалу логарифма правдоподобия добавляются штрафные слагаемые (регуляризаторы), выражающие различные дополнительные требования к модели, не обязательно вероятностного характера. Каждая аддитивная поправка к функционалу приводит к аддитивной поправке в формуле M-шага EM-алгоритма. Это позволяет комбинировать произвольное число требований и строить *многоцелевые тематические модели*. Модели LDA соответствует свой аддитивный регуляризатор. Многочисленным модификациям LDA также соответствуют свои регуляризаторы. Теория АРТМ описывает огромное разнообразие тематических моделей, созданных за последнее десятилетие, не прибегая к избыточным вероятностным допущениям и сложным техникам байесовского вывода.

§1.1 Вероятностная модель коллекции документов

Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) всех употребляемых в них терминов (слов или словосочетаний). Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, \ldots, w_{n_d})$ из словаря W. Термин может повторяться в документе много раз.

Вероятностное пространство и гипотеза независимости. Предполагается, что существует конечное множество тем T, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$, которая не известна. Коллекция документов рассматривается как множество троек (d, w, t), выбранных *случайно и независимо* из дискретного распределения $p(d, w, t)$, заданного на конечном множестве $D \times W \times T$. Документы $d \in D$ и термины $w \in W$ являются наблюдаемыми переменными, тема $t \in T$ является *латентной* (скрытой) переменной.

Гипотеза о независимости элементов выборки эквивалентна предположению, что порядок терминов в документах не важен для выявления тематики, то есть тематику документа можно узнать даже после произвольной перестановки терминов,

хотя для человека такой текст теряет смысл. Это предположение называют гипотезой «мешка слов» (bag of words). Порядок документов в коллекции также не имеет значения; это предположение называют гипотезой «мешка документов».

Приняв гипотезу «мешка слов», можно перейти к более компактному представлению документа как подмножества $d \subset W$, в котором каждому элементу $w \in d$ поставлено в соответствие число n_{dw} вхождений термина w в документ d.

Постановка задачи тематического моделирования. Построить *тематическую модель* коллекции документов D — значит найти множество тем T, распределения $p(w \mid t)$ для всех тем $t \in T$ и распределения $p(t \mid d)$ для всех документов $d \in D$. Можно также говорить о задаче совместной «мягкой» кластеризации множества документов и множества слов по множеству кластеров-тем. *Мягкая кластеризация* означает, что каждый документ или термин не жёстко приписывается какой-то одной теме, а распределяется по нескольким темам.

Найденные распределения используются затем для решения прикладных задач. Распределение $p(t \mid d)$ является удобным признаковым описанием документа в задачах информационного поиска, классификации и категоризации документов.

Гипотеза условной независимости. Будем полагать, что появление слов в документе d, относящихся к теме t, описывается общим для всей коллекции распределением $p(w \mid t)$ и не зависит от документа d. Это предположение, называемое *гипотезой условной независимости*, допускает три эквивалентных представления:

$$\begin{aligned} &p(w \mid d, t) = p(w \mid t); \\ &p(d \mid w, t) = p(d \mid t); \\ &p(d, w \mid t) = p(d \mid t) p(w \mid t). \end{aligned} \tag{1.1}$$

Вероятностная модель порождения данных. Согласно определению условной вероятности, формуле полной вероятности и гипотезе условной независимости

$$p(w \mid d) = \sum_{t \in T} p(t \mid d)\, p(w \mid t). \tag{1.2}$$

Если распределения $p(t \mid d)$ и $p(w \mid t)$ известны, то вероятностная модель (1.2) описывает процесс порождения коллекции D, см. также Алгоритм 1.1 и рис. 1.

Построение тематической модели — это обратная задача: по известной коллекции D требуется восстановить породившие её распределения $p(t \mid d)$ и $p(w \mid t)$.

Гипотеза разреженности. Естественно предполагать, что каждый документ d и каждый термин w связан с небольшим числом тем t. В таком случае значительная часть вероятностей $p(t \mid d)$ и $p(w \mid t)$ должна обращаться в нуль.

Если документ относится к большому числу тем (например, энциклопедия, журнал, сборник статей), то в задачах тематического поиска или классификации документов его имеет смысл разбивать на части, более однородные по тематике.

Если термин относится к большому числу тем, то, скорее всего, это общеупотребительное слово (стоп-слово), бесполезное для определения тематики.

Алгоритм 1.1. Вероятностная модель порождения коллекции документов.

Вход: распределения $p(w \mid t)$, $p(t \mid d)$;
Выход: выборка пар (d_i, w_i), $i = 1, \ldots, n$;

```
для всех d ∈ D
    задать длину n_d документа d;
    для всех i = 1, ..., n_d
        выбрать случайную тему t из распределения p(t | d);
        выбрать случайный термин w из распределения p(w | t);
        добавить в выборку пару (d, w), при этом тема t «забывается»;
```

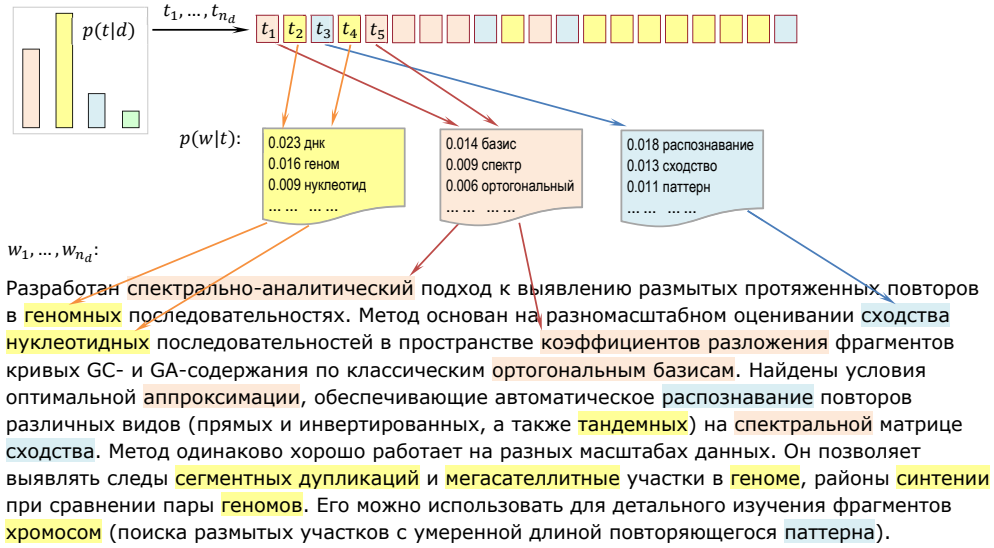


Рис. 1. Процесс порождения текстового документа вероятностной тематической моделью (1.2).

Алгоритмы, в которых нулевые значения не хранятся, намного эффективнее по памяти и по скорости. Поэтому для больших коллекций разреженность должна учитываться обязательно.

Частотные оценки условных вероятностей. Вероятности, связанные с наблюдаемыми переменными d и w, можно оценивать по выборке как частоты (здесь и далее выборочные оценки вероятностей p будем обозначать через $\hat{p}$):

$$\hat{p}(d, w) = \frac{n_{dw}}{n}, \qquad \hat{p}(d) = \frac{n_d}{n}, \qquad \hat{p}(w) = \frac{n_w}{n}, \qquad \hat{p}(w \mid d) = \frac{n_{dw}}{n_d}, \tag{1.3}$$

n_{dw} — число вхождений термина w в документ d;

$n_d = \sum\limits_{w \in W} n_{dw}$ — длина документа d в терминах;

$n_w = \sum\limits_{d \in D} n_{dw}$ — число вхождений термина w во все документы коллекции;

$n = \sum\limits_{d \in D} \sum\limits_{w \in d} n_{dw}$ — длина коллекции в терминах.

Вероятности, связанные со скрытой переменной t, также можно оценивать как частоты, если рассматривать коллекцию документов как выборку троек (d, w, t):

$$\hat{p}(t) = \frac{n_t}{n}, \qquad \hat{p}(w \mid t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(t \mid d) = \frac{n_{dt}}{n_d}, \qquad \hat{p}(t \mid d, w) = \frac{n_{dwt}}{n_{dw}}, \tag{1.4}$$

n_{dwt} — число троек, в которых термин w документа d связан с темой t;

$n_{dt} = \sum\limits_{w \in W} n_{dwt}$ — число троек, в которых термин документа d связан с темой t;

$n_{wt} = \sum\limits_{d \in D} n_{dwt}$ — число троек, в которых термин w связан с темой t;

$n_t = \sum\limits_{d \in D} \sum\limits_{w \in d} n_{dwt}$ — число троек, связанных с темой t.

В пределе $n \to \infty$ частотные оценки $\hat{p}(\cdot)$, определяемые формулами (1.3)–(1.4), стремятся к соответствующим вероятностям $p(\cdot)$, согласно закону больших чисел. Частотная интерпретация даёт ясное понимание всех условных вероятностей, которые будут использоваться в дальнейшем.

Стохастическое матричное разложение. Если число тем $|T|$ много меньше числа документов $|D|$ и числа терминов $|W|$, то равенство (1.2) можно понимать как задачу приближённого представления заданной матрицы частот

$$F = \big(\hat{p}_{wd}\big)_{W \times D}, \quad \hat{p}_{wd} = \hat{p}(w \mid d) = n_{dw}/n_d,$$

в виде произведения $F \approx \Phi\Theta$ двух неизвестных матриц меньшего размера — *матрицы терминов тем* Φ и *матрицы тем документов* Θ:

$$\Phi = (\varphi_{wt})_{W \times T}, \quad \varphi_{wt} = p(w \mid t);$$
$$\Theta = (\theta_{td})_{T \times D}, \quad \theta_{td} = p(t \mid d).$$

Матрицы, столбцы которых неотрицательны и нормированы, следовательно, могут пониматься как дискретные распределения, называются *стохастическими*.

Одно из наиболее известных представлений вида $F \approx \Phi\Theta$ строится из $|T|$ главных компонент сингулярного разложения матрицы F и является решением задачи наименьших квадратов

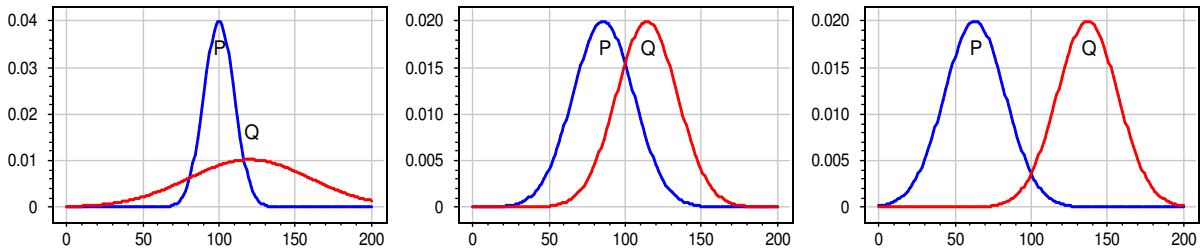
$$\sum_{d \in D} \sum_{w \in W} \big(\hat{p}_{wd} - p(w \mid d)\big)^2 = \sum_{d \in D} \sum_{w \in W} \Big(\hat{p}_{wd} - \sum_{t \in T} \varphi_{wt}\theta_{td}\Big)^2 = \|F - \Phi\Theta\|^2 \to \min_{\Phi,\Theta}. \tag{1.5}$$

Метод главных компонент не подходит для тематического моделирования, так как получаемые с его помощью матрицы Φ, Θ в общем случае не являются стохастическими. Кроме того, квадратичная функция потерь плохо подходит для сравнения вероятностных распределений с «тяжёлыми хвостами».

В вероятностном тематическом моделировании вместо принципа наименьших квадратов используется принцип максимума правдоподобия.

Принцип максимума правдоподобия. Для оценивания параметров Φ, Θ тематической модели по коллекции документов D будем максимизировать правдоподобие (плотность распределения) выборки:

$$p(D; \Phi, \Theta) = C \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}} = \prod_{d \in D} \prod_{w \in d} p(w \mid d)^{n_{dw}} \underbrace{C p(d)^{n_{dw}}}_{\text{const}} \to \max_{\Phi,\Theta},$$



$\mathrm{KL}(P\|Q) = 0.442$ $\mathrm{KL}(Q\|P) = 2.966$ | $\mathrm{KL}(P\|Q) = 0.444$ $\mathrm{KL}(Q\|P) = 0.444$ | $\mathrm{KL}(P\|Q) = 2.969$ $\mathrm{KL}(Q\|P) = 2.969$

Рис. 2. Дивергенция Кульбака–Лейблера $\mathrm{KL}(P\|Q)$ является несимметричной мерой вложенности распределения $P = (p_i)_{i=1}^n$ в распределение $Q = (q_i)_{i=1}^n$. Вложенность P в Q приблизительно одинакова на левом и среднем графиках, вложенность Q в P — на левом и правом графиках.

где C — нормировочный множитель, зависящий только от чисел n_{dw}. Отбросим множители C и $p(d)$, не влияющие на положение точки максимума, подставим выражение для $p(w\,|\,d)$ из (1.2) и воспользуемся обозначениями $\theta_{td} = p(t\,|\,d)$, $\varphi_{wt} = p(w\,|\,t)$. Прологарифмируем $p(D;\Phi,\Theta)$, чтобы превратить произведения в суммы. Получим задачу максимизации логарифма правдоподобия (log-likelihood) при ограничениях неотрицательности и нормированности столбцов матриц Φ и Θ:

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}; \tag{1.6}$$

$$\sum_{w\in W} \varphi_{wt} = 1; \qquad \varphi_{wt} \geqslant 0;$$

$$\sum_{t\in T} \theta_{td} = 1; \qquad \theta_{td} \geqslant 0.$$

Дивергенция Кульбака–Лейблера или *KL-дивергенция* между дискретными распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$ — это несимметричная функция расстояния

$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

Предполагается, что $p_i > 0$ и $q_i > 0$. KL-дивергенция является не вполне адекватной функцией расстояния в случае, когда у распределений P и Q не совпадают носители $\Omega_P = \{i\colon p_i > 0\}$ и $\Omega_Q = \{i\colon q_i > 0\}$.

Перечислим наиболее важные свойства KL-дивергенции.

1. KL-дивергенция неотрицательна. Если $\Omega_P = \Omega_Q$, то KL-дивергенция равна нулю тогда и только тогда, когда распределения совпадают, $p_i \equiv q_i$.

2. KL-дивергенция является мерой вложенности двух распределений. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то распределение P сильнее вложено в Q, чем Q в P, см. рис. 2.

3. Если P — эмпирическое распределение, а $Q(\alpha)$ — параметрическое семейство (модель) распределений, то минимизация KL-дивергенции эквивалентна максимизации правдоподобия:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_\alpha \quad\Longleftrightarrow\quad \sum_{i=1}^n p_i \ln q_i(\alpha) \to \max_\alpha.$$

Максимизация правдоподобия (1.6) эквивалентна минимизации взвешенной суммы дивергенций Кульбака–Лейблера между эмпирическими распределениями $\hat{p}(w \mid d) = n_{dw}/n_d$ и модельными $p(w \mid d)$, по всем документам d из D:

$$\sum_{d \in D} n_d \, \mathrm{KL}_w\Big(\frac{n_{dw}}{n_d} \,\Big\|\, \sum_{t \in T} \varphi_{wt}\theta_{td}\Big) \;\to\; \min_{\Phi,\Theta},$$

где весом документа d является его длина n_d. Если веса n_d убрать, то все документы будут искусственно приведены к одинаковой длине. Такая модификация функционала качества может быть полезна при моделировании коллекций, содержащих документы одинаковой важности, но существенно разной длины.

Лирическое отступление: как решать задачи оптимизации с ограничениями равенствами и неравенствами. Лагранжиан. Теорема Куна–Таккера. Иногда можно игнорировать ограничения-неравенства и, получив решение, доказывать, что оно удовлетворяет этим ограничениям. Просто везение. ToDo[1]

§1.2 Униграммная порождающая модель

Простейшим примером вероятностной порождающей модели является *униграммная модель*, основанная на предположении, что каждое слово появляется в тексте независимо от остальных слов. Модели, в которых учитываются пары, тройки, n-ки слов (обычно соседних), называются, соответственно, *биграммными*, *триграммными*, *n-граммными*. Рассмотрим два варианта униграммной модели.

Униграммная модель документов. Предполагается, что слова каждого документа генерируются случайно и независимо из распределения $p(w|d) = \xi_{dw}$, своего для каждого документа d. Запишем задачу максимизации правдоподобия при ограничениях нормировки и неотрицательности на параметры модели ξ_{dw}:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \xi_{dw} \to \max_{\xi}, \qquad \sum_{w \in W} \xi_{dw} = 1, \qquad \xi_{dw} \geqslant 0.$$

Запишем функцию Лагранжа, проигнорировав ограничения-неравенства (потом, получив решение, убедимся, что они выполнены автоматически):

$$\mathscr{L} = \sum_{d \in D} \Big(\sum_{w \in d} n_{dw} \ln \xi_{dw} - \lambda_d \Big(\sum_{w \in W} \xi_{dw} - 1 \Big) \Big);$$

приравняем нулю производные по переменным ξ_{dw}:

$$\frac{\partial \mathscr{L}}{\partial \xi_{dw}} = \frac{n_{dw}}{\xi_{dw}} - \lambda_d = 0.$$

Суммируя по $w \in d$, получим значение двойственных переменных $\lambda_d = n_d$, и, подставляя их обратно в уравнение для ξ_{dw}, найдём, что искомый параметр ξ_{dw} равен частотной оценке условной вероятности слова w в документе d:

$$\xi_{dw} = \hat{p}(w|d) = n_{dw}/n_d. \tag{1.7}$$

Униграммная модель коллекции. Предполагается, что слова каждого документа генерируются случайно и независимо из распределения $p(w|d) = \xi_w$, общего для всех документов коллекции. По аналогии с предыдущим случаем,

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\ln\xi_w \to \max_{\xi}, \qquad \sum_{w\in W}\xi_w = 1, \qquad \xi_w \geqslant 0.$$
$$\mathscr{L} = \sum_{d\in D}\sum_{w\in d} n_{dw}\ln\xi_w - \lambda\Big(\sum_{w\in W}\xi_w - 1\Big);$$
$$\frac{\partial\mathscr{L}}{\partial\xi_w} = \frac{n_w}{\xi_w} - \lambda = 0;$$

откуда следует, что $\lambda = n$, и искомый параметр ξ_w равен частотной оценке вероятности слова w во всей коллекции:

$$\xi_w = \hat{p}(w) = n_w/n. \tag{1.8}$$

Обе униграммные модели имеют простые, интуитивно очевидные решения (1.7) и (1.8), но не являются тематическими. Тематическая модель (1.2) занимает между ними промежуточное положение. Набор её параметров богаче униграммной модели коллекции, но беднее униграммной модели документов.

Лирическое отступление: распределения с тяжёлыми хвостами. Законы Ципфа, Ципфа–Мандельброта, Хипса. ToDo[2]

§1.3 Предварительная обработка текстовых данных

Понятие «термина» может изменяться в зависимости от целей построения тематической модели и таких особенностей задачи, как язык документов, средняя длина документов, тематика коллекции.

Лемматизация и стемминг. При построении тематической модели нет смысла различать формы (склонения, спряжения) одного и того же слова. Это приведёт к неоправданному разрастанию словаря, дроблению статистики, увеличению ресурсоёмкости и снижению качества модели.

Лемматизация — это приведение каждого слова в документе к его нормальной форме. В русском языке нормальными формами считаются: для существительных — именительный падеж, единственное число; для прилагательных — именительный падеж, единственное число, мужской род; для глаголов, причастий, деепричастий — глагол в инфинитиве. Разработка хорошего *лемматизатора* (lemmatizer) требует составления грамматического словаря со всеми формами слов, либо аккуратной формализации правил языка со всеми исключениями, что является трудоёмким проектом. Известные лемматизаторы совершенствуются постепенно. Их недостатком является неполнота словарей, особенно по части специальной терминологии и неологизмов, которые во многих приложениях как раз и представляют наибольший интерес.

ссылка на рекомендуемые русский и английский лемматизаторы ToDo[3]

Стемминг — это более простая технология, которая состоит в отбрасывании изменяемых частей слов, главным образом, окончаний. Она не требует хранения словаря всех слов и основана на правилах морфологии языка. Недостатком стемминга

является большее число ошибок. Стемминг хорошо подходит для английского языка, но хуже подходит для русского.

ссылка на рекомендуемые русский и английский стеммеры ToDo[4]

Отбрасывание стоп-слов. Слова, встречающиеся во многих текстах различной тематики, бесполезны для тематического моделирования, и могут быть отброшены. К ним относятся предлоги, союзы, числительные, местоимения, некоторые глаголы, прилагательные и наречия. Число таких слов обычно варьируется в пределах нескольких сотен. Их отбрасывание почти не влияет на длину словаря, но может приводить к заметному сокращению длины некоторых текстов.

Отбрасывание редких слов. Слова, встречающиеся в длинном документе слишком редко, например, только один раз, также можно отбрасывать, полагая, что данное слово не характеризует тематику данного документа. При обработке коллекций коротких новостных сообщений этот приём лучше не использовать.

Выделение ключевых фраз. При обработке коллекций научных, юридических или других специальных текстов вместо отдельных слов выделяют *ключевые фразы* — словосочетания, являющиеся терминами предметной области. Это отдельная довольно сложная задача, для решения которой используются тезаурусы, составленные экспертами [4], либо методы машинного обучения [44, 69], при этом для формирования обучающих выборок всё равно приходится привлекать экспертов.

Далее будем полагать, что словарь W получен в результате предварительной обработки всех документов коллекции D и может содержать как отдельные слова, так и ключевые фразы. Элементы словаря $w \in W$ будем называть «терминами».

2 Вероятностный латентный семантический анализ

Вероятностный латентный семантический анализ (probabilistic latent semantic analysis, PLSA) был предложен Томасом Хофманном в [22].

Вероятностная модель появления пары «документ–термин» (d, w) записывается тремя эквивалентными способами:

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

где $p(t)$ — распределение тем во всей коллекции. Первое представление называется симметричным, второе и третье — несимметричными. Они приводят к немного разным итерационным процессам обучения тематической модели. Сейчас возьмём за основу второе представление, совпадающее с (1.2).

§2.1 EM-алгоритм

Для решения задачи (1.6) в PLSA применяется итерационный процесс, в котором каждая итерация состоит из двух шагов — E (expectation) и M (maximization) [15]. Перед первой итерацией выбирается начальное приближение параметров φ_{wt}, θ_{td}.

На E-шаге по текущим значениям параметров φ_{wt}, θ_{td} с помощью формулы Байеса вычисляются условные вероятности $p(t \mid d, w)$ всех тем $t \in T$ для каждого термина $w \in d$ в каждом документе d:

$$H_{dwt} = p(t \mid d, w) = \frac{p(w \mid t)p(t \mid d)}{p(w \mid d)} = \frac{\varphi_{wt}\theta_{td}}{\sum\limits_{s \in T} \varphi_{ws}\theta_{sd}}. \tag{2.1}$$

На M-шаге, наоборот, по условным вероятностям тем H_{dwt} вычисляется новое приближение параметров φ_{wt}, θ_{td}. Это легко сделать, если заметить, что величина

$$\hat{n}_{dwt} = n_{dw}p(t \mid d, w) = n_{dw}H_{dwt} \tag{2.2}$$

оценивает (не обязательно целое) число n_{dwt} вхождений термина w в документ d, связанных с темой t. Просуммировав $\hat{n}_{dwt}$ по документам d и по терминам w, получим оценки $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, и через них, согласно (1.4), — частотные оценки условных вероятностей φ_{wt}, θ_{td}:

$$\varphi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \qquad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt}, \qquad \hat{n}_{wt} = \sum_{d \in D} n_{dw}H_{dwt}. \tag{2.3}$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \qquad \hat{n}_d = \sum_{t \in T} \hat{n}_{dt}, \qquad \hat{n}_{dt} = \sum_{w \in d} n_{dw}H_{dwt}. \tag{2.4}$$

Эти простые, но не вполне строгие рассуждения поясняют суть EM-алгоритма. Покажем теперь, что оценки (2.3)–(2.4) действительно являются решением задачи максимизации правдоподобия (1.6) при фиксированных H_{dwt}.

Запишем лагранжиан задачи (1.6) при ограничениях нормировки, проигнорировав ограничения неотрицательности (позже убедимся, что решение действительно неотрицательно):

$$\mathscr{L}(\Phi, \Theta) = \sum_{d \in D}\sum_{w \in d} n_{dw} \ln \underbrace{\sum_{t \in T} \varphi_{wt}\theta_{td}}_{p(w \mid d)} - \sum_{t \in T} \lambda_t \left(\sum_{w \in W} \varphi_{wt} - 1 \right) - \sum_{d \in D} \mu_d \left(\sum_{t \in T} \theta_{td} - 1 \right).$$

Продифференцировав лагранжиан по φ_{wt} и приравняв нулю производную, получим

$$\lambda_t = \sum_{d \in D} n_{dw} \frac{\theta_{td}}{p(w \mid d)}. \tag{2.5}$$

Домножим обе части этого равенства на φ_{wt}, просуммируем по всем терминам $w \in W$, применим условие нормировки вероятностей φ_{wt} в левой части и выделим переменную H_{dwt} в правой части. Получим

$$\lambda_t = \sum_{d \in D}\sum_{w \in W} n_{dw}H_{dwt}.$$

Снова домножим обе части (2.5) на φ_{wt}, выделим переменную H_{dwt} в правой части и выразим φ_{wt} из левой части, подставив уже известное выражение для λ_t. Получим

$$\varphi_{wt} = \frac{\sum\limits_{d \in D} n_{dw}H_{dwt}}{\sum\limits_{w' \in W}\sum\limits_{d \in D} n_{dw'}H_{dw't}}.$$

Алгоритм 2.1. PLSA-EM: рациональный EM-алгоритм для модели PLSA.

Вход: коллекция документов D, число тем $|T|$, начальные приближения Θ, Φ;
Выход: распределения Θ и Φ;

повторять
  обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D,\ w \in W,\ t \in T$;
  для всех $d \in D,\ w \in d$
    $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$;
    для всех $t \in T$ таких, что $\varphi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на $\delta = n_{dw}\varphi_{wt}\theta_{td}/Z$;
  $\varphi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D,\ t \in T$;
пока Θ и Φ не сойдутся;

Обозначив числитель через $\hat{n}_{wt}$, получим (2.3). Проделав аналогичные действия с производной лагранжиана по θ_{td}, получим (2.4).

Заметим, что если начальные приближения θ_{td} и φ_{wt} положительны, то и после каждой итерации они будут оставаться положительными, несмотря на то, что ограничение неотрицательности было проигнорировано в ходе решения.

Эффективность EM-алгоритма по времени и по памяти. Число операций растёт линейно по длине коллекции n, числу тем T и числу итераций.

Перебор всех терминов w во всех документах d можно организовать очень эффективно, если хранить каждый документ d в виде последовательности пар (w, n_{dw}).

Рациональный EM-алгоритм. Вычисление переменных $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на M-шаге требует однократного прохода всей коллекции в цикле по всем документам $d \in D$ и всем терминам $w \in d$. Внутри этого цикла переменные H_{dwt} можно вычислять непосредственно в тот момент, когда они понадобятся. От этого результат алгоритма не изменяется, E-шаг встраивается внутрь M-шага без дополнительных вычислительных затрат, отпадает необходимость хранения трёхмерной матрицы H_{dwt}. Заметим также, что переменную $\hat{n}_d$ можно не вычислять, поскольку $\hat{n}_d = n_d$. Этот вариант реализации EM-алгоритма будем называть *рациональным*; он показан в Алгоритме 2.1.

§2.2 Обобщённый EM-алгоритм

В EM-алгоритме нет необходимости сверхточно решать задачу максимизации правдоподобия на M-шаге. Достаточно ещё немного приблизиться к точке максимума правдоподобия и снова выполнить E-шаг. Это связано с тем, что сам функционал правдоподобия известен не точно — он зависит от приближённых значений H_{dwt}, полученных на E-шаге. EM-алгоритм с сокращённым M-шагом называется *обобщённым EM-алгоритмом* (generalized EM-algorithm, GEM). Для него справедливы те же доказательства сходимости, что и для основного варианта EM-алгоритма [15].

Другое обобщение состоит в том, что E-шаг выполняется только для части скрытых переменных H_{dwt}. После этого M-шаг выполняется только для тех основных

Алгоритм 2.2. PLSA-GEM: обобщённый EM-алгоритм для модели PLSA.

Вход: коллекция документов D, число тем $|T|$, начальные приближения Θ, Φ;
Выход: распределения Θ и Φ;

1 обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$, n_{dwt} для всех $d \in D,\ w \in W,\ t \in T$;
2 **повторять**
3 **для всех** $d \in D,\ w \in d$
4 $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$;
5 **для всех** $t \in T$ таких, что $n_{dwt} > 0$ или $\varphi_{wt}\theta_{td} > 0$
6 $\delta := n_{dw}\varphi_{wt}\theta_{td}/Z$;
7 увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$ на $(\delta - n_{dwt})$;
8 $n_{dwt} := \delta$;
9 **если** пора обновить параметры Φ, Θ **то**
10 $\varphi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W,\ t \in T$ таких, что $\hat{n}_{wt}$ изменился;
11 $\theta_{td} := \hat{n}_{dt}/\hat{n}_d$ для всех $d \in D,\ t \in T$ таких, что $\hat{n}_{dt}$ изменился;
12 **пока** Θ и Φ не сойдутся;

переменных φ_{wt}, θ_{td}, которые зависят от изменившихся скрытых переменных. Для этого случая также имеются обоснования сходимости [38].

В случае PLSA сокращение M-шага сводится к более частому обновлению параметров θ_{td} и φ_{wt} по значениям счётчиков $\hat{n}_{wt}$ и $\hat{n}_{dt}$. В Алгоритме 2.1 обновления происходят после каждого прохода коллекции. Возможные и другие варианты обновлений: после каждого документа, после каждого термина (d, w), после заданного числа терминов, после каждого вхождения термина. В Алгоритме 2.2 моменты обновления выбираются на шаге 9.

На больших коллекциях частые обновления повышают скорость сходимости. Проведённые в [2] эксперименты показывают, что частота обновления влияет именно на скорость сходимости и почти не влияет на значение правдоподобия в конце итераций. Отсюда следует практическая рекомендация делать обновления после каждого термина, при этом каждый термин документа обрабатывается только один раз. Этот способ имеет дополнительное преимущество — внутри алгоритма можно отказаться от хранения матриц Θ и Φ, поскольку значения θ_{td} и φ_{wt} требуются только на шаге 6, и их можно вычислять «на лету» по тем же формулам, что на шаге 10 и 11. Обновления после каждого вхождения термина являются избыточно частыми, в этом случае каждый термин документа приходится обрабатывать n_{dw} раз.

На первой итерации (т. е. при первом проходе коллекции) частые обновления не делаются, чтобы в счётчиках накопилась информация по всей коллекции. В противном случае оценки параметров θ_{td} и φ_{wt} по начальному фрагменту выборки могут оказаться хуже начального приближения. Начиная со второй итерации, для каждой пары (d, w) из счётчиков $\hat{n}_{wt}$ и $\hat{n}_{dt}$ вычитается n_{dwt} — то самое значение δ, которое было к ним прибавлено при обработке пары (d, w) на предыдущей итерации. Таким образом, счётчики $\hat{n}_{wt}$ и $\hat{n}_{dt}$ всегда содержат результат последнего однократного прохода всей коллекции.

Необходимость хранения трёхмерной матрицы n_{dwt} делает Алгоритм 2.2 неприменимым к большим коллекциям. Этот недостаток устраняется путём реорганизации итераций, либо применением сэмплирования. Далее рассматриваются оба способа.

§2.3 Онлайновый EM-алгоритм

На больших коллекциях Алгоритмы 2.1 и 2.2 могут сходиться очень медленно. Причина в том, что за однократный проход по всем документам коллекции оценки распределений терминов в темах $\varphi_{wt} = \hat{n}_{wt}/\hat{n}_t$ уточняются огромное число раз и успевают сойтись, в то же время распределения тем в документах θ_d проходят лишь одну итерацию. На начальных итерациях, пока распределения θ_d не сошлись, вычислительный ресурс тратится впустую на достижение сходимости φ_t к приближениям, далёким от оптимальных. Суть этой проблемы в том, что параметры θ_{td} привязаны к отдельным документам d, а параметры φ_{wt} — ко всей коллекции. Проблема решается реорганизацией шагов итерационного процесса.

Пакетный алгоритм. Проход каждого документа $d \in D$ производится несколько раз подряд. На каждом проходе документа выполняется E-шаг и обновляется распределение θ_d. Обновление распределений φ_t производится после каждого прохода коллекции, как в Алгоритме 2.1. В результате распределения φ_t и θ_d сходятся более согласованно. Кроме того, такая реорганизация итерационного процесса позволяет отказаться от хранения трёхмерных массивов H_{dwt} или n_{dwt}.

Можно также отказаться от двумерных массивов, размер которых пропорционален $|D|$, что позволит обрабатывать очень большие коллекции документов. Вероятности θ_{td} всех тем $t \in T$ документа d не нужны по окончании обработки документа d, поэтому двумерный массив $(\theta_{td})_{T\times D}$ можно заменить одномерным $(\theta_t)_T$.

Хранение двумерного массива $(\theta_{td})_{T\times D}$ всё же имеет смысл, если размер коллекции $|D|$ относительно невелик, и на шаге 5 инициализация θ_{td} производится только во время первого прохода коллекции, а при последующих проходах используется текущая оценка θ_{td}, оставшаяся с предыдущего прохода. Хорошее начальное приближение обеспечивает сходимость θ_d за меньшее число итераций.

Скорость сходимости зависит от выбора числа итераций на внутреннем цикле по документу и внешнем цикле по коллекции. На начальных итерациях внешнего цикла можно делать меньше итераций внутреннего цикла, поскольку нет смысла добиваться сходимости θ_d, пока распределения φ_t далеки от оптимальных.

Ещё одна идея ускорения сходимости состоит в том, чтобы начальные итерации провести не по всей коллекции, а по случайному подмножеству (пакету) документов $D' \subseteq D$. Если коллекция имеет избыточный размер, то для получения хорошего приближения Φ достаточно будет просмотреть относительно небольшую её часть.

Алгоритм 2.3 назван *пакетным* (batch algorithm), так как он может обрабатывать коллекцию по частям. Развитие этой идеи приводит к онлайновому алгоритму [21], одному из самых быстрых в тематическом моделировании. Он реализован в библиотеке онлайновых алгоритмов Vowpal Wabbit Джона Лэнгфорда.

Онлайновый алгоритм. В машинном обучении *онлайновыми* принято называть алгоритмы, способные адекватно настраивать параметры модели за один проход по выборке. Онлайновые алгоритмы используются для обработки потоковых данных.

Алгоритм 2.3. PLSA-BatchEM: пакетный EM-алгоритм для модели PLSA.

Вход: коллекция документов D, число тем $|T|$;
Выход: распределения Θ и Φ;

инициализировать φ_{wt} для всех $w \in W$, $t \in T$;
повторять
  $\hat{n}_{wt} := 0$; $\hat{n}_t := 0$ для всех $w \in W$, $t \in T$;
  для всех $d \in D$
    инициализировать θ_{td} для всех $t \in T$;
    повторять
      $Z_w := \sum_{t \in T} \varphi_{wt}\theta_{td}$ для всех $w \in d$;
      $\theta_{td} := \frac{1}{n_d} \sum_{w \in d} n_{dw}\varphi_{wt}\theta_{td}/Z_w$ для всех $t \in T$;
    пока θ_d не сойдётся;
    увеличить $\hat{n}_{wt}$, $\hat{n}_t$ на $n_{dw}\varphi_{wt}\theta_{td}/Z_w$ для всех $w \in d$, $t \in T$;
  $\varphi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$;
пока Φ не сойдутся;

Во многих приложениях тематического моделирования коллекция документов пополняется динамически, и требуется обновлять модель, не обрабатывая заново всю коллекцию. Обновление модели предполагает вычисление распределения $\theta_{td} = p(t \mid d)$ для нового документа d и уточнение распределений $\varphi_{wt} = p(w \mid t)$ для всех тем $t \in T$, имеющих ненулевые вероятности для слов документа d. Если коллекция уже имеет большой объём, то добавление документа не сильно влияет на Φ. Чем больше коллекция, тем лучше текущее приближение Φ, и тем меньше итераций потребуется для добавления нового документа.

Стратегии ускорения сходимости. Первый пакет — особенный! Для первого пакета: ToDo[5]
1) число итераций θ_d ограничить сверху числом прошедших итераций Φ +несколько.
2) инициализировать двумерный массив θ_{td} только при 1-м проходе.

Онлайновый Алгоритм 2.4 является модицификацией пакетного Алгоритма 2.3. Теперь вся коллекция разбивается на пакеты документов $D_1, D_2, \ldots, D_j, \ldots$ Способ разбиения остаётся на усмотрение разработчика, в частности, пакеты могут пересекаться либо не пересекаться, просматриваться по одному разу либо многократно, выбираться случайно, по времени поступления или публикации документов, и т. д. Желательно, чтобы размер первого пакета $|D_1|$ был достаточным для оценивания матрицы Φ с приемлемой точностью. Обработка каждого пакета производится пакетным Алгоритмом 2.3 при фиксированных φ_{wt}. Затем счётчики $\tilde{n}_{wt}$, вычисленные по обработанному пакету документов, складываются со счётчиками $\hat{n}_{wt}$, вычисленными по всем предыдущим пакетам.

Режимы работы онлайнового алгоритма отличаются для больших и малых, динамических и статических коллекций. Для большой коллекции достаточно одного прохода, так как матрица Φ стабилизируется после нескольких тысяч первых документов и далее почти не меняется. Малые коллекции, наоборот, требуют многократных проходов. Динамические коллекции (например, новостные сообщения) имеют

Алгоритм 2.4. PLSA-OEM: онлайновый EM-алгоритм для модели PLSA.

Вход: коллекция документов D, число тем $|T|$, параметр ρ_j;
Выход: распределения Θ и Φ;

инициализировать φ_{wt} для всех $w \in W,\ t \in T$;
$\hat{n}_{wt} := 0,\ \hat{n}_t := 0$ для всех $w \in W,\ t \in T$;
для всех пакетов $D_j,\ j = 1, \ldots, J$
  повторять
    $\tilde{n}_{wt} := 0,\ \tilde{n}_t := 0$ для всех $w \in W,\ t \in T$;
    для всех $d \in D_j$
      инициализировать θ_{td} для всех $t \in T$;
      повторять
        $Z_w := \sum_{t \in T} \varphi_{wt}\theta_{td}$ для всех $w \in d$;
        $\theta_{td} := \frac{1}{n_d} \sum_{w \in d} n_{dw}\varphi_{wt}\theta_{td}/Z_w$ для всех $t \in T$;
      пока θ_d не сойдётся;
      увеличить $\tilde{n}_{wt}$, $\tilde{n}_t$ на $n_{dw}\varphi_{wt}\theta_{td}/Z_w$ для всех $w \in d,\ t \in T$;
    $\varphi_{wt} := \dfrac{\rho_j\hat{n}_{wt} + \tilde{n}_{wt}}{\rho_j\hat{n}_t + \tilde{n}_t}$ для всех $w \in W,\ t \in T$ таких, что $\tilde{n}_{wt} > 0$;
  пока Φ не сойдутся;
  $\hat{n}_{wt} := \rho_j\hat{n}_{wt} + \tilde{n}_{wt}$ для всех $w \in W,\ t \in T$;
  $\hat{n}_t := \rho_j\hat{n}_t + \tilde{n}_t$ для всех $t \in T$;

изменчивую во времени тематику. В случаях малых и динамических коллекций значимость пакетов убывает по мере поступления новых. Поэтому вводится параметр $\rho_j \in (0, 1]$, управляющий скоростью забывания старых оценок. При поступлении каждого нового пакета D_j частоты слов в старых пакетах уменьшаются:

$$\hat{n}_{wt} := \rho_j\hat{n}_{wt} + \tilde{n}_{wt}.$$

В случае большой статической коллекции, для которой достаточно одного прохода, можно взять $\rho_j = 1$. Тогда по окончании первого прохода φ_{wt} будут обычными частотными оценками условных вероятностей. Несколько десятков первых пакетов всё же лучше учесть с меньшими значениями ρ_j, чтобы устранить влияние самых первых пакетов, когда матрица Φ оценивалась ещё слишком грубо.

Уменьшать параметр ρ_j имеет смысл в случаях малых коллекций, чтобы быстрее забывать оценки предыдущих проходов, а также в случаях больших пакетов $|D_j|$, когда матрица Φ неплохо оценивается по каждому отдельному пакету.

§2.4 Стохастический EM-алгоритм

В Алгоритме 2.2 для каждой пары (d, w) происходит распределение n_{dw} вхождений термина w в документ d между всеми $|T|$ темами пропорционально вероятностям $p(t \mid d, w)$. При этом приходится хранить массив значений n_{dwt} для всех тем $t \in T$. Расход памяти объёма $O(n|T|)$ может оказаться неприемлемым даже при небольшом

числе тем. В то же время, согласно гипотезе разреженности, употребление термина w в документе d связано, скорее всего, с небольшим числом тем.

Можно было бы оставлять только несколько наибольших значений n_{dwt} на каждом шаге. Однако эксперименты показывают, что эта эвристика приводит к накоплению систематической ошибки и смещению модели [2]..

Проблема разреживания условного распределения $p(t \mid d, w)$ адекватно решается с помощью стохастического EM-алгоритма (stochastic EM-algorithm, SEM) [11]. Распределение скрытой переменной t, вычисленное на E-шаге, не используется непосредственно на M-шаге. Вместо этого из него сэмплируется искусственная выборка, по этой выборке вычисляется эмпирическое распределение, и оно уже используется в формулах M-шага. Это позволяет упростить задачу M-шага, сохранив свойства несмещённости оценок и сходимости EM-алгоритма. Размер сэмплируемой выборки является параметром метода.

В случае PLSA реализация SEM сводится к следующему: для каждой пары (d, w) сэмплируются s случайных тем t_{dwi}, $i = 1, \ldots, s$ из распределения $p(t \mid d, w)$, возможно, повторяющихся. В формулах M-шага вместо распределения $p(t \mid d, w)$ используется его несмещённая эмпирическая оценка:

$$\hat{p}(t \mid d, w) = \frac{1}{s} \sum_{i=1}^{s} \big[t_{dwi} = t\big]. \tag{2.6}$$

Модификация Алгоритма 2.2, трансформирующая его в стохастический обобщённый EM-алгоритм (PLSA-SEM), состоит из трёх изменений:

1) перед шагом 5 сэмплируется s тем t_{dwi}, $i = 1, \ldots, s$ из $p(t \mid d, w)$;

2) на шаге 5 цикл по всем $t \in T$ заменяется циклом по $t = t_{dwi}$, $i = 1, \ldots, s$;

3) на шаге 6 вычисляется $\delta := n_{dw}/s$.

При $s = n_{dw}$ стохастический EM-алгоритм соответствует *сэмплированию Гиббса* [62] — одному из основных методов обучения вероятностных тематических моделей. В [1, 2] предложено *экономное сэмплирование*, когда s уменьшается до 3–5 тем, что приводит к большему разреживанию и экономии вычислительных ресурсов без существенной потери качества тематической модели.

Сэмплирование выборки из дискретного распределения. Пусть имеется дискретное распределение $p(t)$, $t \in T$, и требуется получить реализацию случайной величины из этого распределения. Для этого разобьём отрезок $[0, 1]$ на $|T|$ частей длины $p(t)$ каждый. С помощью датчика случайных чисел сгенерируем случайное число r из равномерного распределения на отрезке $[0, 1]$. Номер отрезка t, в который попало число r, и будет искомой реализацией. Чтобы сгенерировать выборку из s независимых наблюдений, повторим эту процедуру s раз.

§2.5 Начальные приближения

Начальные приближения φ_t и θ_d можно задавать нормированными случайными векторами из равномерного распределения.

Другая распространённая рекомендация — пройти по всей коллекции, выбрать для каждой пары (d, w) случайную тему t и вычислить частотные оценки (1.4) вероятностей φ_{wt} и θ_{td} для всех $d \in D$, $w \in W$, $t \in T$.

Инициализация с частичным обучением применяется в случаях, когда темы известны заранее и имеются дополнительные данные о привязке некоторых документов или терминов к темам. Учёт этих данных улучшает интерпретируемость тем.

Если известно, что документ d относится к подмножеству тем $T_d \subset T$, то в качестве начального θ_{td} можно взять равномерное распределение на этом подмножестве:

$$\theta_{td}^0 = \frac{1}{|T_d|}\big[t \in T_d\big]. \tag{2.7}$$

Если известно, что подмножество терминов $W_t \subset W$ относится к теме t, то в качестве начального φ_{wt} можно взять равномерное распределение на W_t:

$$\varphi_{wt}^0 = \frac{1}{|W_t|}\big[w \in W_t\big]. \tag{2.8}$$

Если известно, что подмножество документов $D_t \subset D$ относится к теме t, то можно взять эмпирическое распределение слов в объединённом документе:

$$\varphi_{wt}^0 = \frac{\sum_{d \in D_t} n_{dw}}{\sum_{d \in D_t} n_d}.$$

Если нет никакой априорной информации о связи документов с темами, то последнюю формулу можно применить к случайным подмножествам документов D_t. В [20] предлагается брать один случайный документ.

Инициализация Θ по Φ. Если для всех тем известны начальные приближения φ_{wt}^0, то первая итерация EM-алгоритма при равномерном распределении $\theta_{td}^0 = 1/|T|$ даёт ещё одну интуитивно очевидную формулу инициализации:

$$\theta_{td} = \frac{1}{n_d}\sum_{w \in d} n_{dw} H_{dwt} = \sum_{w \in d} \frac{n_{dw}}{n_d} \frac{\varphi_{wt}}{\sum_s \varphi_{ws}} = \sum_{w \in d} \hat{p}(w \,|\, d)\, \hat{p}(t \,|\, w). \tag{2.9}$$

Здесь распределение тем в документе d оценивается путём усреднения распределений тем $p(t \,|\, w)$ по словам документа d, вычисленных по формуле Байеса.

Сглаживание. Если полученное начальное приближение φ_{wt}^0 или θ_{td}^0 содержит нулевые вероятности, то его необходимо сгладить, смешав с каким-нибудь неразреженным распределением. Например, φ_{wt}^0 смешивается с эмпирическим распределением слов во всей коллекции и со случайным распределением $\rho(w)$, при некоторых значениях параметров смеси τ_1 и τ_2:

$$\varphi_{wt} = (1 - \tau_1 - \tau_2)\varphi_{wt}^0 + \tau_1 n_w / n + \tau_2 \rho(w).$$

Эксперименты и рекомендации по способам задания начального приближения. ToDo[6]

3 Латентное размещение Дирихле

Основным недостатком PLSA считается высокая размерность пространства параметров, вызывающая переобучение [9]. В задачах машинного обучения для сокращения размерности обычно используется либо *отбор признаков*, приводящий к уменьшению числа параметров, либо *регуляризация* — наложение дополнительных ограничений на параметры. В частности, *байесовская регуляризация* основана на введении априорного распределения вероятности в пространстве параметров.

Тематическая модель *латентного размещения Дирихле* (latent Dirichlet allocation, LDA) [9] основана на разложении (1.2) при дополнительном предположении, что векторы документов $\theta_d = (\theta_{td}) \in \mathbb{R}^{|T|}$ и векторы тем $\varphi_t = (\varphi_{wt}) \in \mathbb{R}^{|W|}$ порождаются распределениями Дирихле с параметрами $\alpha \in \mathbb{R}^{|T|}$ и $\beta \in \mathbb{R}^{|W|}$ соответственно:

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

$$\operatorname{Dir}(\varphi_t;\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \varphi_{wt}^{\beta_w - 1}, \quad \beta_w > 0, \quad \beta_0 = \sum_w \beta_w, \quad \varphi_{wt} > 0, \quad \sum_w \varphi_{wt} = 1.$$

где $\Gamma(z)$ — гамма-функция.

Некоторые свойства распределения Дирихле. Математическое ожидание и дисперсия t-й координаты вектора θ_d равны, соответственно,

$$\mathsf{E}\theta_{td} = \int \theta_{td} \operatorname{Dir}(\theta_d;\alpha)\, d\theta_d = \frac{\alpha_t}{\alpha_0}, \qquad \mathsf{D}\theta_{td} = \frac{\alpha_t(\alpha_0 - \alpha_t)}{\alpha_0^2(\alpha_0 + 1)}. \tag{3.1}$$

Векторный параметр α определяет степень разреженности векторов θ_d, порождаемых распределением $\operatorname{Dir}(\theta;\alpha)$. Если $\alpha_t = 1$ для всех t, то распределение Дирихле переходит в равномерное. Чем больше α_0, тем меньше дисперсия, и тем сильнее векторы θ_d концентрируется вокруг вектора математического ожидания $\mathsf{E}\theta_d$. Чем меньше α_t, тем сильнее значения θ_{td} концентрируются вокруг нуля. Чем меньше α_0, тем более разрежен вектор θ_d. Поэтому α_t называют *параметрами контраста*.

Обоснования. Есть несколько доводов в пользу распределения Дирихле как байесовского регуляризатора вероятностных тематических моделей.

Во-первых, это достаточно широкое параметрическое семейство распределений на единичном симплексе, которое описывает как разреженные, так и сконцентрированные дискретные распределения.

Во-вторых, модель LDA хорошо подходит для описания кластерных структур. Чем меньше значения гиперпараметров α и β, тем сильнее разрежено распределение Дирихле, и тем дальше отстоят друг от друга порождаемые им векторы. Чем меньше α_0, тем сильнее различаются документы θ_d. Чем меньше β_0, тем сильнее различаются темы φ_t. Векторы $\varphi_t = p(w \mid t)$ в пространстве терминов $\mathbb{R}^{|W|}$ представляют центры тематических кластеров. Элементами кластеров являются векторы документов с эмпирическими распределениями $\hat{p}(w \mid d, t)$. Чем меньше гиперпараметры β, тем больше межкластерные расстояния по сравнению с внутрикластерными. Таким образом, гиперпараметры позволяют моделировать тематические кластерные структуры различной степени выраженности.

В-третьих, распределение Дирихле является сопряжённым к мультиномиальному, что упрощает вывод апостериорных оценок вероятностей θ_{td} и φ_{wt}. Именно математическое удобство распределения Дирихле в значительной степени определяет его популярность в тематическом моделировании.

Недостатки. Основной недостаток распределения Дирихле — отсутствие убедительных лингвистических обоснований. Предположение, что все распределения θ_d, $d \in D$ порождаются распределением Дирихле, да ещё и одним и тем же, кажется весьма произвольным. То же можно сказать и о порождении множества распределений φ_t для всех тем $t \in T$. Второй недостаток заключается в том, что параметры θ_{td} и φ_{wt} не могут обращаться в нуль, что противоречит гипотезе разреженности.

§3.1 Байесовский вывод

Рассмотрим процесс порождения документа d как выборки n_d пар тема–термин $X_d = \{(t_1, w_1), \ldots, (t_{n_d}, w_{n_d})\}$. В каждой паре (t_i, w_i) тема t_i выбирается из дискретного распределения $p(t \,|\, d) = \theta_{td}$. Следовательно, вероятность встретить каждую из тем t ровно n_{td} раз подчиняется мультиномиальному распределению:

$$p(X_d|\theta_d) = \frac{n_d!}{\prod_t n_{td}!} \prod_t \theta_{td}^{n_{td}}.$$

Распределение Дирихле является *сопряжённым* к мультиномиальному. Это означает, что при априорном распределении Дирихле $\theta_d \sim \mathrm{Dir}(\theta; \alpha)$ апостериорное распределение вектора θ_d принадлежит тому же семейству распределений, но с другим значением параметра: $\theta_d|X_d \sim \mathrm{Dir}(\theta; \alpha')$. Действительно, по формуле Байеса

$$p(\theta_d|X_d, \alpha) = \frac{p(X_d|\theta_d)\,\mathrm{Dir}(\theta_d; \alpha)}{p(X_d)} = C \prod_t \theta_{td}^{n_{td}} \theta_{td}^{\alpha_t - 1} = \mathrm{Dir}(\theta_d; \alpha'), \quad \alpha'_t = \alpha_t + n_{td},$$

где C — нормировочная константа, не зависящая от θ_d.

Оценим случайную величину θ_{td} её математическим ожиданием (3.1) по апостериорному распределению:

$$p(t|d, X_d, \alpha) = \int p(t|d) p(\theta_d|X_d, \alpha)\, d\theta_d = \int \theta_{td}\, \mathrm{Dir}(\theta_d, \alpha')\, d\theta_d = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}. \tag{3.2}$$

Заменив величину n_{td} её оценкой $\hat{n}_{td}$, получим байесовскую оценку параметра θ_{td} для EM-алгоритма, отличающуюся от оценки максимума правдоподобия (2.4) сглаживающими слагаемыми в числителе и знаменателе:

$$\theta_{td} = \frac{\hat{n}_{dt} + \alpha_t}{\hat{n}_d + \alpha_0}. \tag{3.3}$$

Аналогично выводится сглаженная байесовская оценка и для φ_{wt}:

$$\varphi_{wt} = \frac{\hat{n}_{wt} + \beta_w}{\hat{n}_t + \beta_0}. \tag{3.4}$$

Замена в обобщённом EM-алгоритме частотных оценок условных вероятностей (2.3) и (2.4) сглаженными оценками (3.3) и (3.4) трансформирует PLSA в LDA. Более строгое обоснование EM-подобных алгоритмов приводится в [50, 62] для метода сэмплирования Гиббса и в [54] для метода вариационной байесовской аппроксимации.

В [8] показано, что эти и другие известные алгоритмы обучения LDA являются вариантами EM-алгоритма и отличаются, главным образом, формулой сглаживания частотных оценок вероятностей. Оптимизация гиперпараметров α и β, предложенная в [58, 59], ещё сильнее нивелирует различия между моделями. Согласно экспериментам на 7 текстовых коллекциях [8], более эффективным по качеству и по времени является алгоритм *свёрнутой вариационной байесовской аппроксимации* CVB0 (collapsed variational Bayes). В нашей нотации ему наиболее близок LDA-GEM.

Байесовский подход на данный момент доминирует в тематическом моделировании. Большинство специализированных моделей строятся на основе LDA. Такая популярность вызвана не лингвистической обоснованностью LDA, а математическим удобством — сопряжённые распределения допускают аналитическое интегрирование по пространству параметров модели. Однако при попытке усложнения модели, построения многофункциональных или композитных моделей, байесовский вывод усложняется настолько, что начинает сдерживать развитие тематического моделирования. Ниже, в разделе 5, мы рассмотрим новый подход — *аддитивную регуляризацию тематических моделей*, в котором те же оценки получаются гораздо проще, без байесовского вывода, априорных распределений Дирихле и интегрирования.

§3.2 Сэмплирование Гиббса

Сэмплирование Гиббса (Gibbs sampling, GS) применяется для решения задач статистического оценивания, когда вычисление или хранение функции распределении слишком ресурсоёмко, в то же время, генерация случайной выборки из этого распределения не вызывает затруднений. Тогда вместо исходного распределения используется его несмещённая эмпирическая оценка по сэмплированной выборке.

Применение GS к тематической модели LDA предложено в [50], см. Алгоритм 3.1. Строгие выкладки приводятся в отчёте [62]. Однако LDA-GS можно понимать и намного проще — как специальный случай стохастического EM-алгоритма PLSA-SEM, в котором для каждой пары (d, w) сэмплируется ровно $s = n_{dw}$ тем, а параметры φ_{wt}, θ_{td} обновляются после каждого вхождения термина w в документ d.

Существенное отличие LDA-GS от PLSA-SEM — в использовании сглаженных оценок условных вероятностей (3.3) и (3.4). В экспериментах сэмплирование Гиббса действительно плохо работает без сглаживания [2]. При сэмплировании на шаге 6 все темы должны иметь ненулевые шансы быть выбранными из распределения $p(t \mid d, w)$, особенно на начальных итерациях. На выходе Алгоритма 3.1, наоборот, можно выдать несмещённые оценки искомых условных вероятностей, см. шаги 9, 10.

Ещё одно, чисто техническое, отличие LDA-GS от PLSA-SEM — на шаге 5 счётчики уменьшаются на единицу, тем самым i-е вхождение термина w в документ d не учитывается в оценке распределения $p(t \mid d, w)$, из которого сэмплируется тема t_{dwi}. Из теории следует, что эта деталь исключительно важна [62]. Однако в экспериментах она практически не влияет на качество модели [1, 2]. Счётчики можно одновременно уменьшать для старой темы и увеличивать для новой, как в Алгоритме 2.2.

Таким образом, LDA-GS отличается от PLSA-GEM тремя эвристиками: частотой обновления параметров, сэмплированием и сглаживанием. Эти эвристики не связаны друг с другом и могут применяться в любых сочетаниях, порождая целое семейство алгоритмов тематического моделирования.

Алгоритм 3.1. LDA-GS: сэмплирование Гиббса для тематической модели LDA.

Вход: коллекция D, число тем $|T|$, гиперпараметры α, β;
Выход: распределения Θ и Φ;
обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D,\ w \in W,\ t \in T$;
повторять
  для всех $d \in D,\ w \in d,\ i = 1, \ldots, n_{dw}$
    если не первая итерация **то**
      $t := t_{dwi}$; уменьшить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на 1;
    сэмплировать тему t_{dwi} из $p(t \mid d, w) \propto \dfrac{\hat{n}_{wt} + \beta_w}{\hat{n}_t + \beta_0} \dfrac{\hat{n}_{dt} + \alpha_t}{n_d + \alpha_0}$;
    $t := t_{dwi}$; увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на 1;
пока $\hat{n}_{wt}$, $\hat{n}_{dt}$ не стабилизируются;
$\varphi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $t \in T,\ w \in W$;
$\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D,\ t \in T$;

§3.3 Оптимизация гиперпараметров

В первых работах по LDA [9] и сэмплированию Гиббса [50], а также в последовавших за ними исследованиях использовались симметричные распределения Дирихле с гиперпараметрами $\alpha = (a, \ldots, a)$ и $\beta = (b, \ldots, b)$. Скалярные гиперпараметры a и b либо фиксировались (одна из стандартных рекомендаций: $a = 50/|T|$, $b = 0.01$), либо настраивались путём перебора по сетке значений.

Позже были предложены численные методы оптимизации гиперпараметров, их обзор и сравнение приводится в диссертации [58]. Большинство методов оптимизации гиперпараметров основаны на максимизации *обоснованности* (evidence) модели, определяемой по всей коллекции $X = (X_d)_{d \in D}$:

$$P(X|\alpha) = \int P(X|\theta) p(\theta|\alpha)\, d\theta = \int \prod_{d \in D} P(X_d|\theta_d)\, \mathrm{Dir}(\theta_d; \alpha)\, d\theta_d = \\ = \prod_{d \in D} \frac{\Gamma(\alpha_0)}{\Gamma(n_d + \alpha_0)} \prod_{t \in T} \frac{\Gamma(n_{td} + \alpha_t)}{\Gamma(\alpha_t)} \to \max_{\alpha}$$

Метод неподвижной точки [36] — один из самых простых, но не самый лучший — представляет собой итерационный процесс:

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

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

Этот или другой аналогичный итерационный процесс встраивается между проходами по коллекции, когда значения счётчиков n_{td} и n_d вычислены и фиксированы. Он выполняется намного быстрее одного прохода коллекции. Таким образом, оптимизация гиперпараметров является вычислительно эффективной.

Эксперименты показали, что оптимизация гиперпараметров существенно улучшает качество тематической модели [59]. Оказалось, что априорное распределение

$\mathrm{Dir}(\theta;\alpha)$ лучше брать несимметричным и оптимизировать вектор гиперпараметров $\alpha = (\alpha_1, \ldots, \alpha_{|T|})$, а распределение $\mathrm{Dir}(\varphi;\beta)$ лучше брать симметричным и оптимизировать скалярный гиперпараметр b, причём $0 < b \ll 1$.

Здесь есть неточность, т.к. в статье вводилось двухэтажное распределение Дирихле. ToDo[7]

Лирическое отступление: метод простых итераций. Понятие неподвижной точки. Условия сходимости. ToDo[8]

§3.4 Действительно ли сглаживание необходимо?

Согласно экспериментам [9], качество модели LDA существенно превосходит PLSA. По аналогии с задачами классификации и регрессии отсюда был сделан стандартный вывод, что модель PLSA имеет слишком много параметров θ_{td}, φ_{wt}, что и приводит к переобучению. Байесовская регуляризация накладывает ограничения на параметры, следовательно, должна сокращать эффективную размерность и уменьшать переобучение. Однако корректное сравнение PLSA и LDA показывает, что регуляризация Дирихле в тематических моделях играет совсем другую роль.

Регуляризация Дирихле приводит к сглаживанию частотных оценок условных вероятностей (3.3)–(3.4), что является единственным принципиальным отличием LDA от PLSA. В экспериментах на реальных данных оптимальные значения гиперпараметров α_t и β_w оказываются близки к нулю [59]. Оценки параметров φ_{wt} и θ_{td} в PLSA и LDA заметно отличаются только для терминов, редких в теме, и тем, редких в документе, которые не несут статистически значимой информации о тематике коллекции. Вообще, вероятностное тематическое моделирование основано на предположении, что «тема» — это статистическое явление, связанное с *частым* употреблением определённых терминов. *Редкие* темы и термины должны были бы игнорироваться как шум, но вместо этого LDA, наоборот, повышает оценку их вероятности.

Утверждение о том, что LDA сокращает эффективную размерность пространства параметров [9], звучит неубедительно. PLSA и LDA оценивают параметры φ_{wt} и θ_{td} по одним и тем же формулам (3.3)–(3.4). Более того, в LDA вводятся дополнительные гиперпараметры α_t, β_w, которые тоже приходится оценивать.

Утверждение о том, что LDA гораздо меньше переобучается [9], также нуждается в перепроверке. Качество тематических моделей принято сравнивать по контрольной перплексии, которая может резко повышаться при появлении в контрольных документах редких терминов, имеющих вероятность $p(w \mid d)$, близкую к нулю. В PLSA эта вероятность может вообще оказаться равной нулю, и тогда перплексия формально будет равна $+\infty$. Это выглядит как переобучение, но свидетельствует скорее о неадекватности меры качества и самой модели PLSA, которая исключает возможность появления в текстах нетематических терминов. Модель LDA менее чувствительна к шуму за счёт завышенных частотных оценок условных вероятностей φ_{wt} и θ_{td}. Однако это не решает проблему выделения шума, а лишь скрывает её.

Если из контрольных документов убрать небольшое число наиболее редких терминов или если использовать робастные модели PLSA и LDA, то контрольные перплексии PLSA и LDA практически совпадают [1, 2]. Исследования [33, 63, 31] также подтверждают, что для больших коллекций нет существенных различий в качестве моделей PLSA и LDA. Значимые отличия контрольной перплексии PLSA и LDA в ранних экспериментах [9] объясняются различиями в реализациях алгоритмов обу-

чения. В экспериментах [1, 2] для обучения моделей PLSA и LDA использовался один и тот же алгоритм, отличавшийся только сглаженными оценками в LDA.

Таким образом, роль регуляризации Дирихле в LDA оказывается весьма скромной. Это не сокращение размерности и не уменьшение переобучения, а всего лишь более осторожное оценивание вероятностей шумовых терминов и тем. Кроме того, сглаживание необходимо в алгоритме сэмплирования Гиббса, чтобы все темы из распределения $p(t \mid d, w)$ имели шансы реализоваться; однако это требование скорее техническое и связано с конкретным методом.

Переобучение — не единственный мнимый недостаток PLSA, на который указывает Д. Блэй в [9], мотивируя переход к «более прогрессивной» модели LDA. Вторая претензия заключается в том, что модель PLSA якобы неадекватно описывает новые документы. Действительно, для добавления (folding-in) новых документов в коллекцию Т. Хофманн предлагал в [22] фиксировать матрицу Ф, найденную по всем предыдущим документам, и определять только вектор θ_d для нового документа. Эта эвристика основана на предположении, что коллекция достаточно велика, и один документ d не может существенно повлиять на оценки распределений φ_t. Оно может не выполняться, если документ d содержит значительное число новых терминов или относится к темам, слабо представленным в коллекции. Фиксация распределений φ_t в момент оценивания распределения θ_d является, по сути дела, подменой вероятностной порождающей модели. Однако проблема легко решается путём реорганизации итерационного процесса обучения модели, что и сделано в онлайновом EM-алгоритме. Единственным существенным отличием LDA от PLSA является регуляризация Дирихле, а варианты EM-алгоритма (рациональный, обобщённый, стохастический, онлайновый и другие) могут быть применены к обеим моделям. В работе Хофманна 1999 года эти варианты просто не рассматривались, но это вовсе не означает, что модель не позволяет описывать новые документы. Выше мы рассмотрели онлайновый вариант EM-алгоритма как раз для модели PLSA.

Рассуждения Д. Блэя о недостатках PLSA были впоследствии растиражированы в многочисленных работах его последователей и учеников, без попыток критического переосмысления или какой-либо перепроверки, см. обзоры [51, 14]. Генеральной линией тематического моделирования стало развитие модели LDA на основе математического аппарата графических моделей и байесовского вывода. Однако априорные распределения Дирихле и их обобщения — процессы Дирихле и Питмана-Йора не имеют убедительных лингвистических обоснований. Более того, переход от порождающей модели к алгоритму настройки её параметров требует весьма громоздких выкладок, которые резко усложняются при введении более сложных априорных распределений или совместном моделировании нескольких языковых явлений.

По этим причинам мы будем придерживаться более простого подхода, основанного исключительно на принципе максимума правдоподобия. Мы также будем избегать везде, где это возможно, избыточных вероятностных допущений и строить «полувероятностные» модели, в которых дополнительные данные или лингвистические знания могут учитываться не только в порождающей модели, но и путём модификации функционала или непосредственно метода его оптимизации.

4 Робастная тематическая модель

Согласно вероятностной модели (1.2), каждый термин w в каждом документе d порождается некоторой темой t. Однако появление отдельных терминов может объясняться не только тематикой документа. Возможны, как минимум, ещё два альтернативных объяснения, условно называемых фоном и шумом.

Фон — это общеупотребительные слова, в частности, стоп-слова, не отброшенные на стадии предварительной обработки. Фоновые слова имеют значимые вероятности во многих темах, снижая релевантность тематического поиска.

Шум — это термины, специфичные для конкретного документа, либо редкие термины, относящиеся к темам, слабо представленным в данной коллекции. Тематическая модель даёт слишком низкие значения вероятности $p(w\,|\,d)$ для таких терминов, то есть не способна объяснить их появление в документах коллекции.

Описание фона и шума тематической моделью лишено смысла. Необходимо каким-то образом исключать их из функционала правдоподобия. При этом неизвестно, какие именно термины являются фоновыми и шумовыми, но известно, что в целом по коллекции их относительно немного.

§4.1 Тематическая модель с шумом и фоном

Робастная вероятностная тематическая модель SWB (special words with background) представляет собой вероятностную смесь трёх компонент — тематической, шумовой и фоновой [12]:

$$p(w\,|\,d) = \frac{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}{1+\gamma+\varepsilon}; \qquad Z_{dw} = \sum_{t\in T}\varphi_{wt}\theta_{td}. \tag{4.1}$$

где *шумовая компонента* $\pi_{dw} \equiv p_{\text{ш}}(w\,|\,d)$ — неизвестное распределение терминов в документе d, *фоновая компонента* $\pi_w \equiv p_{\text{ф}}(w)$ — неизвестное распределение терминов во всей коллекции. Априорные вероятности тематической, шумовой и фоновой компонент модели обозначим, соответственно, $q_{\text{т}} = \frac{1}{1+\gamma+\varepsilon}$, $q_{\text{ш}} = \frac{\gamma}{1+\gamma+\varepsilon}$, $q_{\text{ф}} = \frac{\varepsilon}{1+\gamma+\varepsilon}$, где γ и ε — неотрицательные параметры.

Суть робастной модели в том, что если тематическая компонента Z_{dw} плохо объясняет избыточную частоту n_{dw} некоторого термина w в некотором документе d, то она может быть объяснена альтернативным образом либо шумовой компонентной π_{dw}, либо фоновой π_w. Таким образом, редкие и общеупотребительные слова в явном виде описываются как нетематические.

Требуется найти значения вероятностей φ_{wt}, θ_{td}, π_{dw}, π_w, при которых логарифм правдоподобия достигает максимума:

$$L(\Phi,\Theta,\Pi) = \sum_{d\in D}\sum_{w\in d} n_{dw}\ln\frac{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}{1+\gamma+\varepsilon} \;\to\; \max_{\Phi,\Theta,\Pi}, \tag{4.2}$$

при ограничениях неотрицательности $\pi_{dw} \geqslant 0$, $\pi_w \geqslant 0$ и нормировки

$$\sum_{w\in W}\varphi_{wt} = 1, \quad \sum_{t\in T}\theta_{td} = 1, \quad \sum_{w\in d}\pi_{dw} = 1, \quad \sum_{w\in W}\pi_w = 1.$$

Чтобы получить формулы М-шага, запишем лагранжиан данной задачи при ограничениях нормировки и неотрицательности π_{dw}, π_w, проигнорировав ограничения неотрицательности θ_{td} и φ_{wt}, которые будут выполнены автоматически.

$$\mathscr{L}(\Phi,\Theta,\Pi) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \frac{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}{1+\gamma+\varepsilon} - \\ - \sum_{t\in T}\lambda_t\left(\sum_{w\in W}\varphi_{wt} - 1\right) - \sum_{d\in D}\mu_d\left(\sum_{t\in T}\theta_{td} - 1\right) - \\ - \sum_{d\in D}\nu_d\left(\sum_{w\in d}\pi_{dw} - 1\right) + \sum_{d\in D}\sum_{w\in d}\kappa_{dw}\pi_{dw} - \\ - \nu'\left(\sum_{w\in W}\pi_w - 1\right) + \sum_{w\in W}\kappa'_w\pi_w.$$

Двойственные переменные κ_{dw}, соответствующие ограничениям $\pi_{dw} \geqslant 0$, должны быть неотрицательны и удовлетворять условиям дополняющей нежёсткости

$$\kappa_{dw}\pi_{dw} = 0, \quad d \in D,\ w \in d.$$

Аналогично, для двойственных переменных κ'_w, соответствующих $\pi_w \geqslant 0$:

$$\kappa'_w\pi_w = 0, \quad w \in W.$$

По аналогии со стандартным EM-алгоритмом, на Е-шаге для каждой пары (d,w) вычисляются по формуле Байеса условные вероятности тем $H_{dwt} = p(t \mid d,w)$:

$$H_{dwt} = \frac{\varphi_{wt}\theta_{td}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}, \quad t \in T, \tag{4.3}$$

а также условные вероятности того, что термин w является шумом H_{dw} и фоном H'_{dw}:

$$H_{dw} = \frac{\gamma\pi_{dw}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}; \qquad H'_{dw} = \frac{\varepsilon\pi_w}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}. \tag{4.4}$$

Продифференцировав лагранжиан по переменным θ_{td} и φ_{wt} и приравняв нулю производные, получим прежние формулы для φ_{wt} (2.3) и θ_{td} (2.4), с единственным отличием, что теперь H_{dwt} вычисляются по новой формуле (4.3).

Продифференцируем лагранжиан по π_{dw} и приравняем нулю производную:

$$\nu_d = \frac{n_{dw}\gamma}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w} + \kappa_{dw}. \tag{4.5}$$

Домножим обе части этого равенства на π_{dw}, просуммируем по всем терминам $w \in W$, применим условие нормировки вероятностей π_{dw} в левой части и условие дополняющей нежёсткости в правой части. Получим выражение двойственной переменной ν_d через все основные переменные:

$$\nu_d = \sum_{w\in d} n_{dw}\frac{\gamma\pi_{dw}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w} = \sum_{w\in d} n_{dw}H_{dw}. \tag{4.6}$$

Поскольку H_{dw} есть апостериорная вероятность того, что термин w в документе d является шумом, величина ν_d интерпретируется как оценка числа шумовых терминов в документе d.

Проделав аналогичные действия для фоновой компоненты, получим

$$\nu' = \sum_{d\in D} n_{dw} \frac{\varepsilon}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w} + \kappa'_w,$$
$$\nu' = \sum_{d\in D}\sum_{w\in d} n_{dw} \frac{\varepsilon\pi_w}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w} = \sum_{d\in D}\sum_{w\in d} n_{dw} H'_{dw},$$

где ν' интерпретируется как оценка числа фоновых терминов во всей коллекции.

Мультипликативный М-шаг. Домножим обе части (4.5) на π_{dw}, но не будем суммировать по w. Получим формулу М-шага для шумовой компоненты:

$$\pi_{dw} = \frac{1}{\nu_d} n_{dw} \frac{\gamma\pi_{dw}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w} = \frac{n_{dw} H_{dw}}{\sum\limits_{w'\in d} n_{dw'} H_{dw'}}.$$

Аналогично получается формула М-шага для фоновой компоненты:

$$\pi_w = \frac{1}{\nu'} n_{dw} \frac{\varepsilon\pi_w}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w} = \frac{\sum\limits_{d\in D} n_{dw} H'_{dw}}{\sum\limits_{d\in D}\sum\limits_{w'\in d} n_{dw'} H'_{dw'}}.$$

Неотрицательность решения π_{dw}, π_w гарантируется, коль скоро начальные приближения π_{dw}, π_w неотрицательны. Мультипликативный М-шаг приводит к аналогичной проблеме разреженности для переменных π_{dw} и π_w, что и для переменных φ_{wt} и θ_{td}. Если в начальном приближении значение π_{dw} или π_w равно нулю, то оно сохранится и далее на протяжении итераций. Если в начальном приближении π_{dw} или π_w не равно нулю, то оно так и останется ненулевым.

Аддитивный М-шаг решает проблему разреживания шумовой компоненты [1]. Перепишем (4.5) в другом виде:

$$n_{dw}\gamma = (\nu_d - \kappa_{dw})(Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w).$$

Согласно условиям дополняющей нежёсткости, хотя бы одна из двух неотрицательных переменных κ_{dw}, π_{dw} должна быть равна нулю. Отсюда следует, что если $n_{dw}\gamma < \nu_d(Z_{dw} + \varepsilon\pi_w)$, то $\pi_{dw} = 0$ и $\kappa_{dw} > 0$. Если же имеет место противоположное неравенство, то $\kappa_{dw} = 0$ и π_{dw} находится из уравнения $n_{dw}\gamma = \nu_d(Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w)$. Объединяя оба эти случая, получаем итоговое выражение для π_{dw}:

$$\pi_{dw} = \left(\frac{n_{dw}}{\nu_d} - \frac{Z_{dw} + \varepsilon\pi_w}{\gamma}\right)_+. \tag{4.7}$$

Таким образом, если термин w в документе d встречается существенно чаще, чем предсказывают тематическая и фоновая компоненты модели, то его появление объясняется особенностями данного документа, и тогда $\pi_{dw} > 0$.

Аддитивный М-шаг, в отличие от мультипликативного, приводит к автоматическому выбору структуры разреженности матрицы $(\pi_{dw})_{D\times W}$.

Алгоритм 4.1. PLSA-ROEM: робастный онлайновый EM-алгоритм.

Вход: коллекция документов D, число тем $|T|$;
Выход: распределения Θ, Φ, Π;

инициализировать φ_{wt}, π_w для всех $w \in W,\ t \in T$;
$\hat{n}_{wt} := 0,\ \hat{n}_t := 0,\ \hat{n}'_w := 0,\ \hat{n}' := 0$ для всех $w \in W,\ t \in T$;
для всех пакетов $D_j,\ j = 1, \ldots, J$
  повторять
    $\tilde{n}_{wt} := 0,\ \tilde{n}_t := 0,\ \tilde{n}'_w := 0,\ \tilde{n}' := 0$ для всех $w \in W,\ t \in T$;
    для всех $d \in D_j$
      инициализировать θ_{td} для всех $t \in T$;
      повторять
        $Z_w := \sum_{t \in T} \varphi_{wt}\theta_{td} + \gamma\pi_{dw} + \varepsilon\pi_w$ для всех $w \in d$;
        $\nu_d := \sum_{w \in d} n_{dw}\gamma\pi_{dw}/Z_w$;
        $\tilde{n}_t := \sum_{w \in d} n_{dw}\varphi_{wt}\theta_{td}/Z_w$ для всех $t \in T$;
        $\theta_{td} := \tilde{n}_t \,/\, \sum_{s \in T} \tilde{n}_s$ для всех $t \in T$;
        $\pi_{dw} := \big(\pi_{dw} + n_{dw}/\nu_d - Z_w/\gamma\big)_+$ для всех $w \in d$;
      пока θ_{td} и π_{dw} для данного d не сойдутся;
      увеличить $\tilde{n}_{wt}$, $\tilde{n}_t$ на $n_{dw}\varphi_{wt}\theta_{td}/Z_w$ для всех $w \in d,\ t \in T$;
      увеличить $\tilde{n}'_w$, $\tilde{n}'$ на $n_{dw}\varepsilon\pi_w/Z_w$ для всех $w \in d$;
    $\varphi_{wt} := \dfrac{\rho_j\hat{n}_{wt} + \tilde{n}_{wt}}{\rho_j\hat{n}_t + \tilde{n}_t}$ для всех $w \in W,\ t \in T$;
    $\pi_w := \dfrac{\rho_j\hat{n}'_w + \tilde{n}'_w}{\rho_j\hat{n}' + \tilde{n}'}$ для всех $w \in W$;
  пока Φ не сойдутся;
  $\hat{n}_{wt} := \rho_j\hat{n}_{wt} + \tilde{n}_{wt}$ для всех $w \in W,\ t \in T$;
  $\hat{n}_t := \rho_j\hat{n}_t + \tilde{n}_t$ для всех $t \in T$;
  $\hat{n}'_w := \rho_j\hat{n}'_w + \tilde{n}'_w$ для всех $w \in W$;
  $\hat{n}' := \rho_j\hat{n}' + \tilde{n}'$;

§4.2 Робастный онлайновый EM-алгоритм

Робастная модификация PLSA-ROEM онлайнового PLSA-OEM показана в Алгоритме 4.1. Главное отличие от обычного PLSA в том, что теперь n_{dw} вхождений термина w в документ d распределяются не только между темами $t \in T$, но также между шумовой и фоновой компонентами, пропорционально вероятностям

$$\tilde{H}_{dw} = \big(\tfrac{1}{Z}\varphi_{wt}\theta_{td},\ t \in T;\ \ \tfrac{1}{Z}\gamma\pi_{dw};\ \ \tfrac{1}{Z}\varepsilon\pi_w\big),$$

где Z — нормирующий множитель.

Возможны различные варианты алгоритма PLSA-ROEM: только с шумовой компонентой ($\varepsilon = 0$), только с фоновой компонентой ($\gamma = 0$), с аддитивным и мультипликативным M-шагом. В Алгоритме 4.1 показан вариант с шумом и фоном, аддитивным M-шагом, без сэмплирования и без сглаживания.

Сглаживание вводится в Алгоритм 4.1 заменой частотных оценок (2.3)–(2.4) параметров φ_{wt}, θ_{td} на шагах 17, 12 байесовскими оценками (3.3)–(3.4).

Сэмплирование вводится заменой распределения $\tilde{H}_{dw}$ его эмпирической оценкой, аналогичной (2.6).

О невозможности оптимизации априорных вероятностей шума и фона. Приравняв нулю производные лагранжиана по γ и ε, нетрудно получить формулы для обновления γ и ε. Однако эксперименты показывают, что с итерациями $\gamma \to \infty$, $\varepsilon \to 0$, что приводит к полному вырождению тематической модели в униграммную модель документов. Поэтому параметры γ и ε необходимо фиксировать.

§4.3 Упрощённый робастный алгоритм

Недостатком предыдущей модели является необходимость подбирать параметры γ, ε и хранить параметры π_{dw}, число которых сопоставимо с размером коллекции. Рассмотрим упрощённую робастную модель, которая вообще не требует дополнительных затрат памяти или времени. Фоновая компонента в ней отсутствует, а шумовая компонента π_{dw} включается только когда $Z_{dw} = 0$, то есть когда термин w в документе d оказывается *нетематическим*:

$$p(w \mid d) = \nu_d Z_{dw} + \big[Z_{dw}{=}0\big]\pi_{dw}, \qquad Z_{dw} = \sum_{t\in T} \varphi_{wt}\theta_{td}, \tag{4.8}$$

где π_{dw} — новые параметры модели, $\pi_{dw} > 0$ тогда и только тогда, когда $Z_{dw} = 0$; параметр ν_d определяется из условия нормировки $\sum_{w\in W} p(w \mid d) = 1$.

Хотя ситуация $Z_{dw} = 0$ интерпретируется как вполне нормальная, последствия оказываются катастрофическими для стандартной вероятностной модели (1.2): в функционале правдоподобия (1.6) под логарифмом появляется нуль, распределение тем (2.1) для данного слова не существует.

Максимизация правдоподобия (1.6) для модели (4.8) снова приводит к частотным оценкам условных вероятностей (2.3)–(2.4), но теперь H_{dwt} и $\hat{n}_{dwt}$ оцениваются только по тематическим терминам:

$$\hat{n}_{dwt} = \big[Z_{dw} > 0\big] n_{dw} H_{dwt}.$$

Таким образом, при вычислении параметров φ_{wt} и θ_{td} все нетематические термины просто игнорируются.

Оптимальное значение π_{dw} достаточно определять только для тех (d, w), при которых $Z_{dw} = 0$. Оно также выражается аналитически и совпадает с *униграммной оценкой* условной вероятности $p(w \mid d)$:

$$\pi_{dw} = n_{dw}/n_d.$$

Нормировочный множитель ν_d равен доле тематических терминов в документе:

$$\nu_d = \sum_{w\in W} \big[Z_{dw} > 0\big]\pi_{dw} = \frac{1}{n_d} \sum_{w\in d} \big[Z_{dw} > 0\big] n_{dw}.$$

Значения параметров π_{dw} и ν_d не нужны для вычисления тематической компоненты модели — матриц Φ и Θ, но могут понадобиться при вычислении функционалов качества модели, непосредственно зависящих от $p(w \mid d)$.

§4.4 Выделение стоп-слов

5 Регуляризация тематических моделей

Искомое стохастическое матричное разложение $\Phi\Theta$ определено не единственным образом, а с точностью до невырожденного преобразования: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$, при условии, что матрицы $\Phi' = \Phi S$ и $\Theta' = S^{-1}\Theta$ также стохастические. Задача тематического моделирования имеет в общем случае бесконечно много решений. Неединственность решения влечёт неустойчивость ЕМ-алгоритма. Стартуя из различных начальных приближений, он будет сходиться к различным точкам бесконечного множества решений. Соответствующие эксперименты описаны в §13.2.

Задачи, решение которых неединственно или неустойчиво, называются *некорректно поставленными*. Общий подход к их решению называется *регуляризацией* [7]. Он заключается в том, чтобы некоторым разумным образом ввести дополнительные ограничения на Φ, Θ, сузив тем самым множество решений.

Допустим, что наряду с правдоподобием (1.6) требуется максимизировать ещё n критериев $R_i(\Phi, \Theta)$, $i = 1, \ldots, n$, называемых *регуляризаторами*. Для решения задачи многокритериальной оптимизации будем максимизировать линейную комбинацию критериев L и R_i с неотрицательными *коэффициентами регуляризации* τ_i, при условии неотрицательности и нормировки столбцов матриц Φ и Θ:

$$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta), \qquad L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}, \tag{5.1}$$

Решение этой задачи приводит к обобщению формул М-шага в ЕМ-алгоритме:

$$\varphi_{wt} = \frac{\big(\hat{n}_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}(\Phi, \Theta)\big)_+}{\sum\limits_{u \in W}\big(\hat{n}_{ut} + \varphi_{ut}\frac{\partial R}{\partial \varphi_{ut}}(\Phi, \Theta)\big)_+}, \qquad \theta_{td} = \frac{\big(\hat{n}_{dt} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}(\Phi, \Theta)\big)_+}{\sum\limits_{s \in T}\big(\hat{n}_{ds} + \theta_{sd}\frac{\partial R}{\partial \theta_{sd}}(\Phi, \Theta)\big)_+}, \tag{5.2}$$

где $\hat{n}_{wt}$, $\hat{n}_{dt}$ определяются по прежним формулам (2.3)–(2.4).

Можно ли обосновать сходимость по аналогии с обычным ЕМ? ToDo[9]
Есть ли что-то похожее в литературе?

Знаменатель в этих формулах нужен только для нормировки. Поэтому используется также сокращённая запись через знак пропорциональности $\propto$:

$$\varphi_{wt} \propto \left(\hat{n}_{wt} + \varphi_{wt}\frac{\partial R(\Phi, \Theta)}{\partial \varphi_{wt}}\right)_+, \qquad \theta_{td} \propto \left(\hat{n}_{dt} + \theta_{td}\frac{\partial R(\Phi, \Theta)}{\partial \theta_{td}}\right)_+.$$

Добавление ещё одного регуляризатора приводит к добавлению соответствующей поправки в формуле М-шага. Это позволяет единообразно строить *многоцелевые тематические модели*, сочетающие в себе большое число дополнительных требований, а также *композитные тематические модели*, объединяющие в себе несколько более простых моделей.

Полувероятностный подход. Многие регуляризаторы допускают вероятностную интерпретацию. Если регуляризатор с точностью до константного сомножителя является логарифмом некоторого априорного распределения, то задача (5.1) эквивалентна максимизации апостериорной вероятности.

Мы будем придерживаться более гибкой концепции *полувероятностного подхода*. Наличие вероятностной интерпретации — желательное, но не обязательное свойство регуляризатора. Более ценной является возможность комбинирования различных регуляризаторов, независимо от того, какие они имеют обоснования.

Принцип регуляризации некорректно поставленных задач не нуждается в дополнительных вероятностных обоснованиях. Достаточно того, что левая часть функционала $L(\Phi,\Theta)$ является логарифмом правдоподобия, а правая часть $R(\Phi,\Theta)$ позволяет выбрать из множества решений наиболее подходящее.

Далее рассматриваются многочисленные примеры полезных регуляризаторов.

§5.1 Сглаживание и разреживание

Сглаживающий регуляризатор Дирихле. Сглаженные частотные оценки условных вероятностей (3.3)–(3.4), обычно получаемые через априорные распределения Дирихле и байесовский вывод, могут быть получены также через регуляризатор.

Зададим дискретные распределения на множестве терминов $\tilde{\beta} = (\tilde{\beta}_w)_{w\in W}$ и на множестве тем $\tilde{\alpha} = (\tilde{\alpha}_t)_{t\in T}$. Потребуем, чтобы распределения φ_t были похожи на $\tilde{\beta}$, распределения φ_d — на $\tilde{\alpha}$. Для этого будем минимизировать суммы KL-дивергенций:

$$\sum_{t\in T} \mathrm{KL}_w(\tilde{\beta}_w \| \varphi_{wt}) \to \min_{\Phi}; \qquad \sum_{d\in D} \mathrm{KL}_t(\tilde{\alpha}_t \| \theta_{td}) \to \min_{\Theta}.$$

Для одновременной минимизации обеих сумм KL-дивергенций введём сумму двух регуляризаторов с коэффициентами β_0 и α_0:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T}\sum_{w\in W} \tilde{\beta}_w \ln \varphi_{wt} + \alpha_0 \sum_{d\in D}\sum_{t\in T} \tilde{\alpha}_t \ln \theta_{td} \to \max.$$

Обозначим $\beta_w = \beta_0\tilde{\beta}_w$ и $\alpha_t = \alpha_0\tilde{\alpha}_t$. Применив общую формулу М-шага (5.2), получим сглаженные байесовские оценки (3.3)–(3.4) в модели LDA:

$$\varphi_{wt} \propto \hat{n}_{wt} + \beta_w, \qquad \theta_{td} \propto \hat{n}_{dt} + \alpha_t.$$

Будем называть данный регуляризатор *сглаживающим регуляризатором Дирихле*, хотя он не использует распределение Дирихле. Название подчёркивает его связь с моделью латентного размещения Дирихле LDA, на основе которой в настоящее время строится подавляющее большинство более сложных тематических моделей. Популярность модели LDA объясняется тем, что главным способом оценивания параметров Φ и Θ до сих пор был байесовский вывод с его чисто техническим требованием, чтобы априорное распределение было сопряжено с мультиномиальным, то есть было распределением Дирихле.

В теории регуляризации тематических моделей распределение Дирихле утрачивает не только лингвистическую, но и математическую целесообразность. Это лишь один из возможных регуляризаторов, не самый лучший и не настолько универсальный, как принято считать. В качестве базовой модели логичнее брать более простую модель PLSA, которая не имеет собственных регуляризаторов, и добавлять к ней регуляризаторы, адекватные конкретной задаче.

Разреживающий энтропийный регуляризатор. Недостатком сглаживающего регуляризатора является его явное противоречие с гипотезой разреженности. Для практических целей (классификации, категоризации, информационного поиска и т. д.) было бы полезно иметь тематическую модель с сильно разреженными матрицами Φ и Θ, в которых доля нулевых значений превышает 90%.

Чем сильнее разрежено распределение, тем ниже его энтропия. Максимальной энтропией обладает равномерное распределение. Поэтому будем максимизировать KL-дивергенцию между равномерным распределением и искомыми распределениями φ_t и θ_d. Назовём *энтропийным регуляризатором* сумму дивергенций по всем темам t и всем документам d с коэффициентами регуляризации β и α:

$$R(\Phi,\Theta) = -\beta \sum_{t\in T}\sum_{w\in W} \ln \varphi_{wt} - \alpha \sum_{d\in D}\sum_{t\in T} \ln \theta_{td} \to \max.$$

Применив общую формулу М-шага (5.2), получим:

$$\varphi_{wt} \propto \big(\hat{n}_{wt} - \beta\big)_+, \qquad \theta_{td} \propto \big(\hat{n}_{dt} - \alpha\big)_+.$$

Идея энтропийной регуляризации была предложена в динамической тематической модели PLSA для обработки видеопотоков [56]. В данной задаче документами являются видеозаписи, терминами — признаки на изображениях, темами — появление определённого объекта в течение определённого времени, например, проезд автомобиля. Сильно разреженное распределение было необходимо для описания моментов возникновения тем. Удивительно, что авторы не заметили возможность применения этой же техники для разреживания распределений φ_t и θ_d.

Заметим, что сглаживание и разреживание могут быть описаны одной и той же формулой, без ограничений на знаки параметров β и α.

Максимизация апостериорной вероятности — это байесовский подход, позволяющий совместить разреживание и сглаживание, хотя и с некоторыми неестественными ограничениями. Рассмотрим задачу максимизации совместного правдоподобия выборки (коллекции) D и модели Φ, Θ при фиксированных гиперпараметрах β, α:

$$p(D \mid \Phi,\Theta) p(\Phi;\beta) p(\Theta;\alpha) \;\to\; \max_{\Phi,\Theta}.$$

Полагая, что столбцы матриц Φ и Θ являются независимыми случайными векторами из распределений Дирихле, запишем задачу максимизации правдоподобия:

$$\prod_{d\in D}\prod_{w\in d} p(d,w)^{n_{dw}} \prod_{t\in T} \mathrm{Dir}(\varphi_t;\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d;\alpha) \;\to\; \max_{\Phi,\Theta}.$$

Прологарифмировав это произведение и отбросив слагаемые, не влияющие на положение точки максимума, получим задачу максимизации

$$\begin{aligned} L(\Phi,\Theta) = &\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + \\ &+ \sum_{t\in T}\sum_{w\in W} (\beta_w - 1)\,[\varphi_{wt} > 0] \ln \varphi_{wt} + \\ &+ \sum_{d\in D}\sum_{t\in T} (\alpha_t - 1)\,[\theta_{td} > 0] \ln \theta_{td} \;\to\; \max, \end{aligned} \tag{5.3}$$

при стандартных ограничениях неотрицательности $\varphi_{wt} \geqslant 0$, $\theta_{td} \geqslant 0$ и нормировки

$$\sum_{w \in W} \varphi_{wt} = 1, \quad \sum_{t \in T} \theta_{td} = 1.$$

Эта задача формально совпадает с задачей максимизации обычного логарифма правдоподобия (первое слагаемое) с аддитивным регуляризатором (второе и третье слагаемые). Поэтому к ней применима общая формула (5.2).

Условные сомножители $[\varphi_{wt} > 0]$ и $[\theta_{td} > 0]$ появились в (5.3) из следующих соображений. Распределения Дирихле $\mathrm{Dir}(\varphi_t; \beta)$ и $\mathrm{Dir}(\theta_d; \alpha)$ не определены при $\varphi_{wt} = 0$ и $\theta_{td} = 0$ соответственно. Однако исключать возможность их обнуления нельзя согласно гипотезе разреженности. Поэтому мы допускаем обнуление отдельных координат векторов φ_t и θ_d, исключая соответствующие размерности из распределений Дирихле. Уменьшение размерности влияет на нормировочные множители, которые образуют аддитивную поправку к логарифму правдоподобия. Решение задачи максимизации (5.3) не зависит от этой поправки, поэтому она сразу отбрасывается.

Для решения задачи (5.3) применим EM-алгоритм и воспользуемся тем, что уравнение $x = a \cdot [x > 0]$ имеет решение $x = a_+$. Получим формулы M-шага:

$$\varphi_{wt} \propto \big(\hat{n}_{wt} + \beta_w - 1\big)_+, \qquad \theta_{td} \propto \big(\hat{n}_{dt} + \alpha_t - 1\big)_+. \tag{5.4}$$

Формулы (5.4) охватывают два противоположных типа регуляризации — сглаживание и разреживание.

При $\alpha_t > 1$ ($\beta_w > 1$) полученные формулы эквивалентны байесовским сглаженными оценками (3.3)–(3.4) с точностью до чисто технического преобразования гиперпараметров (уменьшения на единицу).

Если $\alpha_t = 1$ ($\beta_w = 1$), то априорные распределения Дирихле симметричны и равномерны, регуляризация отключается и формулы (5.4) переходят в оценки максимума правдоподобия (2.3)–(2.4), применяемые в PLSA.

При $0 < \alpha_t < 1$ ($0 < \beta_w < 1$) малые значения вероятностей θ_{td} (φ_{wt}) могут обнуляться. Однако требование неотрицательности гиперпараметров β (α) сильно ограничивает возможности разреживания. В частности, при использовании сэмплирования Гиббса счётчики $\hat{n}_{wt}$ ($\hat{n}_{dt}$) могут принимать только целые значения, поэтому φ_{wt} (θ_{td}) смогут обратиться в нуль только при $\hat{n}_{wt} = 0$ ($\hat{n}_{dt} = 0$). Этого совершенно не достаточно для сильного разреживания матриц Φ и Θ, особенно в случаях коллекций большого размера.

Таким образом, апостериорная вероятность — недостаточно гибкий регуляризатор, поскольку она поощряет сглаживание в большей степени, чем разреживание.

§5.2 Частичное обучение

Детальный анализ и интерпретация построенной тематической модели может порождать дополнительные обучающие данные о том, что некоторый документ или термин релевантен или не релевантен определённой теме. Эксперты могут также просматривать ранжированные списки документов или терминов по темам и формировать обучающие данные о том, что некоторый документ или термин должен быть ранжирован выше или ниже какого-то другого документа или термина. Привязки документов и терминов к темам помогают фиксировать интерпретации тем,

избежать их перемешивания в ходе EM-итераций, повышают устойчивость тематической модели. Поскольку такие привязки возникают лишь для небольшого числа документов, терминов и тем, задача использования этих данных относится к области *частичного обучения* (semi-supervised learning).

Данные о релевантности документов темам. Пусть для некоторых документов d задано распределение θ^0_{td} на множестве тем. В частности, это может быть равномерное распределение на подмножестве тем $T_d \subset T$, к которым относится документ d, см. (2.7). Когда задаются такие требования, обычно неизвестно, относится ли документ ещё к каким-то темам, и как распределены вероятности θ_{td} между темами из T_d. Поэтому явное ограничение $\theta_{td} = \theta^0_{td}$ является избыточно жёстким. Введём регуляризатор, максимизирующий ковариацию между распределениями θ^0_{td} и θ_{td}. Для большей общности введём непрерывно дифференцируемую возрастающую функцию μ, и вместо θ_{td} запишем $\mu(\theta_{td})$:

$$R(\Phi, \Theta) = \tau \sum_{d \in D} m_d \sum_{t \in T} \theta^0_{td} \mu(\theta_{td}) \to \max,$$

где m_d — вес или степень важности документа d. Вес можно полагать равным длине документа, $m_d = n_d$, либо единице, если все документы одинаково важны.

Формула для θ_{td}, согласно (5.2), принимает вид

$$\theta_{td} \propto \hat{n}_{dt} + \tau m_d \theta^0_{td} \theta_{td} \mu'(\theta_{td}).$$

Смысл этой формулы в том, чтобы на каждой итерации EM-алгоритма ещё немного увеличивать оценки условной вероятности $\theta_{td} = p(t \mid d)$, если известно, что документ d относится к теме t. Похожая модификация EM-алгоритма для задач классификации текстов с частичным обучением предлагалась в [43].

Это тоже сглаживание, но, в отличие от LDA, оно производится только для тех θ_{td} и φ_{wt}, по которым имеются обучающие данные.

Выбор возрастающей функции μ в значительной степени произволен.

При $\mu(z) = z$ максимизируется взвешенная сумма ковариаций $\operatorname{cov}(\theta^0_d, \theta_d)$. Если распределение θ^0_{td} равномерно на T_d, то ковариация не накладывает никаких ограничений на распределение вероятностей θ_{td} между темами из T_d.

При $\mu(z) = \ln z$ максимизация R эквивалентна минимизации взвешенной суммы дивергенций $\mathrm{KL}(\theta^0_d \| \theta_d)$. В этом случае распределение θ_{td} стремится к θ^0_{td}. Логарифм — это единственная функция μ, для которой $z\mu'(z) \equiv 1$ и формула M-шага имеет наиболее простой вид с правой частью, не зависящей от переменных θ_{td}:

$$\theta_{td} \propto \hat{n}_{dt} + \tau m_d \theta^0_{td}.$$

Полученная формула интерпретируется как добавление в документ d «виртуального термина» с частотой $\tau m_d \theta^0_{td}$, который всегда относится к теме t.

Найти похожие подходы в литературе. ToDo[10]

Данные о релевантности терминов темам. Пусть для некоторых тем t задана функция распределения φ^0_{wt} на множестве терминов. В частности, это может быть

равномерное распределение на подмножестве терминов $W_t \subset W$ относящихся к теме t, см. (2.8). Введём регуляризатор

$$R(\Phi, \Theta) = \tau \sum_{t \in T} m_t \sum_{w \in W} \varphi^0_{wt} \mu(\varphi_{wt}) \to \max,$$

где m_t — вес темы, который можно полагать равным единице.

Формула для φ_{wt}, согласно (5.2), принимает вид

$$\varphi_{wt} \propto \hat{n}_{wt} + \tau m_t \, \varphi^0_{wt} \varphi_{wt} \mu'(\varphi_{wt}).$$

При $\mu(z) = \ln z$ максимизация R эквивалентна минимизации взвешенной суммы KL-дивергенций $\mathrm{KL}(\varphi^0_t \| \varphi_t)$, и формула M-шага приобретает наиболее простой вид:

$$\varphi_{wt} \propto \hat{n}_{wt} + \tau m_t \varphi^0_{wt}.$$

что интерпретируется как добавление в коллекцию D «виртуального документа», в котором термин w с частотой $\tau m_t \varphi^0_{wt}$ всегда относится к теме t.

Найти похожие подходы в литературе. ToDo[11]

Данные о переранжировании. Допустим, эксперты имеют возможность просмотреть список тем по любому документу d, ранжированный по убыванию частот $\hat{n}_{dt}$. Эксперт может перенести любую тему на то место в списке, которое он считает наиболее релевантным.

Эти данные, полученные от экспертов, легко учесть в EM-алгоритме. Чтобы тема t оказалась на k-м месте в списке тем документа d, достаточно сделать значение $\hat{n}_{dt}$ немного бóльшим k-го значения $\hat{n}^{(k)}_{dt}$ и скорректировать счётчик $\hat{n}_d$:

если тема t для документа d должна быть на k-м месте **то**
$\hat{n}'_{dt} := \frac{1}{2}\big(\hat{n}^{(k)}_{dt} + \hat{n}^{(k-1)}_{dt}\big)[k > 1] + \frac{1}{2}\big(3\hat{n}^{(k)}_{dt} - \hat{n}^{(k+1)}_{dt}\big)[k = 1]$;
$\hat{n}_d := \hat{n}_d - \hat{n}_{dt} + \hat{n}'_{dt}$;
$\hat{n}_{dt} := \hat{n}'_{dt}$;

Аналогичным образом возможно собрать и учесть данные о переранжировании терминов. Допустим, по теме t имеется список терминов, ранжированный по убыванию частот n_{wt}, и эксперт может проделать с ним аналогичную работу. Чтобы термин w оказался на k-м месте в списке терминов темы t, достаточно сделать значение $\hat{n}_{wt}$ немного бóльшим k-го значения $\hat{n}^{(k)}_{wt}$ и скорректировать счётчик $\hat{n}_t$:

если термин w для темы t должен быть на k-м месте **то**
$\hat{n}'_{wt} := \frac{1}{2}\big(\hat{n}^{(k)}_{wt} + \hat{n}^{(k-1)}_{wt}\big)[k > 1] + \frac{1}{2}\big(3\hat{n}^{(k)}_{wt} - \hat{n}^{(k+1)}_{wt}\big)[k = 1]$;
$\hat{n}_t := \hat{n}_t - \hat{n}_{wt} + \hat{n}'_{wt}$;
$\hat{n}_{wt} := \hat{n}'_{wt}$;

Таким образом, различные виды априорной информации о связях документов и терминов с темами формализуются либо с помощью регуляризаторов, либо непосредственно через модификацию значений счётчиков. В обоих случаях они приводят к модификации формул M-шага.

Найти похожие подходы в литературе. ToDo[12]

§5.3 Разреживание как L_0-регуляризация

Рассмотрим ещё один разреживающий регуляризатор, равный числу нулевых параметров в матрицах Φ и Θ:

$$R(\Phi,\Theta) = \tau_\varphi \sum_{w\in W}\sum_{t\in T}\big[\varphi_{wt}=0\big] + \tau_\theta \sum_{d\in D}\sum_{t\in T}\big[\theta_{td}=0\big] \to \max. \tag{5.5}$$

Данный критерий не является гладким, поэтому воспользоваться общими формулами (5.2) не удастся. Выбор максимального подмножества коэффициентов для обнуления — это задача комбинаторной оптимизации. Для её решения предлагается эвристический алгоритм постепенного принудительного разреживания.

Принудительное разреживание. Допустим, что EM-алгоритм сошёлся в точку локального максимума правдоподобия $L(\Phi,\Theta)$, и первые производные правдоподобия (1.6) по всем параметрам φ_{wt}, θ_{td} равны нулю. Зададимся вопросом: обнуление каких параметров меньше всего повлияет на значение правдоподобия? Применим ту же технику, которая используется в методе оптимального разреживания многослойных нейронных сетей OBD (Optimal Brain Damage) [29]. Разложив правдоподобие в ряд Тейлора в окрестности точки максимума, получим квадратичную форму по приращениям параметров $\Delta\varphi_{wt}$, $\Delta\theta_{td}$ (здесь выписаны не все частные производные, однако нетрудно показать, что остальные частные производные либо равны нулю, либо в сумме дают нулевой вклад в квадратичную форму):

$$L(\Phi{+}\Delta\Phi,\Theta{+}\Delta\Theta) = L(\Phi,\Theta) + \frac{1}{2}\sum_{w\in W}\sum_{t\in T}\sum_{s\in T}\Delta\varphi_{wt}\Delta\varphi_{ws}\frac{\partial^2 L(\Phi,\Theta)}{\partial\varphi_{wt}\partial\varphi_{ws}} + \\ + \frac{1}{2}\sum_{d\in D}\sum_{t\in T}\sum_{s\in T}\Delta\theta_{td}\Delta\theta_{sd}\frac{\partial^2 L(\Phi,\Theta)}{\partial\theta_{td}\partial\theta_{sd}} + o(\Delta\Phi,\Delta\Theta).$$

Обнулить параметр φ_{wt} означает положить $\varphi_{wt}+\Delta\varphi_{wt}=0$, откуда следует $\Delta\varphi_{wt}=-\varphi_{wt}$. Аналогично, $\Delta\theta_{td}=-\theta_{td}$. Возьмём вторые производные правдоподобия и перегруппируем слагаемые:

$$L(\Phi{+}\Delta\Phi,\Theta{+}\Delta\Theta) = L(\Phi,\Theta) - \frac{1}{2}\sum_{t\in T}\hat{n}_t\sum_{w\in W}\varphi_{wt} - \frac{1}{2}\sum_{d\in D}\hat{n}_d\sum_{t\in T}\theta_{td} + o(\Delta\Phi,\Delta\Theta).$$

Переделать. Расписать случаи, когда обнуляется: ToDo[13]
1) одна ячейка теты или фи,
2) одна строка фи, за исключением небольшого числа ячеек,
3) одна строка теты, за исключением небольшого числа ячеек.

Заметим, что в стандартном методе OBD обычно пренебрегают смешанными частными производными, чтобы упростить вид квадратичной формы. В данном случае квадратичная форма упрощается благодаря специальному виду функционала, и делать какие-либо приближения нет необходимости.

Из полученной формулы следует интуитивно очевидная стратегия разреживания: после каждого прохода коллекции в каждом распределении $\varphi_{wt}=\hat{n}_{wt}/\hat{n}_t$ и $\theta_{td}=\hat{n}_{dt}/\hat{n}_d$ обнуляются наименьшие значения вероятностей, для которых сумма

счётчиков $\hat{n}_{wt}$ и $\hat{n}_{dt}$, соответственно, не превышает некоторый порог. Варьирование этого порога эквивалентно варьированию коэффициента регуляризации τ.

Эвристические стратегии разреживания и результаты экспериментов подробно обсуждаются в разделе §13.4.

Сокращение тематики и словаря. Принудительное разреживание может приводить к обнулению строк целиком в матрицах Θ и Φ.

Обнуление t-й строки в матрице Θ означает, что тема t исключается из тематической модели. Это позволяет построить процедуру оптимизации числа тем, если задавать изначально избыточное число тем и постепенно в процессе итераций избавляться от лишних тем.

Проверить работоспособность этой идеи в эксперименте. ToDo[14]

Обнуление w-й строки в матрице Φ означает, что слово w не существенно для тематической модели. Сокращение словаря полезно с точки зрения вычислительной эффективности и повышения интерпретируемости модели за счёт отбрасывания незначимых слов. Однако эксперименты показали, что отбрасываются в основном наименее частотные слова (с наименьшими n_w), что практически эквивалентно грубой фильтрации словаря, вообще не требующей построения тематической модели.

Пока есть только предварительный эксперимент [L.Evans, 2013]. Посчитать процент совпадения множеств слов, отброшенных разреживанием и грубым частотным фильтром. Скорее всего, они почти совпадут, и это будет поводом для перехода к ковариационному регуляризатору. ToDo[15]

Далее рассматривается метод максимизации ковариации тем, который также приводит к разреживанию, но не имеет этого недостатка.

§5.4 Повышение различности тем

Тематическая модель тем полезнее, чем более различные темы она находит. Это предположение приводит к дополнительному требованию увеличивать различность тем. Можно по-разному формализовать понятие различности тем как дискретных распределений $\varphi_{wt} = p(t \mid w)$ или нормированных векторов $\varphi_w = (\varphi_{wt})_{t \in T}$. Остановимся на естественной мере различности — ковариации:

$$R(\Phi, \Theta) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \operatorname{cov}(\varphi_t, \varphi_s) \to \max, \qquad \operatorname{cov}(\varphi_t, \varphi_s) = \sum_{w \in W} \varphi_{wt} \varphi_{ws}.$$

Этот критерий не зависит от Θ, поэтому для θ_{td} формулы М-шага не меняются. Формула для φ_{wt}, согласно (5.2), принимает вид

$$\varphi_{wt} \propto \Big(\hat{n}_{wt} - \tau \varphi_{wt} \sum_{s \in T \setminus t} \varphi_{ws} \Big)_+ .$$

Смысл этой формулы в том, что условные вероятности $\varphi_{wt} = p(w \mid t)$ постепенно уменьшаются для тех слов w, которые имеют бóльшие значения вероятности φ_{ws} в других темах. В процессе итераций ЕМ-алгоритма для каждого слова вероятности наиболее значимых тем приобретают всё большие значения, а вероятности менее

значимых тем уменьшаются и могут обращаться в нуль. При достаточно больших τ требование некоррелированности приводит к разреживанию матрицы Ф.

Регуляризатор, описанный в предыдущем параграфе, также приводит к разреживанию, но минимизация ковариаций не так агрессивно обнуляет строки матрицы Ф, соответствующие редким словам. Кроме того, данный регуляризатор обладает дополнительным полезным свойством выделять стоп-слова в отдельные темы [52].

Реализовать и сравнить в экспериментах с принудительным разреживанием. Проверить, что действительно «не так агрессивно обнуляет строки» — пока это только гипотеза. Кажется, что от принудительного разреживания откажемся в пользу антикоррелятора. ToDo[16]

Вариант реализации: как только изменяется φ_{wt}, надо пробежать по всем φ_{ws}, которые от него зависят, и прибавить разность. Эксперимент: то ли это надо сделать обязательно, то ли это просто ускорит сходимость, то ли это без разницы. ToDo[17]

§5.5 Повышение когерентности тем

Тема называется *когерентной*, если термины, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции [40, 41, 42]. Для повышения когерентности темы необходимо заранее вычислить оценки C_{uv} совместной встречаемости пар слов (u, v). Они могут оцениваться как по сторонней коллекции, например, по Википедии [39], так по коллекции, для которой строится модель [35]. Будем полагать, что $C_{uv} \geqslant 0$, причём если слова u и v совместно не встречаются, то $C_{uv} = 0$. Для экономии времени и памяти оценки C_{uv} можно вычислять только для самых частых слов или сохранять только те значения C_{uv}, которые превышают некоторый порог. Обозначим через Q множество пар слов (u, v), для которых имеется оценка C_{uv}. Будем называть такие пары слов *когерентными*.

Оценки совместной встречаемости принято вычислять на основе документной частоты терминов. Введём следующие обозначения:

N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом, как правило, в окне ширины $h = 10$ слов [39];

N_u — число документов, в которых термин u встречается хотя бы один раз;

$N = |D|$ — число документов в коллекции;

$\mathrm{LCP}_{uv} = \ln(N_{uv}/N_u)$ — логарифм условной вероятности (log conditional probability), мера связанности слова v со словом u;

$\mathrm{PMI}_{uv} = \ln(N_{uv}N/N_uN_v)$ — поточечная взаимная информация (pointwise mutual information), мера неслучайности совместного употребления слов u и v; если появление слов u и v — статистически независимые события, то $\mathrm{PMI}_{uv} \approx 0$;

$\mathrm{IDF}_u = \ln(N/N_u)$ — инвертированная документная частота (inverted document frequency), в информационном поиске применяется для понижения веса слишком частных слов, в том числе стоп-слов.

Наиболее адекватной мерой когерентности отдельных тем и тематической модели в целом принято считать среднее значение LCP_{uv}, по всем парам из 10 наиболее частых слов в каждой теме [35, 28]. Среднее значение PMI_{uv}, использовавшаяся ранее, признано менее удачной мерой [35]. Определение когерентности рассматривается более подробно в §12.6, стр. 64. Вопросы о том, какую величину C_{uv} взять для опти-

мизации когерентности, и как построить критерий регуляризации, пока не выяснены столь однозначно и требуют дополнительных исследований.

Регуляризатор ковариационного типа. В работе [39] предлагается использовать оценку совместной встречаемости $C_{uv} = N_{uv}\big[\mathrm{PMI}_{uv} > 0\big]$ и регуляризатор

$$R(\Phi,\Theta) = \tau \sum_{t\in T} \ln \sum_{(u,v)\in Q} C_{uv}\varphi_{ut}\varphi_{vt} \to \max.$$

Формула для φ_{wt}, согласно (5.2), имеет вид

$$\varphi_{wt} \propto \hat{n}_{wt} + \tau\varphi_{wt} \frac{\sum\limits_{(u,w)\in Q} C_{uw}\varphi_{ut} + \sum\limits_{(w,v)\in Q} C_{wv}\varphi_{vt}}{\sum\limits_{(u,v)\in Q} C_{uv}\varphi_{ut}\varphi_{vt}}.$$

Таким образом, условные вероятности φ_{wt} увеличиваются для тех слов w, которые имеют когерентные им слова u с высокими вероятностями φ_{ut}.

Упрощённые варианты регуляризаторов ковариационного типа. Можно использовать и другие меры близости векторов $(C_{uv})_{(u,v)}$ и $(\varphi_{ut}\varphi_{vt})_{(u,v)}$. При этом вероятности перераспределяются между когерентными словами несколько иначе.

Убрав логарифм из $R(\Phi,\Theta)$, получим более простую формулу для φ_{wt}:

$$\varphi_{wt} \propto \hat{n}_{wt} + \tau\varphi_{wt}\Big(\sum_{(u,w)\in Q} C_{uw}\varphi_{ut} + \sum_{(w,v)\in Q} C_{wv}\varphi_{vt}\Big).$$

Если в качестве регуляризатора взять сумму KL-дивергенций для пар слов,

$$R(\Phi,\Theta) = \tau \sum_{t\in T} \sum_{(u,v)\in Q} C_{uv} \ln(\varphi_{ut}\varphi_{vt}) \to \max,$$

то формула для φ_{wt} упростится ещё больше:

$$\varphi_{wt} \propto \hat{n}_{wt} + \tau\Big(\sum_{(u,w)\in Q} C_{uw} + \sum_{(w,v)\in Q} C_{wv}\Big).$$

Регуляризатор, соответствующий обобщённой урновой схеме Пойя. В работе [35] предлагается нормированная оценка

$$C_{uv} = \frac{\tilde{C}_{uv}}{\sum\limits_{w\in W} \tilde{C}_{wv}}; \qquad \tilde{C}_{uv} = \begin{cases} N_u \mathrm{IDF}_u, & u = v; \\ N_{uv}\mathrm{IDF}_u\big[\mathrm{IDF}_u \geqslant 3\big], & u \neq v. \end{cases}$$

Про регуляризацию в явном виде не говорится. Обоснование проводится для алгоритма сэмплирования Гиббса с помощью обобщённой урновой схемы Пойя, что, на наш взгляд, только затрудняет понимание. В результате получается несколько необычная формула для φ_{wt}:

$$\varphi_{wt} = \frac{\sum\limits_{u\in W} C_{wu}\hat{n}_{ut} + \beta}{\hat{n}_t + |W|\beta}.$$

Необычность в том, что коэффициент при $\hat{n}_{wt}$ равен не 1, а величине C_{ww}, которая после нормировки имеет неясную интерпретацию.

Вычисления по этой формуле в алгоритме сэмплирования Гиббса организуются следующим образом. Всякий раз, когда счётчик $\hat{n}_{ut}$ увеличивается на $\delta = 1$, в цикле по всем словам w счётчик $\hat{n}_{wt}$ увеличивается на $C_{wu}\delta$. Чтобы это не занимало много времени, матрица C_{uv} должна быть сильно разреженной. Для этого вводится эвристика $\mathrm{IDF}_u \geqslant 3$, возможно также вводить и другие эвристики, например, $\mathrm{PMI}_{uv} > 0$.

Модифицируем формулу φ_{wt} так, чтобы она приобрела стандартный вид, оставаясь в то же время реализацией обобщённой урновой схемы Пойя. Введём управляющий параметр τ и уберём параметр β, который является очевидным следствием вездесущей регуляризации Дирихле:

$$\varphi_{wt} \propto \hat{n}_{wt} + \tau \sum_{u \in W \setminus w} C_{uw} \hat{n}_{ut}.$$

Нетрудно показать, что эта формула соответствует регуляризатору

$$R(\Phi, \Theta) = \tau \sum_{t \in T} \sum_{(u,v) \in Q} C_{uv} \hat{n}_{ut} \ln \varphi_{vt} \to \max,$$

который также несколько необычен тем, что коэффициенты $\hat{n}_{ut}$ изменяются в ходе итераций. Тем не менее, этот регуляризатор имеет совершенно ясную интерпретацию. Возьмём $C_{uv} = N_{uv}/N_v$ — оценку условной вероятности слова v при условии u. Тогда данный регуляризатор минимизирует сумму дивергенций Кульбака-Лейблера

$$\sum_{t \in T} \sum_{v \in W} \mathrm{KL}(\hat{n}_{v|t} \| \varphi_{vt}) \to \min$$

между распределением $\varphi_{vt} = p(v \,|\, t)$ и эмпирической оценкой $\hat{n}_{v|t}$ частоты слова v в теме t, вычисленной по оценкам частот всех когерентных слов:

$$\hat{n}_{v|t} = \hat{p}(v \,|\, t)\hat{n}_t = \sum_{u\colon (u,v) \in Q} \hat{p}(v \,|\, u)\hat{p}(u \,|\, t)\hat{n}_t = \sum_{u\colon (u,v) \in Q} C_{uv} \hat{n}_{ut}.$$

Таким образом, алгоритм обобщённой урновой схемы Пойя [35] также можно считать разновидностью регуляризации.

§5.6 Учёт связей между документами

Ещё до построения тематической модели может быть известно, что какие-то документы имеют схожую тематику. Это могут быть документы, относящиеся к одной рубрике тематического рубрикатора, или документы, сгруппированные пользователем электронной библиотеки по тематике, или документы, где-то упоминавшиеся вместе. В частности, это могут быть документы, ссылающиеся друг на друга [16]. Научные статьи ссылаются на другие статьи через списки литературы. Веб-страницы или статьи Википедии используют для этого гиперссылки. Одна и та же ссылка может многократно упоминаться в тексте, и эту важную частотную информацию также необходимо учитывать.

Тематические наборы документов. Обозначим через $D_y \subset D$ *тематические наборы* документов, предположительно имеющих схожую тематику, где индекс y пробегает заданное конечное множество Y. Тематических наборов может быть много, и они могут противоречить друг другу. Поэтому требование, чтобы документы d, d' из одного тематического набора D_y имели похожие распределения θ_d, $\theta_{d'}$, должно учитываться не жёстко. Один из возможных вариантов регуляризации — максимизация суммы ковариаций:

$$R(\Phi,\Theta) = \frac{\tau}{2}\sum_{y\in Y}\sum_{d,d'\in D_y}\operatorname{cov}(\theta_d,\theta_{d'}) \to \max, \qquad \operatorname{cov}(\theta_d,\theta_{d'}) = \sum_{t\in T}\theta_{td}\theta_{td'}.$$

Этот критерий не зависит от Φ, поэтому для φ_{wt} формулы М-шага не меняются. Формула для θ_{td}, согласно (5.2), принимает вид

$$\theta_{td} \propto \hat{n}_{dt} + \tau\theta_{td}\sum_{y\in Y}\sum_{d'\in D_y\setminus d}\theta_{td'}.$$

Смысл этой формулы в том, что условные распределения $\theta_{td} = p(t\,|\,d)$ документов d, d', принадлежащих одной тематической подборке, в ходе итераций постепенно приближаются друг к другу.

Найти похожие работы в литературе. ToDo[18]

Ссылки и гиперссылки. Ковариационный регуляризатор легко обобщается на случай, когда на множестве документов задан направленный граф связей $G = \langle D, E\rangle$, где E — множество рёбер графа. Ребро графа $(d,c) \in E$ означает, что документ d ссылается на документ c или цитирует его [16]. Будем считать, что в таком случае тематика документа c близка к тематике документа d. При этом сходство тематики пропорционально n_{dc} — числу ссылок на документ c в документе d. Формализуем эти предположения с помощью регуляризатора

$$R(\Phi,\Theta) = \tau\sum_{(d,c)\in E} n_{dc}\operatorname{cov}(\theta_d,\theta_c) \to \max.$$

В [16] предложена похожая модель LDA-JS, в которой вместо максимизации ковариации минимизируется дивергенция Йенсена-Шеннона между θ_d и θ_c.

Обозначим через $N_d = \{c \in D\colon (d,c) \in E\}$ множество вершин, смежных с d.

Формула М-шага для θ_{td}, согласно (5.2), принимает вид

$$\theta_{td} \propto \hat{n}_{dt} + \tau\theta_{td}\sum_{c\in N_d} n_{dc}\theta_{tc}.$$

Таким образом, условные распределения $\theta_{td} = p(t\,|\,d)$ в ходе итераций приближаются к распределениям θ_{tc} документов, связанных с d.

§5.7 Траектория регуляризации

При использовании линейной комбинации регуляризаторов R_i возникает проблема выбора вектора коэффициентов $\tau = (\tau_i)_{i=1}^n$. Аналогичная проблема эффективно решается в эластичных сетях (elastic net) при совмещении L_1- и L_2-регуляризации

для задач регрессии и классификации [70]. Там решение вычисляется одновременно для множества векторов (regularization path), за время, сравнимое со временем решения одной задачи при фиксированном векторе коэффициентов [19]. Общие методы многопараметрической регуляризации предложены в [23].

В задачах тематического моделирования регуляризаторы могут влиять друг на друга. Разреживание влияет на большинство регуляризаторов. Согласно экспериментам, некоторые регуляризаторы могут ухудшать сходимость, если включать их слишком рано или слишком резко. Поэтому предлагается увеличивать коэффициенты регуляризации постепенно и в определённой последовательности, выстраивая в ходе итераций EM-алгоритма траекторию в пространстве коэффициентов регуляризации (regularization trajectory).

Есть ли в литературе такой подход? Внимательнее посмотреть [23]. ToDo[19]

6 Тематические модели классификации

Многие текстовые коллекции содержат для каждого документа d дополнительную информацию, называемую также *метаинформацией*, например:

— время y_d создания или публикации документа d;
— список авторов A_d документа d;
— список документов D'_d, на которые ссылается d;
— список авторов A'_d, на которых ссылается d;
— список документов D''_d, которые ссылаются на d;
— список авторов A''_d, которые ссылаются на d;
— список категорий C_d рубрикатора, к которым относится d;
— список сущностей E_d, упоминаемых в документе d;
— список ярлыков L_d, присвоенных пользователями документу d.
— список пользователей U_d документа d.

Для этих и других подобных типов информации задача формализуется единообразно. Каждому документу соответствует набор элементов из конечного множества C, называемых *классами* (class), *метками классов* или просто *метками* (label). Предполагается, что если документы имеют одинаковые метки, то они имеют также схожую тематику. Поэтому учёт меток может улучшать интерпретируемость тем, даже если между классами и темами нет однозначного соответствия.

Задача заключается в том, чтобы выявить связи между классами и темами, улучшить качество тематической модели и построить алгоритм классификации новых документов, для которых метки ещё не проставлены.

Сложность задачи в том, что стандартные алгоритмы классификации показывают неудовлетворительные результаты на больших текстовых коллекциях с большим числом несбалансированных, пересекающихся, взаимозависимых классов [48]. *Несбалансированность* означает, что классы могут содержать как очень малое, так и очень большое число документов. В случае *пересекающихся* классов документ может относиться как к одному классу, так и к очень большому числу классов. *Взаимозависимые* классы имеют схожие множества характерных терминов, и при классификации документа вступают в конкуренцию.

Тематические модели лучше справляются с такими задачами, поскольку они учитывают все классы одновременно [48]. Кроме того, в процессе построения модели они приписывают метки классов каждому термину w в каждом документе d, что даёт полезную дополнительную информацию о структуре документов.

Далее мы рассмотрим несколько тематических моделей для классификации документов, в порядке усложнения и отказа от избыточно сильных ограничений.

§6.1 Моделирование классов темами

Начнём с простой тематической модели классификации, основанной на двух довольно сильных ограничениях.

1. Темы отождествляются с классами, $C \equiv T$. Это сильное предположение о классах, так как требование условной независимости $p(w \mid d, c) = p(w \mid c)$, постулируемое для латентных тем, может не выполняться для наблюдаемых классов.

2. Для каждого документа d точно известно множество всех классов $C_d \subset C$, к которым он относится. Это предположение подходит лишь для некоторых типов задач. Для времени и авторов — подходит; для ссылок, категорий, пользователей и большинства других типов — не подходит.

При сделанных предположениях можно использовать стандартную тематическую модель (1.2), в которой фиксирована *структура разреженности* матрицы Θ: $\theta_{cd} = p(c \mid d) = 0$ для всех $c \notin C_d$,

$$p(w \mid d) = \sum_{c \in C_d} p(w \mid c) p(c \mid d) = \sum_{c \in C_d} \varphi_{wc} \theta_{cd}.$$

Формулы E-шага и M-шага также модифицируются не сильно: $p(c \mid d, w)$, $\hat{n}_{dc}$, θ_{cd} вычисляются не для всех $c \in C$, а только для классов документа $c \in C_d$.

Для классификации новых документов d применяется стандартная модель PLSA/LDA без ограничений на θ_{cd}. Документ относится к тем классам c, для которых условные вероятности θ_{cd} максимальны.

Данная модель известна в литературе как Flat-LDA [48] и Labeled-LDA [45]. Выразительные возможности данной модели существенно беднее, чем у PLSA/LDA, так как значительная доля элементов матрицы Θ фиксированы и равны нулю.

§6.2 Моделирование классов распределениями тем

Далее будем полагать, что классы $c \in C$ описываются не отдельными темами, а неизвестными условными распределениями $p(t \mid c)$ на множестве тем T.

Чтобы говорить о вероятностях классов, расширим вероятностное пространство до множества $D \times W \times T \times C$. Будем считать, что с каждым словом w в каждом документе d связана не только тема $t \in T$, но и класс $c \in C$.

Рассмотрим стандартную тематическую модель (1.2), в которой распределение вероятности тем документов $p(t \mid d)$ описывается смесью распределений тем классов $p(t \mid c)$ и классов документов $\pi_{cd} = p(c \mid d)$, где новой неизвестной является *матрица*

классификаций документов $\Pi = (\pi_{cd})_{C\times D}$:

$$p(t\,|\,d) = \sum_{c\in C} p(t\,|\,c)p(c\,|\,d) = \sum_{c\in C} \theta_{tc}\pi_{cd};$$
$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t) \sum_{c\in C} p(t\,|\,c)p(c\,|\,d) = \sum_{t\in T} \varphi_{wt} \sum_{c\in C} \theta_{tc}\pi_{cd}. \tag{6.1}$$

Для этой модели постулируются две гипотезы условной независимости:

$p(w\,|\,t,c,d) = p(w\,|\,t)$ — распределение слов полностью определяется тематикой документа и не зависит от самого документа и его классов;

$p(t\,|\,c,d) = p(t\,|\,c)$ — тематика документа d зависит не от самого документа, а только от того, каким классам он принадлежит;

$p(c\,|\,t,d) = p(c\,|\,t)$ — условие, эквивалентное предыдущему — классификация документа d зависит не от самого документа, а только от его тематики.

EM-алгоритм. Максимизация правдоподобия (1.6) для данной тематической модели $p(w\,|\,d)$ приводит к следующим формулам E-шага и M-шага:

$$H_{dwtc} = p(t,c\,|\,d,w) = \frac{\varphi_{wt}\theta_{tc}\pi_{cd}}{p(w\,|\,d)};$$
$$\varphi_{wt} \propto \hat{n}_{wt} = \sum_{d,c} n_{dw}H_{dwtc};$$
$$\theta_{tc} \propto \hat{n}_{tc} = \sum_{d,w} n_{dw}H_{dwtc};$$
$$\pi_{cd} \propto \hat{n}_{cd} = \sum_{w,t} n_{dw}H_{dwtc}.$$

Информация о классификации документов вводится в модель через матрицу Π, которая может фиксироваться либо вычисляться в зависимости от задачи.

Рассмотрим три основных случая.

Фиксированная матрица классификаций. Если классификации C_d для каждого документа d заданы точно и все классы равнозначны, то в качестве π_{cd} можно взять равномерные распределения на C_d:

$$\hat{p}(c\,|\,d) = \frac{1}{|C_d|}\big[c \in C_d\big]. \tag{6.2}$$

Этот случай в точности соответствует автор-тематической модели Author-Topic Model (ATM) [47], в которой классами являются авторы документов, и каждый документ d связан с множеством его авторов C_d. Равномерное распределение $p(c\,|\,d)$ является формализацией предположения о равном вкладе всех авторов $c \in C_d$ в создание документа d.

Для некоторых типов классов характерны частотные данные m_{dc} — сколько раз документ d был отнесён к классу c. Таким свойством, в частности, обладают:

— ссылки на авторов или документы в документе d;

— авторы или документы, ссылающиеся или цитирующие документ d;

— сущности, упоминаемые в документе d.

В таких случаях матрицу Π естественно заполнить частотными оценками:

$$\pi_{cd} = m_{dc}/m_d; \qquad m_d = \sum_{c \in C} m_{dc}.$$

В EM-алгоритме элементы фиксированной матрицы классификаций Π не вычисляются и не хранятся. Величины H_{dwtc}, $\hat{n}_{cd}$, вычисляются только для классов документа $c \in C_d$; суммирования также производятся только по $c \in C_d$.

Модель Бернулли для бинарных данных m_{dc}. ToDo[20]

Матрица классификаций с фиксированными нулями. Если множества классов C_d для всех документов d заданы точно, но вероятности классов $\pi_{cd} = p(c \,|\, d)$ неизвестны, то в матрице Π фиксируется *структура разреженности*, то есть только нулевые значения $\pi_{cd} = 0$ для всех (c, d) таких, что $c \notin C_d$.

В EM-алгоритме величины H_{dwtc}, $\hat{n}_{cd}$, π_{cd} вычисляются только для классов документа $c \in C_d$; суммирования также производятся только по $c \in C_d$.

Регуляризация матрицы классификаций. Если множества классов C_d заданы неточно или частично, то элементы матрицы Π не фиксируются. Вместо этого вводится регуляризатор, заставляющий распределения $\pi_{cd} = p(c \,|\, d)$ быть ближе к заданным распределениям $\hat{p}(c \,|\, d) = m_{dc}/m_d$:

$$R(\Pi) = \tau \sum_{d \in D} m_{dc} \sum_{c \in C_d} \ln \pi_{cd} \to \max.$$

В EM-алгоритме величины H_{dwtc}, $\hat{n}_{cd}$, π_{cd} вычисляются для всех $c \in C$, суммирования также производятся по всем $c \in C$. Регуляризатор приводит к модификации формулы M-шага только для переменных π_{cd}, по аналогии с общей формулой (5.2),

$$\pi_{cd} \propto \hat{n}_{cd} + \tau m_{dc}.$$

Регуляризация хорошо подходит для задач классификации с *частичным обучением*, в которых не только про вероятности π_{cd} ничего нельзя сказать заранее, но даже нет уверенности, что множества C_d включают в себя все классы, которым документ d действительно принадлежит. Задача частичного обучения в том и состоит, чтобы определить, каким ещё классам принадлежит каждый документ. Типичным примером являются задачи категоризации текстов или определения пользователей, которым можно порекомендовать данный документ.

§6.3 Частотный регуляризатор

В задачах классификации с большим числом классов или с несбалансированными классами хорошо зарекомендовала себя *частотная регуляризация* (label regularization) [32].

Потребуем, чтобы оценка безусловного распределения классов по тематической модели $p(c) = \frac{1}{|D|} \sum_{d \in D} \pi_{cd}$ была близка к наблюдаемым частотам классов $\hat{p}(c) = \frac{1}{|D|}|D_c|$,

где $D_c = \{d \in D\colon c \in C_d\}$ — множество документов, относящихся к классу c. Выразим данное требование с помощью регуляризатора:

$$R(\Pi) = \tau \sum_{c \in C} |D_c| \ln \sum_{d \in D} \pi_{cd} \to \max.$$

Формула М-шага для π_{cd}, по аналогии с (5.2), принимает вид

$$\pi_{cd} \propto \hat{n}_{cd} + \tau \frac{|D_c| \pi_{cd}}{\sum\limits_{d' \in D} \pi_{cd'}}.$$

Частотная регуляризация использовалась в тематической модели Prior-LDA, которая была предложена в [48] как улучшение модели Flat-LDA.

§6.4 Тематическая модель классификации

Тематическая модель (6.1) описывает распределение слов в документах $p(w \,|\, d)$ через распределения $p(t \,|\, c)$ и $p(c \,|\, d)$, связанные с классами. Если в задаче имеется сразу несколько типов классов, например, время, авторы, ссылки, категории, то не совсем ясно, какую из классификаций предпочесть как основную, и уже совсем не ясно, как учесть классификации всех остальных типов.

Решение данной проблемы заключается в том, чтобы моделировать распределение классов документов $p(c \,|\, d)$ через распределение тем документов $\theta_{td} = p(t \,|\, d)$ по аналогии с основной тематической моделью $p(w \,|\, d)$:

$$p(c \,|\, d) = \sum_{t \in T} p(c \,|\, t)\, p(t \,|\, d) = \sum_{t \in T} \psi_{ct} \theta_{td}, \tag{6.3}$$

где новой неизвестной является *матрица классов тем* $\Psi = (\psi_{ct})_{C \times T}$. Для классов постулируется гипотеза условной независимости $p(c \,|\, t, d) = p(c \,|\, t)$, означающая, что для классификации документа d достаточно знать только его тематику.

Будем полагать, что обучающая информация о принадлежности документов классам задаётся числами $m_{dc} \geqslant 0$, интерпретации которых могут быть различными в зависимости от задачи, например:

— частота, сколько раз документ d отнесён к классу c;

— бинарный индикатор $m_{dc} = [c \in C_d]$;

— оценка числа слов в документе, относящихся к классу c: $m_{dc} = n_d [c \in C_d] / |C_d|$.

Для построения регуляризатора будем минимизировать KL-дивергенцию между моделью классификации $p(c \,|\, d)$ и эмпирическим распределением $\hat{p}(c \,|\, d) \propto m_{dc}$:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct} \theta_{td} \to \max, \tag{6.4}$$

где коэффициент регуляризации τ необходим для «приведения к одному масштабу» частот слов n_{dw} и частот классов m_{dc}.

Тематическая модель с таким регуляризатором в сочетании с частотным регуляризатором и регуляризатором Дирихле известна как Dependency LDA [48]. В оригинальной работе приводится намного более громоздкое обоснование этой модели в рамках байесовского подхода, графических моделей и сэмплирования Гиббса.

EM-алгоритм. Задача решается по-прежнему с помощью EM-алгоритма.

На E-шаге дополнительно оценивается условная вероятность $H'_{dct} = p(t \mid d, c)$:

$$H_{dwt} = \frac{\varphi_{wt}\theta_{td}}{\sum\limits_{s\in T}\varphi_{ws}\theta_{sd}}; \qquad H'_{dct} = \frac{\psi_{ct}\theta_{td}}{\sum\limits_{s\in T}\psi_{cs}\theta_{sd}}.$$

На M-шаге оценки φ_{wt} вычисляются по прежним формулам; оценки ψ_{ct}, как и следовало ожидать, аналогичны φ_{wt} с точностью до замены терминов w на классы c и H_{dwt} на H'_{dct}; оценки θ_{td} агрегируют счётчики терминов и классов в документах:

$$\varphi_{wt} \propto \hat{n}_{wt}; \qquad \hat{n}_{wt} = \sum_{d\in D} n_{dw} H_{dwt};$$

$$\psi_{ct} \propto \hat{m}_{ct}; \qquad \hat{m}_{ct} = \sum_{d\in D} m_{dc} H'_{dct};$$

$$\theta_{td} \propto \hat{n}_{dt} + \tau\hat{m}_{dt}; \qquad \hat{n}_{dt} = \sum_{w\in W} n_{dw} H_{dwt}; \qquad \hat{m}_{dt} = \sum_{c\in C} m_{dc} H'_{dct}.$$

Вероятностная интерпретация. В нашей вероятностной модели с каждым словом в документе (d, w) связана как тема t, так и класс c. Примем гипотезу условной независимости терминов и классов в документах: $p(w, c \mid d) = p(w \mid d)\, p(c \mid d)$. Благодаря этому предположению логарифм правдоподобия $\ln \prod_{d\in D}\prod_{w\in d} p(d, w, c)^{n_{dw}}$ распадается на два похожих слагаемых — левое $L(\Phi, \Theta)$ совпадает со стандартным функционалом (1.6), правое $R(\Psi, \Theta)$ соответствует регуляризатору с коэффициентом $\tau = 1$, в котором m_{dc} есть число слов в документе, относящихся к классу c:

$$L(\Phi, \Theta) + R(\Psi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_t \varphi_{wt}\theta_{td} + \tau \sum_{d,c} m_{dc} \ln \sum_t \psi_{ct}\theta_{td} \to \max.$$

Если исходная обучающая информация m_{dc} имеет частотную интерпретацию, то регуляризатор является не только способом повышения устойчивости решения, но и непосредственным следствием принципа максимума правдоподобия.

Разреживание. Если имеется гипотеза, что каждый класс связан с небольшим числом тем, то для распределений ψ_{ct} вводится разреживающий регуляризатор:

$$R'(\Psi) = \tau' \sum_{c\in C}\sum_{t\in T} \big[\psi_{ct} = 0\big] \to \max,$$

для оптимизации которого можно использовать принудительное разреживание (см. §5.3) с одним ограничением: в матрице Ψ не должно быть нулевых строк, так как каждый класс содержит хотя бы один документ, следовательно, хотя бы одну тему.

Возможность оценить распределения $\psi_{ct} = p(c \mid t)$ и связать каждую тему с небольшим числом классов во многих приложениях является преимуществом данного типа моделей, так как позволяет лучше интерпретировать темы.

Тематическая модель цитирования документов LDA-post, предложенная в [16], в точности соответствует описанной модели, если $C = D$ и классами C_d являются документы, процитированные в документе d.

Число ненулевых элементов в строке разреженной матрицы Ψ, соответствующей документу $c \in D$, интерпретируется как число тем, на которые документ c оказывает существенное влияние. Это важный наукометрический показатель, если, конечно, он вычисляется по достаточно полной коллекции научных публикаций.

Тематическая модель цитирования авторов Author Link Topic model (ALT) [24] также аналогична описанной модели, но строится на основе автор-тематической модели (6.1). Классами в (6.1) являются авторы документов d коллекции D, а классами в регуляризаторе — авторы документов, на которые ссылаются d.

Многофункциональные тематические модели, учитывающие много типов классификаций, строятся путём добавления регуляризаторов, по одному на каждый тип. Каждому множеству классов C^j, $j = 1, \ldots, J$, ставится в соответствие модель классификации вида (6.3) и регуляризатор вида (6.4), строится ещё одна матрица классов тем Ψ^j, на М-шаге добавляется вычисление её элементов и вводится ещё одно слагаемое в формулу θ_{td}, на Е-шаге оцениваются условные вероятности $p(t \mid d, c^j)$, $c^j \in C^j$. Таким образом, добавление ещё одного типа классификаций приводит к линейному по числу классов $|C^j|$ увеличению затрат времени и памяти.

§6.5 Тематическая модель категоризации

Категоризация текстовых документов — это специальный случай классификации. Категории создаются с целью структуризации знаний. Каждая категория объединяет схожие по смыслу документы. Категории могут разделяться на подкатегории, образуя иерархическую структуру [3]. Наиболее известными примерами являются международный Универсальный десятичный классификатор УДК и российский Библиотечно-библиографический классификатор ББК. Это иерархические классификаторы, содержащие десятки тысяч категорий, создававшиеся сотнями экспертов по всем областям знания на протяжения многих десятилетий. Большая часть издаваемых книг расписаны по этим или другим подобным классификаторам. При создании новых категорий к ним приписывают ключевые термины. Имеются словари терминов по категориям, называемые *алфавитно-предметными указателями*.

Тематическая модель отдельных категорий. Понятия категории и темы во многом схожи, но не совпадают. Темы и категории имеют схожее назначение — группировать близкие по смыслу документы. Категории создаются экспертами на основе слабо формализованных критериев. Темы определяются формально как дискретные распределения на множестве терминов, удовлетворяющие гипотезе условной независимости на документах коллекции. Нет никаких гарантий, что категории являются темами. В лучшем случае какие-то из категорий могут быть темами. Более разумной представляется гипотеза, что распределение терминов $p(w \mid c)$ в каждой категории c является смесью распределений $p(w \mid t)$ небольшого числа тем t.

Будем строить тематическую модель категоризации на основе модели (6.3), считая, что классы C — это категории, и $|C| < |T|$.

Обобщим регуляризатор (6.4), заменив логарифм на непрерывно дифференцируемую возрастающую функцию μ:

$$R(\Psi,\Theta) = \tau \sum_{d\in D}\sum_{c\in C} m_{dc}\mu\left(\sum_{t\in T}\psi_{ct}\theta_{td}\right) \to \max.$$

Оказывается, что при определённом выборе функции μ у каждой темы возникает единственная *родительская категория* $c(t)$, следовательно, каждая категория разбивается на непересекающиеся темы. Эта особенность модели позволяет получать ответы на ряд практически важных вопросов:

— насколько однородны категории?

— какие категории пора разбивать на подкатегории?

— сколько подкатегорий можно было бы выделить в каждой категории?

Ответы на эти вопросы необходимы экспертам для улучшения качества категоризации, особенно в условиях постоянно растущей коллекции документов.

Ковариационный регуляризатор. При $\mu(z) = z$ и $m_{dc} = n_d\hat{p}(c\,|\,d)$ регуляризатор $R(\Psi,\Theta)$ равен взвешенной сумме ковариаций дискретных распределений $\hat{p}(c\,|\,d)$ и $p(c\,|\,d)$. Возьмём в качестве $\hat{p}(c\,|\,d)$ равномерное распределение (6.2). Тогда регуляризатор будет нечувствителен к тому, в каких долях модель $p(c\,|\,d)$ распределяет документ d по категориям из C_d.

Перегруппируем слагаемые в регуляризаторе:

$$L(\Phi,\Theta) + R(\Psi,\Theta) = \sum_{d,w} n_{dw} \ln \sum_t \varphi_{wt}\theta_{td} + \tau \sum_{c,t}\psi_{ct}\sum_d m_{dc}\theta_{td} \to \max.$$

Решение по переменным ψ_{ct} находится аналитически благодаря тому, что максимизируемый функционал линеен по ψ_{ct} на симплексе $\psi_{ct} \geqslant 0$, $\sum\limits_{c\in C}\psi_{ct} = 1$.

Чтобы функционал достигал максимума, для каждой темы t, независимо от остальных тем, переменная ψ_{ct} должна принимать максимальное значение, равное 1, при таком $c = c^*(t)$, для которого коэффициент при ψ_{ct} максимален:

$$c^*(t) = \arg\max_{c\in C}\sum_{d\in D} m_{dc}\theta_{td} = \arg\max_{c\in C}\sum_{d\in D}\frac{[c\in C_d]}{|C_d|}n_d\theta_{td}.$$

В силу нормировки, значения ψ_{ct} должны обращаться в нуль для остальных c. Следовательно, распределение ψ_{ct} в каждом столбце матрицы Ψ является вырожденным:

$$\psi_{ct} = \big[c = c^*(t)\big].$$

Таким образом, каждой теме t соответствует только одна *родительская* категория $c^*(t)$, и ковариационный регуляризатор действительно решает поставленную задачу разбиения категорий на темы.

Подставим найденное решение ψ_{ct} в регуляризатор, который теперь будет зависеть только от матрицы Θ:

$$R(\Theta) = \tau \sum_{t\in T}\sum_{d\in D}\frac{[c^*(t)\in C_d]}{|C_d|}n_d\theta_{td} \to \max.$$

Воспользуемся формулой М-шага (5.2). Формулы для φ_{wt} ничем не отличаются от стандартных, поскольку регуляризатор не зависит от Φ. Формулы для θ_{td} требуют предварительно найти родительскую категорию $c^*(t)$ для каждой темы t:

$$\theta_{td} \propto \hat{n}_{dt} + \tau \frac{n_d}{|C_d|}[c^*(t) \in C_d].$$

Данная формула имеет прозрачную интерпретацию: если документ d относится к родительской категории темы t, то к нему добавляется «виртуальный термин» темы t с частотой $\tau n_d/|C_d|$.

Категоризация новых документов после того, как построена тематическая модель, является тривиальной задачей. Она сводится к переходу от распределения тем в документе $p(t \mid d) = \theta_{td}$ к распределению категорий в документе $p(c \mid d) = \sum_t \psi_{ct}\theta_{td}$. Разреженность матриц Ψ и Θ позволяет находить это распределение для любого документа практически мгновенно.

7 Динамические тематические модели

Динамические (temporal) тематические модели учитывают дополнительную информацию о времени создания или публикации документов. Будем полагать, что каждый документ d связан с моментом времени y_d из заданного конечного линейно упорядоченного множества Y. Для научных статей это может быть год публикации. Для новостных сообщений используются более мелкие отсчёты времени — неделя, день или час. При построении динамических моделей наряду со стандартными распределениями $\varphi_{wt} = p(w \mid t)$ и $\theta_{td} = p(t \mid d)$ оцениваются распределения каждой темы во времени $\xi_{yt} = p(y \mid t)$. Они нужны для того, чтобы отобразить динамику изменения тем во времени. Примеры таких визуализаций можно найти в работах [68, 55].

Расширим вероятностное пространство до множества $D \times W \times T \times Y$. Будем считать, что с каждым словом в документе (d, w) связана не только тема $t \in T$, но и момент $y \in Y$. Введём обозначение Y_+ для множества Y без начального (нулевого) момента времени. Обозначим через $\Xi = (\xi_{yt})_{Y \times T}$ матрицу новых неизвестных.

Примем гипотезу условной независимости: $p(y \mid d, t) = p(y \mid t)$, означающую, что модель должна объяснять наблюдаемую отметку времени y_d исходя только из тематики документа d.

Заметим, что, поскольку множество моментов времени Y дискретно, все вероятностные предположения с точностью до обозначений те же, что и в тематической модели классификации.

Рассмотрим две динамические модели. *Модель с фиксированной тематикой* основана на предположении, что темы не меняются со временем, $p(w \mid t, y) = p(w \mid t)$. *Модель с медленно меняющейся тематикой* ослабляет это предположение и в явном виде оценивает распределения слов в темах $p(w \mid t, y)$ как зависящие от времени.

§7.1 Модель с фиксированной тематикой

Запишем распределение моментов времени в произвольном документе d как вероятностную смесь распределений моментов в темах и тем в документе:

$$p(y \mid d) = \sum_{t \in T} p(y \mid t)\, p(t \mid d) = \sum_{t \in T} \xi_{yt} \theta_{td}.$$

Будем считать, что момент времени y_d является случайным наблюдением, выбранным из распределения $p(y \mid d)$ для каждого слова документа d. Тогда эмпирические распределения $\hat{p}(y \mid d) = [y = y_d]$ являются вырожденными. Обозначим через $m_{dy} = [y = y_d] n_d$ число терминов документа d, относящихся к моменту времени y.

Возьмём в качестве регуляризатора логарифм правдоподобия модели $p(y \mid d)$:

$$R_1(\Xi, \Theta) = \tau_1 \sum_{d \in D} n_d \sum_{y \in Y} \hat{p}(y \mid d) \ln p(y \mid d) = \tau_1 \sum_{d \in D} m_{dy} \ln \sum_{t \in T} \xi_{yt} \theta_{td} \to \max.$$

Введённая модель $p(y \mid d)$ с регуляризатором R_1 ничем по сути не отличается от модели классификации (6.3) с регуляризатором (6.4).

Второй регуляризатор более специфичен для динамических задач. Он формализует предположение, что тематика меняется медленно, поэтому вероятности ξ_{yt} в последовательные моменты времени должны быть близки:

$$R_2(\Xi) = -\frac{\tau_2}{2} \sum_{y \in Y_+} \sum_{t \in T} \big(\xi_{yt} - \xi_{y-1,t}\big)^2 \to \max.$$

Задачу максимизации функционала $L(\Phi, \Theta) + R_1(\Xi, \Theta) + R_2(\Xi)$ будем решать с помощью EM-алгоритма.

На E-шаге дополнительно оценивается условная вероятность

$$H'_{dyt} = p(t \mid d, y) = \frac{\xi_{yt} \theta_{td}}{p(y \mid d)} = \frac{\xi_{yt} \theta_{td}}{\sum\limits_{s \in T} \xi_{ys} \theta_{sd}}.$$

На M-шаге φ_{wt} оценивается по прежним формулам:

$$\varphi_{wt} \propto \hat{n}_{wt}, \qquad \hat{n}_{wt} = \sum_{d \in D} n_{dw} H_{dwt}.$$

Оценки ξ_{yt} аналогичны оценкам φ_{wt} после замены w на y и H_{dwt} на H'_{dyt}. Кроме того, на оценки ξ_{yt} влияет регуляризатор R_2, выполняющий функцию сглаживания. Он увеличивает значение ξ_{yt}, если оно меньше полусуммы соседних вероятностей $\xi_{y-1,t}$, $\xi_{y+1,t}$, и уменьшает, если оно больше их полусуммы:

$$\xi_{yt} \propto \tau_1 \hat{m}_{yt} + \tau_2 \xi_{yt} \big(\xi_{y-1,t} + \xi_{y+1,t} - 2\xi_{yt}\big), \quad \hat{m}_{yt} = \sum_{d \in D} m_{dy} H'_{dyt}.$$

Оценки θ_{td} агрегируют счётчики терминов и моментов времени:

$$\theta_{td} \propto \hat{n}_{dt} + \tau_1 \hat{m}_{dt}, \qquad \hat{n}_{dt} = \sum_{w \in d} n_{dw} H_{dwt}, \qquad \hat{m}_{dt} = \sum_{y \in Y} m_{dy} H'_{dyt}.$$

Если есть основания полагать, что большинство тем существуют только в ограниченном интервале времени, то можно ввести дополнительный регуляризатор

$$R_3(\Xi) = \tau_3 \sum_{y \in Y} \sum_{t \in T} \big[\xi_{yt} = 0\big] \to \max$$

и использовать принудительное разреживание (§5.3).

§7.2 Модель с медленно меняющейся тематикой

§7.3 Модели с непрерывным временем

8 Иерархические тематические модели

Этот раздел устарел и будет целиком переписан ToDo[21]

Для больших коллекций текстовых документов естественно строить иерархии вложенных друг в друга тем (называемых также категориями или рубриками), чтобы упростить поиск документов. Иерархия — это общепринятый способ структуризации знаний. Однако разделение тем на более узкие подтемы субъективно, неоднозначно и часто вызывает споры среди специалистов.

В статье [66] приводится обзор иерархических тематических моделей и отмечается, что оптимизация структуры иерархии по коллекции документов является открытой проблемой; более того, разработка объективной количественной оценки качества иерархии — также открытая проблема.

Многие иерархические модели имеют те или иные неестественные ограничения: либо фиксируется число уровней, либо фиксируется число подтем в каждой теме или на каждом уровне, либо документ не может относиться к темам из различных ветвей дерева, либо темы не могут иметь общую подтему, либо темам во внутренних узлах не сопоставляется распределение на множестве терминов.

§8.1 Определение тематического дерева

Гипотеза о существовании тематического дерева. Рассмотрим дерево с множеством вершин V и корнем $t_0 \in V$. Вершины дерева соответствуют темам. Каждой теме $t \in V$ соответствует множество её подтем — дочерних вершин в дереве $S_t \subset V$. Каждое ребро дерева соответствует паре «тема–подтема» (t, s), $s \in S_t$. Если $S_t = \varnothing$, то тема t называется *терминальной* или *листом* тематического дерева. Для каждой вершины t в дереве V существует только одна родительская вершина, следовательно, только один путь $(t_0, \ldots, t)$ от корня дерева t_0 до темы t.

Ранее мы предполагали, что каждое вхождение термина w в документ d связано только с одной темой t. Теперь примем за аксиому другие предположения:

1) если пара (d, w) связана с темой t, то она связана и со всеми темами выше вершины t на пути до корня t_0;

2) если пара (d, w) не связана с темой t, то она не связана и со всеми подтемами в поддереве ниже вершины t.

Этих двух предположений, совершенно не вероятностного характера, достаточно, чтобы построить иерархическую вероятностную тематическую модель.

Вероятностная интерпретация отношения «тема–подтема». Каждому ребру тематического дерева (t, s) соответствует условная вероятность $p(s \mid t)$ того, что термин документа, связанный с темой t, связан также с подтемой $s \in S_t$:

$$p(s \mid t) = \frac{p(t, s)}{p(t)} = \frac{p(s)}{p(t)}. \tag{8.1}$$

Если рассматривать коллекцию документов как выборку троек (d, w, t), то частотной оценкой этой условной вероятности будет $\hat{p}(s \mid t) = n_s/n_t$ — доля троек, связанных с подтемой s, среди всех троек, связанных с темой t.

Условные вероятности подтем удовлетворяют ограничениям нормировки, которые, в силу (8.1), допускают две эквивалентные записи:

$$\sum_{s \in S_t} p(s \mid t) = 1, \quad \sum_{s \in S_t} p(s) = p(t), \quad t \in V. \tag{8.2}$$

Обозначим через T множество тем, соответствующих терминальным вершинам дерева V. Условие нормировки

$$\sum_{t \in T} p(t) = 1. \tag{8.3}$$

выполняется именно для этого множества, а не для всего множества тем в дереве V. Из (8.2) следует, что условие нормировки (8.3) останется в силе, если заменить любое из множеств $S_t \subseteq T$ его родительской темой t, а также если делать такие замены многократно в произвольном порядке, вплоть до корневой темы t_0, $p(t_0) = 1$.

При разделении темы t на подтемы $s \in S_t$ условные распределения для подтем $\varphi_{ws} = p(w \mid s)$ и $\theta_{sd} = p(s \mid d)$ должны удовлетворять требованиям нормировки

$$\sum_{w \in W} \varphi_{ws} = 1, \quad s \in S_t; \qquad \sum_{s \in S_t} \theta_{sd} = \theta_{td}, \quad d \in D. \tag{8.4}$$

Распределения $p(s \mid w) = \varphi_{ws} \frac{p(s)}{p(w)}$ и $p(d \mid s) = \theta_{sd} \frac{p(d)}{p(s)}$ также должны быть нормированы, откуда следуют ещё две серии тождеств:

$$\sum_{s \in S_t} \varphi_{ws}\, p(s) = \varphi_{wt}\, p(t), \quad w \in W; \qquad \sum_{d \in D} \theta_{sd}\, p(d) = p(s), \quad s \in S_t. \tag{8.5}$$

Документы во внутренних вершинах. В некоторых приложениях важно, чтобы документы и термины могли относиться не только к терминальным вершинам, но и к любым внутренним вершинам тематического дерева. В частности, это могут быть документы, относящиеся сразу к нескольким подтемам, либо новые документы, которые пока не выделились в отдельную подтему.

Для каждой внутренней вершины $t \in V \backslash T$ создаётся выделенная терминальная вершина — подтема $s_0 \in S_t$. Если документ или термин попадает в s_0, то считается, что он остался в теме t. В терминах кластеризации выделенная подтема s_0 — это специальный «фоновый» кластер, к которому относится всё, что не удалось с уверенностью отнести к другим кластерам — подтемам темы t.

К выделенной подтеме s_0 естественно предъявлять требование минимизации числа документов и описывать её тем же распределением, что и родительскую тему t.

§8.2 Фиксированная иерархия

Рассмотрим случай, когда структура тематического дерева $\{S_t \colon t \in V\}$ фиксирована. Чтобы каждая тема $t \in T$ имела интерпретацию, к ней привязывается

множество документов $D_t \subset D$ и множество терминов $W_t \subset W$. Одно из этих множеств может быть пустым. Таким образом, ставится задача *частичного обучения* (semi-supervised learning) иерархической тематической модели.

Распределение $\theta_{td} = p(t \mid d)$, полученное в результате тематического моделирования, непосредственно решает задачу категоризации — документ d относится к тем темам (категориям), для которых вероятность θ_{td} превышает заданный порог. Для решения задач категоризации лучше подходят разреженные модели, в которых малые вероятности θ_{td} обнуляются в процессе построения модели. Обнуление или игнорирование малых вероятностей в уже построенной модели может приводить к менее адекватным результатам.

Для категоризации текстов часто применяется другой подход: для каждой пары тема–подтема строится классификатор на два класса [49]. Каждый классификатор обучается по выборке документов, относящихся к родительской теме, что требует больших затрат времени и памяти. Иерархическая тематическая модель, очевидно, является более естественным инструментом категоризации.

Иерархический Алгоритм 8.1 основан на онлайновом Алгоритме 2.4. Основное отличие PLSA-HOEM в том, что для вычисления распределения θ_{td} тем t в документе d производится спуск по дереву от корня к терминальным вершинам.

Модификация формул M-шага для иерархической модели. Пусть $T \subset V$ — подмножество вершин дерева, удовлетворяющее условию нормировки (8.3), и для всех тем $t \in T$ известны значения параметров φ_{wt}, θ_{td}. Сначала T состоит из единственной корневой вершины t_0, для которой $\varphi_{wt_0} = p(w)$ и $\theta_{t_0 d} = 1$. Спуск по дереву — это итерационный процесс, на каждом шаге которого выбирается некоторое подмножество тем $R \subseteq T$, и для каждой вершины $s \in S$ из множества всех их подтем $S = \bigcup_{t \in R} S_t$, вычисляются значения параметров φ_{ws}, θ_{sd}. После этого множество T заменяется на $(T \backslash R) \cup S$ и начинается следующий шаг. Спуск продолжается, пока T не совпадёт с множеством терминальных вершин дерева.

Рассмотрим вероятностную модель (1.2) при ограничениях нормировки (8.4). Параметры φ_{wt} и θ_{td} для всех тем $t \in T \backslash R$ будем считать фиксированными. Обозначим через σ_{dw} фиксированную часть вероятностной тематической модели:

$$\sigma_{dw} = \sum_{t \in T \backslash R} \varphi_{wt} \theta_{td}, \quad d \in D,\ w \in d.$$

Задача оценивания параметров φ_{ws}, θ_{sd}, $s \in S$ сводится к максимизации логарифма правдоподобия, аналогично задаче (1.6), но оптимизируется только часть параметров $\Phi_S = (\varphi_{ws})_{W \times S}$ и $\Theta_S = (\theta_{sd})_{S \times D}$, связанных с темами из S:

$$\begin{aligned} L(\Phi_S, \Theta_S) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \Bigg(\sigma_{dw} + \sum_{s \in S} \varphi_{ws} \theta_{sd} \Bigg) \to \max_{\Phi_S, \Theta_S}; \\ \sum_{w \in W} \varphi_{ws} = 1, \quad s \in S; \\ \sum_{s \in S_t} \theta_{sd} = \theta_{td}, \quad t \in R, \quad d \in D. \end{aligned}$$

Как обычно, нужно записать лагранжиан, приравнять нулю его производные по переменным φ_{ws} и θ_{sd}, из полученных уравнений исключить двойственные переменные и выразить φ_{ws} и θ_{sd} через H_{dws}:

$$H_{dws} = \frac{\varphi_{ws}\theta_{sd}}{\sigma_{dw} + \sum\limits_{s' \in S} \varphi_{ws'}\theta_{s'd}}, \quad d \in D, \;\; w \in d, \;\; s \in S;$$

$$\varphi_{ws} = \frac{\sum\limits_{d \in D} n_{dw} H_{dws}}{\sum\limits_{w' \in W} \sum\limits_{d \in D} n_{dw'} H_{dw's}}, \quad w \in W, \;\; s \in S;$$

$$\theta_{sd} = \theta_{td} \frac{\sum\limits_{w \in d} n_{dw} H_{dws}}{\sum\limits_{s' \in S_t} \sum\limits_{w' \in d} n_{dw'} H_{dw's'}}, \quad d \in D, \;\; s \in S_t, \;\; t \in R;$$

или, в более компактной записи с использованием счётчиков:

$$\varphi_{ws} = \frac{\hat{n}_{ws}}{\hat{n}_s}, \qquad \hat{n}_s = \sum_{w \in W} \hat{n}_{ws}, \qquad \hat{n}_{ws} = \sum_{d \in D} n_{dw} H_{dws}. \tag{8.6}$$

$$\theta_{sd} = \theta_{td} \frac{\hat{n}_{ds}}{\hat{n}_{dt}}, \qquad \hat{n}_{dt} = \sum_{s \in S_t} \hat{n}_{ds}, \qquad \hat{n}_{ds} = \sum_{w \in d} n_{dw} H_{dws}. \tag{8.7}$$

Таким образом, формулы М-шага и Е-шага для иерархического алгоритма лишь немногим отличаются от обычного PLSA-EM.

Инициализация и частичное обучение. Привязки терминов и документов к темам задаются в виде начальных распределений φ^0_{wt} и θ^0_{td}, вычисляемых по формулам из §2.5 с нормировкой (8.4). При инициализации они смешиваются с неразреженными случайными распределениями, на каждом шаге EM-алгоритма — с оценками (8.6) и (8.7). Благодаря тому, что текущие приближения φ_{wt} и θ_{td} немного притягиваются к начальным распределениям φ^0_{wt} и θ^0_{td}, темы не уходят далеко от исходно заданных интерпретаций. Сила этого «притяжения» регулируется параметрами λ и μ.

Регуляризация Дирихле может быть добавлена в Алгоритм 8.1 обычным образом: частотные оценки условных вероятностей (8.6), (8.7) заменяются сглаженными:

$$\varphi_{ws} = \frac{\beta_w + \hat{n}_{ws}}{\sum\limits_{w' \in W} \big(\beta_{w'} + \hat{n}_{w's}\big)} = \frac{\beta_w + \hat{n}_{ws}}{\beta_0 + \hat{n}_s}. \tag{8.8}$$

$$\theta_{sd} = \theta_{td} \frac{\alpha_s + \hat{n}_{ds}}{\sum\limits_{s' \in S_t} \big(\alpha_{s'} + \hat{n}_{ds'}\big)} = \theta_{td} \frac{\alpha_s + \hat{n}_{ds}}{\alpha_t + \hat{n}_{dt}}. \tag{8.9}$$

Заметим, что при разделении темы t на множество подтем S_t расщепляются также и гиперпараметры распределения Дирихле $\mathrm{Dir}(\theta_d; \alpha)$, а их сумма α_0 не меняется. Это следует из условий нормировки (8.4) и свойства (3.1):

$$\sum_{s \in S_t} \theta_{sd} = \theta_{td} \quad \Rightarrow \quad \sum_{s \in S_t} \mathsf{E}\theta_{sd} = \mathsf{E}\theta_{td} \quad \Rightarrow \quad \sum_{s \in S_t} \alpha_s = \alpha_t.$$

Алгоритм 8.1. PLSA-HOEM: иерархический онлайновый EM-алгоритм.

Вход: коллекция документов D; параметры λ и μ,
множество тем V и структура тематического дерева $\{S_t \colon t \in V\}$,
привязки терминов и документов к темам φ^0_{wt} и θ^0_{td};
Выход: распределения Θ и Φ;

инициализировать φ_{wt} с учётом φ^0_{wt} для всех $w \in W$, $t \in V$;
повторять
  $\hat{n}_{wt} := 0$; $\hat{n}_t := 0$ для всех $w \in W$, $t \in V$;
  для всех $d \in D$
    инициализировать θ_{td} с учётом θ^0_{td} для всех $t \in V$;
    $T := \{t_0\}$; $R := \{t_0\}$; $\theta_{t_0 d} = 1$;
    пока множество подтем $S := \bigcup_{t \in R} S_t$ не пусто
      $\sigma_{dw} = \sum_{t \in T \setminus R} \varphi_{wt}\theta_{td}$ для всех $w \in d$;
      повторять
        $Z_w := \sigma_{dw} + \sum_{s \in S} \varphi_{ws}\theta_{sd}$ для всех $w \in d$;
        $n_s := \sum_{w \in d} n_{dw}\varphi_{ws}\theta_{sd}/Z_w$ для всех $w \in d$;
        $n := \sum_{s \in S} n_s$;
        $\theta_{sd} := \mu\theta^0_{sd} + (1-\mu)\theta_{td} n_s/n$ для всех $s \in S_t$, $t \in R$;
      пока θ_{sd} не сойдутся для всех $s \in S$;
      увеличить $\hat{n}_{ws}$, $\hat{n}_s$ на $n_{dw}\varphi_{ws}\theta_{sd}/Z_w$ для всех $w \in d$, $s \in S$;
      $T := (T \setminus R) \cup S$; $R := S$;
  $\varphi_{wt} := \lambda\varphi^0_{wt} + (1-\lambda)\hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in V$;
пока Φ не сойдутся;

§8.3 Реконструкция иерархии

9 Многоязычные тематические модели

§9.1 Параллельные тексты

§9.2 Сопоставимые тексты

§9.3 Регуляризация матрицы переводов слов

10 Модели текста как последовательности слов

§10.1 Коллокации

§10.2 Марковские модели синтаксиса языка

§10.3 Выделение ключевых фраз

§10.4 Тематическая структура документа

11 Многомодальные тематические модели

§11.1 Коллаборативная фильтрация

§11.2 Модель научной социальной сети

§11.3 Персонализация рекламы в Интернете

12 Критерии качества тематических моделей

Этот раздел устарел и будет целиком переписан ToDo[22]

§12.1 Внутренние оценки качества тематических моделей

Оценивание качества тематических моделей является нетривиальной проблемой. В отличие от задач классификации или регрессии здесь нет чёткого понятия «ошибки» или «потери». Стандартные критерии качества кластеризации типа средних внутрикластерных или межкластерных расстояний или их отношений плохо подходят для оценивания «мягкой» совместной кластеризации документов и терминов.

Наиболее распространённым критерием является *перплексия* (perplexity), используемая для оценивания моделей языка в компьютерной лингвистике. Это мера несоответствия или «удивлённости» модели $p(w\,|\,d)$ терминам w, наблюдаемым в документах d коллекции D, определяемая через логарифм правдоподобия (1.6):

$$\mathscr{P}(D;p) = \exp\Big(-\frac{1}{n}L(\Phi,\Theta)\Big) = \exp\bigg(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w\,|\,d)\bigg). \tag{12.1}$$

Чем меньше эта величина, тем лучше модель p предсказывает появление терминов w в документах d коллекции D.

Интерпретация перплексии. Если термины w порождаются из равномерного распределения $p(w) = 1/V$ на словаре мощности V, то перплексия модели p на таком тексте сходится к V с ростом его длины. Чем сильнее распределение p отличается от равномерного, тем меньше перплексия. Чем сильнее модель p отличается от генерирующего распределения, тем больше перплексия. В нашем случае в (12.1) используются условные вероятности терминов $p(w\,|\,d)$, и интерпретация немного другая: если каждый документ генерируется из V равновероятных терминов (возможно, различных в разных документах), то перплексия сходится к V. Опять-таки, чем сильнее распределение отличается от равномерного, тем меньше перплексия.

Чтобы сравнение перплексии двух коллекций было корректным, необходимо, чтобы они имели один и тот же словарь. ToDo[23]

Чтобы перплексия была характеристикой только качества модели, необходимо вводить нормировки, чтобы длины документов и эффективная мощность словаря не влияли на перплексию. ToDo[24]

Перплексия контрольной выборки. Обозначим через $p_D(w\,|\,d)$ модель, построенную по обучающей коллекции документов D. Перплексия обучающей выборки $\mathscr{P}(D;p_D)$ является оптимистично смещённой (заниженной) характеристикой качества модели из-за эффекта переобучения. Обобщающую способность модели принято оценивать *перплексией контрольной выборки* (hold-out perplexity) $\mathscr{P}(D';p_D)$.

Вопрос о том, как разделить исходную коллекцию на обучение D и контроль D', не тривиален. К сожалению, детали этой процедуры во многих статьях опускаются. В [9] предлагается разделять все документы на обучающие и контрольные случайным образом в пропорции $9:1$. Однако в силу гипотез «мешка слов» и «мешка документов» более корректным было бы случайное разбиение каждого документа на обучающую и контрольную части. С другой стороны, во многих приложениях важно проверить способность тематической модели хорошо описывать новые документы.

Новые документы порождают две проблемы: во-первых, для них необходимо оценивать θ_{td}; во-вторых, они могут содержать новые термины w, для которых придётся оценивать также φ_{wt}, увеличивать размерность векторов $\varphi_t = (\varphi_{wt})_{w\in W}$ и перенормировать их. Такая процедура оценивания модели частично включает в себя процедуру обучения, в результате чего оценка качества снова может оказаться оптимистично смещённой.

Частичное решение этой проблемы предлагается в [8]. После обучения модели p_D векторы φ_t фиксируются, векторы θ_d контрольных документов $d\in D'$ оцениваются по первой половине каждого документа, по вторым половинам вычисляется контрольная перплексия. Что такое «половина», не уточняется. Простое разрезание текста на две части может приводить к смещённым оценкам. Например, научные статьи обычно начинаются с введения и обзора, использующих общую терминологию, затем идёт изложение частных результатов. Если в коллекции много таких текстов, то оценка окажется пессимистично смещённой. Противоположный пример неслучайного разбиения текста — когда число вхождений каждого термина n_{dw} делится ровно пополам между обучающей и контрольной выборками. В таком случае обучающая и контрольная половины документа будут неразличимы для тематической модели, и оценка окажется оптимистично смещённой.

В наших экспериментах последовательность терминов $\{w_1,\ldots,w_{n_d}\}$ каждого контрольного документа $d\in D'$ после случайной перестановки разбивается на две части равной длины. Новые слова, попадающие во вторую часть, игнорируются.

Ещё один выход — робастные модели. Новые редкие слова считаются шумом, описываются униграммной моделью и почти не дают вклада в контрольную перплексию. Робастную модель трудно удивить новыми словами, т.к. она трактует их как шум. ToDo[25]

Более сложные процедуры несмещённого оценивания правдоподобия предложены в [60] и улучшены в [10]. Они имеют трудоёмкость, квадратичную по длине документа, и в процессе оценивания используют ту же тематическую модель, качество которой оценивается. Эти недостатки несколько ограничивают их применимость.

§12.2 Критерии условной независимости

Гипотеза условной независимости $p(w\,|\,d,t) = p(w\,|\,t)$ чрезвычайно важна для вероятностных тематических моделей. Именно она обеспечивает переход к компактному представлению данных $F\approx\Phi\Theta$. Для её проверки не требуется выделять контрольную выборку, что является преимуществом данного типа критериев.

Оба распределения оцениваются в EM-алгоритме:

$$\hat{p}(w\,|\,d,t) = \frac{n_{dwt}}{\hat{n}_{dt}}, \quad t \in T,\ d \in D;$$
$$\hat{p}(w\,|\,t) = \frac{\hat{n}_{wt}}{\hat{n}_t}, \quad t \in T.$$

Рассмотрим статистические тесты, проверяющие нулевую гипотезу о том, что различия между этими распределениями незначимы, точнее, что выборка с эмпирическим распределением $\hat{p}(w\,|\,d,t)$ могла быть получена из генеральной совокупности с распределением $\hat{p}(w\,|\,t)$.

Точный тест Фишера. Рассмотрим следующий статистический эксперимент, связанный с каждой тройкой (d,w,t). Имеется последовательность из n_t терминов, в которой термин w встречается ровно n_{wt} раз. Из этой последовательности случайно и независимо выбираются n_{dt} терминов. Какова вероятность, что термин w окажется в числе выбранных не более n_{dwt} раз? В условиях истинности нулевой гипотезы эта вероятность описывается функцией гипергеометрического распределения:

$$P_{dwt} = \sum_{i=0}^{n_{dwt}} \frac{C_{n_{wt}}^{i}\, C_{n_t-n_{wt}}^{n_{dt}-i}}{C_{n_t}^{n_{dt}}}.$$

дописать

ToDo[26]

Критерий X^2 Пирсона основан на вычислении статистики хи-квадрат, которая является естественной мерой различия двух распределений:

$$X_{dt}^2 = \sum_{w \in W_{dt}} \frac{(E_{dwt} - n_{dwt})^2}{E_{dwt}} = \hat{n}_{dt} \sum_{w \in W_{dt}} \frac{\big(\hat{p}(w\,|\,t) - \hat{p}(w\,|\,d,t)\big)^2}{\hat{p}(w\,|\,t)},$$

где $E_{dwt} = \hat{n}_{dt}\hat{p}(w\,|\,t)$ — ожидаемое число вхождений термина w в документ d, связанных с темой t, $W_{dt} = \big\{w \in W\colon E_{dwt} > 0\big\}$.

Если значение X_{dt}^2 превышает $(1-\alpha)$-квантиль распределения хи-квадрат $\chi^2_{k,1-\alpha}$ с числом степеней свободы $k = |W_{dt}| - 1$, то нулевая гипотеза отвергается.

Условием применимости асимптотики χ_k^2 считается наличие достаточного числа наблюдений во всей выборке, $\hat{n}_{dt} \geqslant 50$, а также достаточного ожидаемого числа наблюдений каждого термина, $E_{dwt} \geqslant 5$. Второе требование в типичном случае не выполняется для большинства терминов w, так как распределение $\hat{p}(w\,|\,t)$, как правило, разрежено, более того, мощность словаря W_{dt} может превышать длину документа $\hat{n}_{dt}$. Таким образом, в нашем случае критерий Пирсона применять нельзя. Для случая разреженных распределений больше подходят статистики G^2 и D^2.

Статистика G^2 определяется через дивергенцию Кульбака–Лейблера, и для неё также справедливо асимптотическое распределение χ_k^2 с тем же числом степеней свободы, но при менее жёстких требованиях к числу наблюдений:

$$G_{dt}^2 = 2 \sum_{w \in W_{dt}} n_{dwt} \ln \frac{n_{dwt}}{E_{dwt}}.$$

Статистика D^2 — это поправка к статистике X^2, предложенная Зельтерманом в [67] специально для случая разреженных распределений:

$$D^2_{dt} = \sum_{w \in W_{dt}} \frac{(E_{dwt} - n_{dwt})^2 - n_{dwt}}{E_{dwt}}.$$

Эта статистика имеет асимптотически нормальное распределение. Особенности её применения обсуждаются в [53, 25].

Семейство функций расстояния Кресси–Рида. Для сравнения эмпирической функции вероятности $\hat{p}(w)$, оцененной по выборке длины n, с истинной функцией вероятности $p(w)$ принято использовать функции расстояния, придающие больший вес малым вероятностям:

$$\mathrm{KL}(\hat{p} \,\|\, p) = \sum_w \hat{p}(w) \ln \frac{\hat{p}(w)}{p(w)} \quad \text{— дивергенция Кульбака–Лейблера;} \tag{12.2}$$

$$X^2(\hat{p}, p) = \sum_w \frac{\big(p(w) - \hat{p}(w)\big)^2}{p(w)} \quad \text{— ненормированная } \chi^2\text{-статистика;} \tag{12.3}$$

$$H^2(\hat{p}, p) = \sum_w \Big(\sqrt{p(w)} - \sqrt{\hat{p}(w)}\Big)^2 \quad \text{— расстояние Хеллингера.} \tag{12.4}$$

Эти и другие «разумные» функции расстояния обобщаются (с точностью до константного множителя) параметрическим семейством дивергенций Кресси–Рида [13, 46]:

$$\mathrm{CR}_\lambda(\hat{p} : p) = \frac{2}{\lambda(\lambda+1)} \sum_w \hat{p}(w) \left(\left(\frac{\hat{p}(w)}{p(w)} \right)^\lambda - 1 \right).$$

дописать, сделать эксперименты

ToDo[27]

Перестановочный тест основан на использовании эмпирического распределения статистики, полученного путём сэмплирования большого числа выборок в условиях истинности нулевой гипотезы. Перестановочные тесты применяются в тех случаях, когда функция распределения статистики неизвестна или имеет слишком сложный вид или её известные асимптотики не достаточно точны.

Пусть S — одна из статистик X^2, G^2, D^2. Зафиксируем тему t. Сгенерируем N независимых выборок терминов из распределения $\hat{p}(w \,|\, t)$. Для каждой из них вычислим эмпирическое распределение $\hat{p}(w)$ и значение статистики S. По выборке значений статистики $\{S_1, \ldots, S_N\}$ построим эмпирическое распределение $\hat{F}_t(S)$ и найдём его $(1-\alpha)$-квантиль $\hat{F}_{t,1-\alpha}$. Число N должно быть порядка 10^3 при $\alpha = 0.05$.

Обозначим через S_{dt} значение статистики S, вычисленное по распределению $\hat{p}(w \,|\, d, t)$ для заданных $t \in T$ и $d \in D$. Поскольку распределение $\hat{F}_t(S)$ построено в условиях истинности нулевой гипотезы, неравенство $S_{dt} > \hat{F}_{t,1-\alpha}$ является критерием отклонения нулевой гипотезы для документа d на уровне значимости α.

Заметим, что квантиль $\hat{F}_{t,1-\alpha}$ достаточно вычислить один раз для каждой темы t и использовать для всех документов $d \in D$, что даёт значительную экономию времени. Однако при изменении распределения $\hat{p}(w \,|\, t)$ распределение $\hat{F}_t(S)$ и его квантиль придётся пересчитать заново.

Оценки средней несогласованности для документов и тем. Введём индикатор события «*тема* t *не согласована в документе* d при уровне значимости α»:

$$B_{dt}(\alpha) = \big[S_{dt} > \hat{F}_{t,1-\alpha}\big];$$

Определим *среднюю несогласованность* темы, документа и тематической модели в целом при уровне значимости α:

$$B_t(\alpha) = \sum_{d\in D} \frac{\hat{n}_{dt}}{\hat{n}_t} B_{dt}(\alpha) \quad \text{— средняя несогласованность темы } t;$$

$$B_d(\alpha) = \sum_{t\in T} \frac{\hat{n}_{dt}}{n_d} B_{dt}(\alpha) \quad \text{— средняя несогласованность документа } d;$$

$$B(\alpha) = \sum_{d\in D}\sum_{t\in T} \frac{\hat{n}_{dt}}{n} B_{dt}(\alpha) \quad \text{— средняя несогласованность модели.}$$

Это нормированные величины, принимающие значения из отрезка $[0,1]$. Чем меньше средняя несогласованность, тем лучше модель описывает соответствующую тему t, документ d или всю коллекцию в целом.

Критерий условной независимости. В [34] предлагается ещё один критерий, оценивающий степень несоответствия темы $t \in T$ гипотезе условной независимости. Он основан на дивергенции Кульбака–Лейблера и может быть легко вычислен в EM-алгоритме на каждом проходе коллекции:

$$\mathrm{KL}_t = \mathrm{KL}\Big(\hat{p}(d,w\,|\,t) \;\Big\|\; \hat{p}(d\,|\,t)\,\hat{p}(w\,|\,t)\Big) = \sum_{d,w} \frac{n_{dwt}}{\hat{n}_t} \ln \frac{n_{dwt}}{E_{dwt}}.$$

Статистика $G_t^2 = \sum\limits_{d\in D} G_{dt}^2 = 2\hat{n}_t\,\mathrm{KL}_t$ имеет асимптотически распределение χ^2_k с числом степеней свободы $k = \sum\limits_{d\in D} |W_{dt}| - |W| - |D| + 1$. В силу разреженности распределения $\hat{p}(d,w\,|\,t)$ вместо критерия хи-квадрат лучше применять перестановочный тест. Гипотеза условной независимости принимается для темы t, когда значение статистики G_t^2 меньше критического.

Выделение несогласованных тем. Статистические критерии позволяют находить «неудачные» темы, которые целесообразно разбивать на подтемы, непосредственно во время итераций EM-алгоритма. Темы можно ранжировать и сравнивать по значениям средней согласованности $B_t(\alpha)$ или статистики G_t^2. Заметим, что сравнивать темы по значению дивергенции KL_t некорректно, так как только после умножения на «длину темы» $\hat{n}_t$ получается величина $G_t^2 = 2\hat{n}_t\,\mathrm{KL}_t$, имеющая (асимптотически) одинаковое распределение для всех тем.

Эксперименты Влады Целых ToDo[28]

§12.3 Критерии качества классификации документов

Оценивание качества тематической модели упрощается в тех случаях, когда она строится с целью классификации или поиска документов. Каждый документ

описывается $|T|$-мерным вектором тем $\theta_d = \big(p(t \,|\, d)\big)_{t \in T}$. Качество модели определяется тем, насколько хорошо классифицируются документы, представленные этими векторами.

Пусть каждый документ $d \in D$ относится к классу $y_d \in Y$, алгоритм классификации $a\colon \mathbb{R}^{|T|} \to Y$ относит документ d к классу $a_d = a(\theta_d)$. В задачах информационного поиска и категоризации текстов качество классификации принято измерять в терминах точности и полноты [49].

Точность (precision) относительно класса $y \in Y$ определяется как доля правильно классифицированных документов среди всех документов, отнесённых алгоритмом a к классу y:

$$P_y(a) = \frac{\#\big\{d \in D\colon a_d = y_d = y\big\}}{\#\big\{d \in D\colon a_d = y\big\}}.$$

Полнота (recall) относительно класса $y \in Y$ определяется как доля правильно классифицированных документов среди всех документов класса y:

$$R_y(a) = \frac{\#\big\{d \in D\colon a_d = y_d = y\big\}}{\#\big\{d \in D\colon y_d = y\big\}}.$$

Чем больше значения точности и полноты, тем выше качество классификации.

В задачах информационного поиска обычно рассматривают два класса — документ либо «релевантен», либо «нерелевантен»; точность и полноту определяют только относительно класса релевантных документов.

Задачи категоризации, как правило, являются *многоклассовыми*, $|Y| \gg 2$. В таких случаях точность и полноту усредняют по всем классам.

В качестве агрегированного показателя, объединяющего точность P и полноту R, принято использовать F_1-меру:

$$F_1 = \frac{2PR}{P + R}.$$

§12.4 Критерии качества тематического поиска

Описать идею разбиения каждого документа на части и поиска одних частей по другим. Качество поиска может измеряться с помощью Mean Average Precision. ToDo[29]

§12.5 Интерпретируемость тем

§12.6 Когерентность

Тема называется *когерентной*, если термины, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции [41, 42]. Когерентность может оцениваться по сторонней коллекции (например, по Википедии) [39], либо по той же коллекции, по которой строится модель [35].

Предлагалось несколько оценок когерентности. Сначала использовалась *поточечная взаимная информация* (pointwise mutual information, PMI) [41, 42]:

$$\mathrm{PMI}(t) = \sum_{i=1}^{k-1} \sum_{j=i}^{k} \log \frac{N(w_i, w_j)}{N(w_i) N(w_j)},$$

где w_i — i-й термин в порядке убывания φ_{wt}, $N(w)$ — число документов, в которых термин w встречается хотя бы один раз, $N(w,w')$ — число документов, в которых термины w, w' встречаются рядом хотя бы один раз, число k обычно полагается равным 10. «Встречаются рядом» означает — в окне заданной ширины h, которая является параметром, обычно $h = 10$.

Затем в экспериментах [35] было показано, что более адекватной мерой когерентности является *логарифм условной вероятности* (log conditional probability, LCP), оценивающая вероятность менее частого слова при условии более частого:

$$\mathrm{LCP}(t) = \sum_{i=1}^{k-1} \sum_{j=i}^{k} \log \frac{N(w_i, w_j)}{N(w_i)},$$

§12.7 Точность восстановления модельных данных

Алгоритм 1.1 можно использовать для генерации модельных данных по заданным распределениям $p(w\,|\,t)$ и $p(t\,|\,d)$. Это крайне полезно на стадии тестирования методов обучения тематических моделей, решающих задачу (1.6). Хороший метод должен быть способен восстановить по данным ту самую модель, которая эти данные породила. Модельные данные можно генерировать различной длины n; можно добавлять в них шум — случайные пары (d_i, w_i) из распределения, заведомо плохо приближаемого моделью (1.2); можно задавать распределения $p(w\,|\,t)$, $p(t\,|\,d)$ более различными или более похожими, тем самым делая задачу восстановления модели более лёгкой или более трудной; задавать различное число тем $|T|$, а восстанавливать модель при другом числе тем, либо пытаться его определить. Эксперименты с варьированием модели данных позволяют исследовать устойчивость метода и узнать границы его применимости. Только в случае модельных данных известно, какая тема t_i на самом деле связана с каждой парой (d_i, w_i), что позволяет оценивать качество восстановления модели по данным как долю правильно угаданных тем или как расстояние между восстановленными и истинными распределениями $p(w\,|\,t)$, $p(t\,|\,d)$.

Показать эксперименты на модельных данных ToDo[30]

Лирическое отступление: задача о назначениях и венгерский алгоритм. ToDo[31]

13 Эксперименты с тематическими моделями

§13.1 Экспериментальные текстовые коллекции

Коллекция RuDis содержит $|D| = 2000$ авторефератов диссертаций на русском языке. Суммарная длина коллекции $n \approx 8.7 \cdot 10^6$ слов. Объём словаря $|W| \approx 3 \cdot 10^4$. Контрольная коллекция D' содержит 200 авторефератов. Предварительно сделана лемматизация и отброшены стоп-слова.

Коллекция NIPS содержит $|D| = 1566$ текстов статей научной конференции Neural Information Processing Systems на английском языке. Суммарная длина коллекции $n \approx 2.3 \cdot 10^6$ слов. Объём словаря $|W| \approx 1.3 \cdot 10^4$. Контрольная коллекция D' содержит 174 документов. Предварительно сделан стемминг и отброшены стоп-слова.

§13.2 Неустойчивость LDA (*Глушаченков В. В.*)

Задача тематического моделирования (1.6) является некорректно поставленной. Её решение не единственно, так как стохастическое матричное разложение определено с точностью до невырожденного преобразования: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$. Её решение неустойчиво, так как выбор преобразования S в EM-подобных алгоритмах никак не контролируется и зависит от начального приближения.

Устойчивость решения может повышаться, если исходные данные удовлетворяют гипотезе разреженности. Чем больше нулевых значений в матрицах Φ и Θ, тем меньше остаётся линейных преобразований S столбцов Φ, при которых преобразование S^{-1} над строками Θ оставляет все элементы матрицы Θ неотрицательными. При сильной разреженности разложение может оказаться единственным с точностью до перестановки тем.

Цель эксперимента. Показать, что алгоритмы PLSA и LDA-GS дают неустойчивые решения, зависящие от случайных начальных приближений. Показать, что устойчивость повышается с ростом разреженности исходных матриц Φ и Θ, порождающих коллекцию документов. Проверить, позволяет ли алгоритм постепенного принудительного разреживания повысить устойчивость решения или правильнее определить структуру разреженности матриц Φ и Θ.

Исходные данные и условия эксперимента. Все эксперименты проводились на модельных коллекциях, порождаемых известными матрицами Φ и Θ при $|D| = 500$, $|W| = 1000$, $|T| = 30$, длина документов n_d выбиралась случайно из равномерного распределения на $[100, 600]$.

Для генерации столбцов исходных матриц Φ и Θ применялось два подхода:

1) симметричные распределения Дирихле: $\varphi_t \sim \mathrm{Dir}(\beta)$, $\theta_d \sim \mathrm{Dir}(\alpha)$;

2) равномерные распределения с последующим обнулением заданной доли элементов R_φ, R_θ в каждом столбце матриц Φ и Θ соответственно.

На рис. 3 показана зависимость степени разреженности (доли нулевых элементов) модельных матриц Θ и Φ от гиперпараметров распределения Дирихле. Сильная разреженность (более 60% нулей) возникает при значении гиперпараметра менее 0.1.

Начальные приближения φ_{wt} и θ_{td} задавались путём обхода всей коллекции, при этом каждой паре (d, w) назначалась случайная тема t из равномерного распределения и вычислялись частотные оценки φ_{wt} и θ_{td} согласно (1.4).

Использовался алгоритм принудительного разреживания (см. §5.3, §13.4) с эвристикой упрощённой робастности (см. §4.3). Принудительное разреживание начиналось с i_0-го прохода коллекции, $i_0 = 10$; доля обнуляемых наименьших вероятностей в каждом распределении не превышала $r = 0.1$; сумма обнуляемых значений не превышала $S_\Phi = 0.001$, $S_\Theta = 0.1$.

Отклонение восстановленных распределений $\hat{p}(i \mid j)$ от модельных $p(i \mid j)$ измерялось средним расстоянием Хеллингера

$$H(\hat{p}, p) = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{2}\sum_{i=1}^{n}\Big(\sqrt{\hat{p}(i \mid j)} - \sqrt{p(i \mid j)}\Big)^2},$$

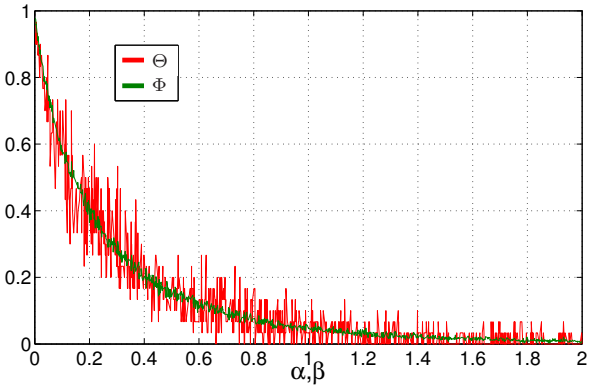
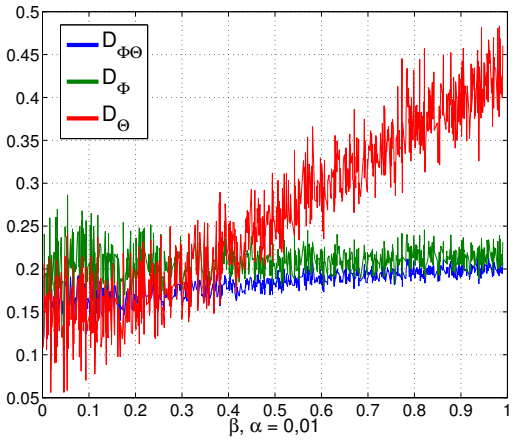
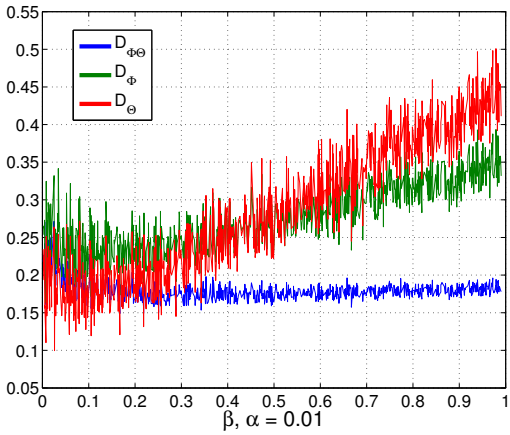


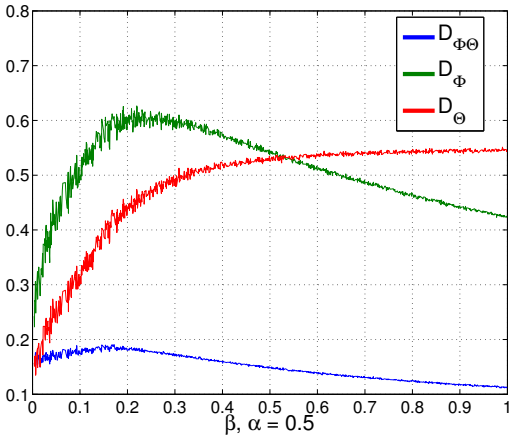
Рис. 3. Зависимость степени разреженности (доли нулевых элементов) модельных матриц Θ и Φ от гиперпараметров распределения Дирихле.



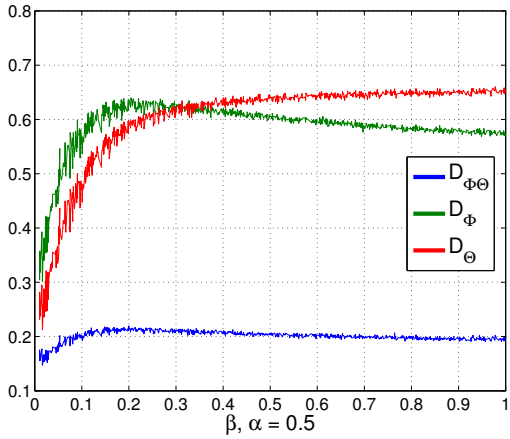
LDA-GS



PLSA-EM



LDA-GS



PLSA-EM

Рис. 4. Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Φ порождающей модели при фиксированной разреженности матрицы Θ с параметрами $\alpha = 0.01,\ 0.5$, для алгоритмов LDA-GS и PLSA-EM.

как для самих матриц Φ и Θ, так и для их произведения:

$$D_\Phi(\hat{\Phi}, \Phi) = H(\hat{\Phi}, \Phi);$$
$$D_\Theta(\hat{\Theta}, \Theta) = H(\hat{\Theta}, \Theta);$$
$$D_{\Phi\Theta}(\hat{\Phi}\hat{\Theta}, \Phi\Theta) = H(\hat{\Phi}\hat{\Theta}, \Phi\Theta).$$

Поскольку матрицы $\hat{\Phi}$, $\hat{\Theta}$ восстанавливаются с точностью до перестановки тем, перед сравнением к ним применялся венгерский алгоритм [37], который ищет перестановочную матрицу Π, минимизирующую функционал

$$f(\Pi) = D_\Phi(\hat{\Phi}\Pi, \Phi) + D_\Theta(\Pi^{-1}\hat{\Theta}, \Theta).$$

Качество восстановления структуры разреженности матриц Φ и Θ измерялось долей ошибок первого рода

$$S^1_\Phi = \frac{1}{|W||T|}\sum_{w,t}\big[\hat{\varphi}_{wt} > 0\big]\big[\varphi_{wt} = 0\big], \qquad S^1_\Theta = \frac{1}{|D||T|}\sum_{d,t}\big[\hat{\theta}_{dt} > 0\big]\big[\theta_{dt} = 0\big],$$

и долей ошибок второго рода

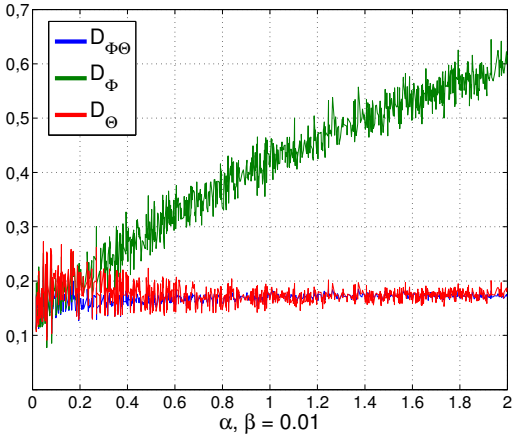
$$S^2_\Phi = \frac{1}{|W||T|}\sum_{w,t}\big[\hat{\varphi}_{wt} = 0\big]\big[\varphi_{wt} > 0\big], \qquad S^2_\Theta = \frac{1}{|D||T|}\sum_{d,t}\big[\hat{\theta}_{dt} = 0\big]\big[\theta_{dt} > 0\big].$$

Результаты. Графики на рис. 4 и рис. 5 показывают зависимости точности восстановления матриц Φ, Θ и $\Phi\Theta$ от разреженности матриц Φ и Θ в порождающей модели. Сравнивались алгоритмы LDA-GS и PLSA-EM. Гиперпараметры α и β в алгоритме LDA-GS полагались равными тем же значениям, которые использовались при генерации модельных данных. Оба алгоритма хорошо восстанавливают произведение $\Phi\Theta$. Однако сами матрицы Θ и Φ хорошо восстанавливаются только когда они сильно разрежены. При уменьшении разреженности оба алгоритма неустойчивы, причём PLSA-EM менее устойчив.

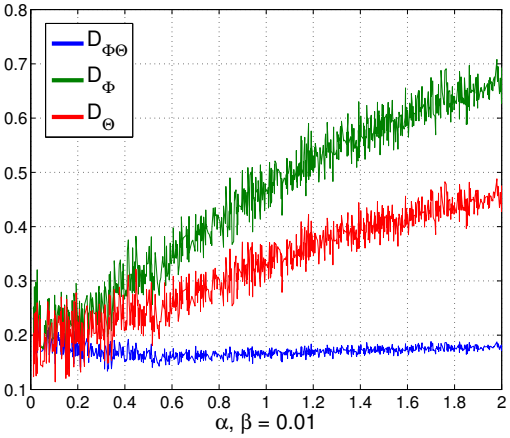
Графики на рис. 6 и рис. 7 показывают зависимости точности восстановления матриц Φ, Θ и $\Phi\Theta$ от разреженности матриц Φ и Θ в порождающей модели. Теперь сравниваются два варианта алгоритма PLSA-EM: стандартный и с принудительным разреживанием. Оба алгоритма хорошо восстанавливают матрицы Φ и Θ только когда они сильно разрежены (на 80% или более). Чем менее разрежены исходные матрицы Φ и Θ, тем меньшую точность даёт принудительное разреживание, что вполне естественно.

Преимущество принудительного разреживания проявляется в том случае, когда требуется восстановить структуру разреженности — узнать, какие именно элементы матриц Φ и Θ равны нулю.

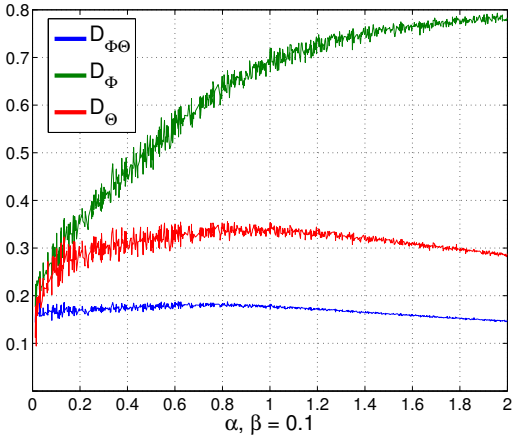
Графики на рис. 8 и рис. 9 показывают зависимости числа ошибок первого и второго рода при определении нулевых элементов в матрицах Φ, Θ. Разреживающий EM-алгоритм действительно лучше восстанавливает структуру разреженности почти на всем интервале. Ошибки первого рода меньше, следовательно разреживающий алгоритм правильнее определяет нулевые элементы. При сильной разреженности исходных данных (80% и более) ошибки второго рода малы у всех алгоритмов.



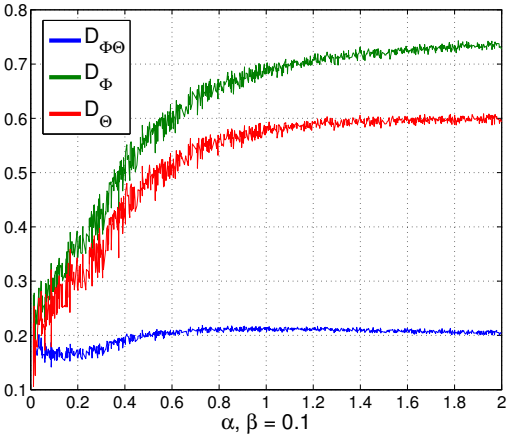
LDA-GS



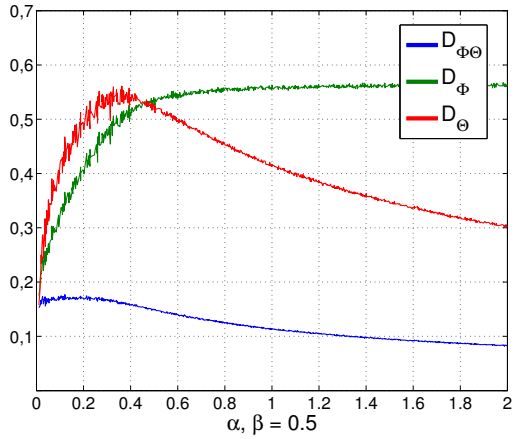
PLSA-EM



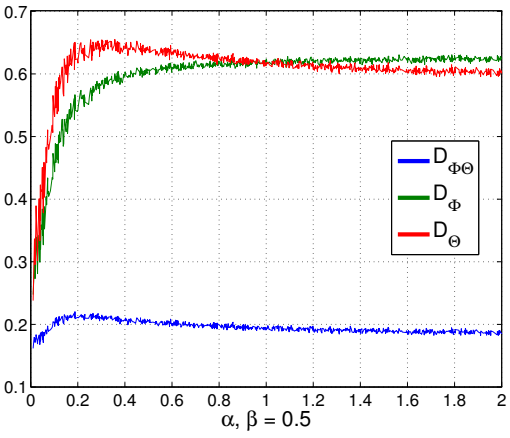
LDA-GS



PLSA-EM

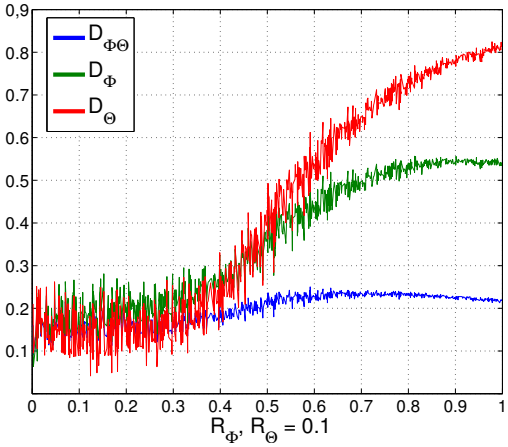


LDA-GS

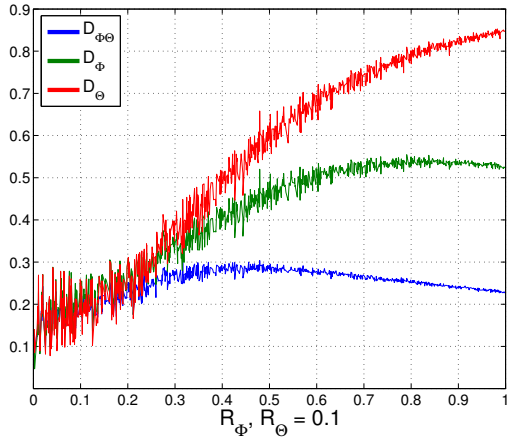


PLSA-EM

Рис. 5. Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Θ порождающей модели при фиксированной разреженности матрицы Φ с параметрами $\beta = 0.01,\ 0.1,\ 0.5$, для алгоритмов LDA-GS и PLSA-EM.

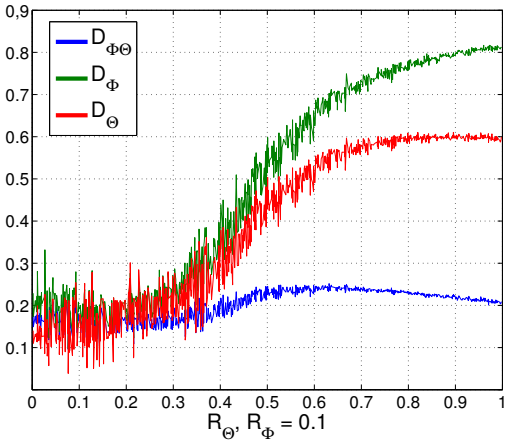


без разреживания

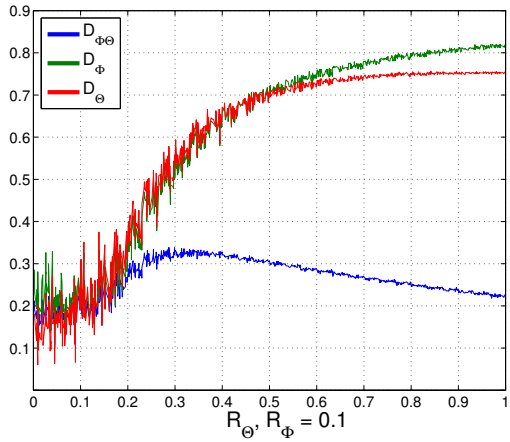


с разреживанием

Рис. 6. Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности R_Φ при фиксированной разреженности $R_\Theta = 0.1$.

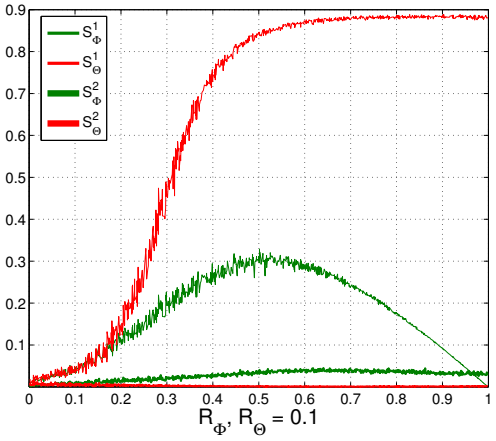


без разреживания

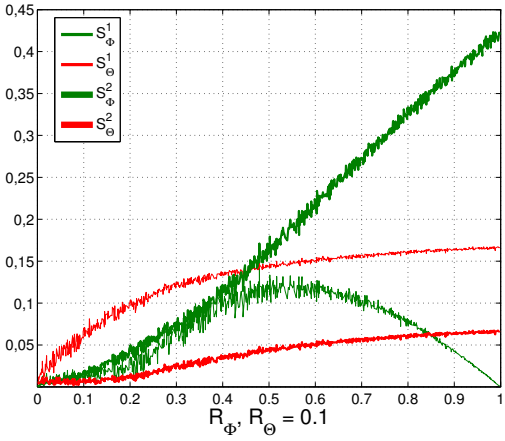


с разреживанием

Рис. 7. Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности R_Θ при фиксированной разреженности $R_\Phi = 0.1$.

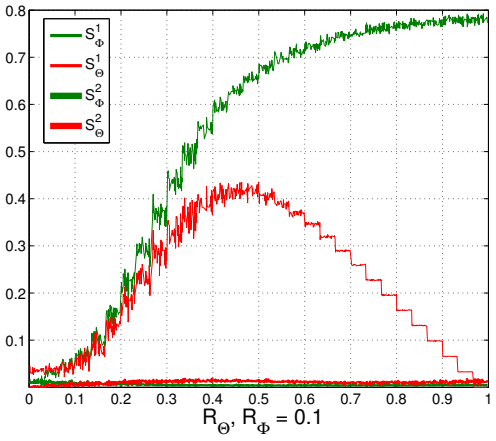


без разреживания

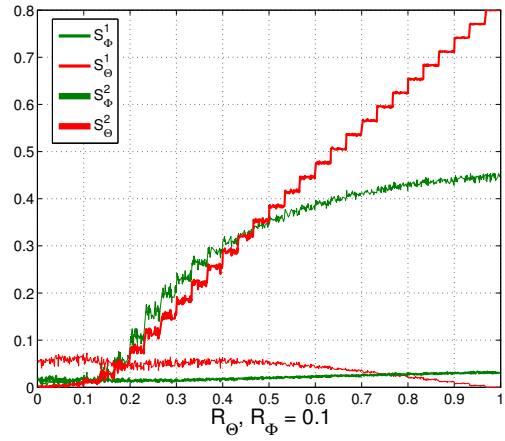


с разреживанием

Рис. 8. Зависимость доли ошибок первого и второго рода при определении нулевых элементов в матрицах Φ, Θ при фиксированной разреженности $R_\Theta = 0.1$.



без разреживания



с разреживанием

Рис. 9. Зависимость доли ошибок первого и второго рода при определении нулевых элементов в матрицах Φ, Θ при фиксированной разреженности $R_\Phi = 0.1$.

Выводы. Произведение $\Phi\Theta$ восстанавливается устойчиво и практически с одинаковой точностью всеми алгоритмами (PLSA-EM и LDA-GS, с разреживанием и без), независимо от степени разреженности исходных данных.

Восстановление матриц Φ и Θ устойчиво только при условии, что справедлива гипотеза разреженности, то есть когда истинные матрицы Φ и Θ, породившие коллекцию документов, разрежены на 80% или более.

Алгоритм постепенного принудительного разреживания не улучшает точность восстановление матриц Φ и Θ по метрике Хеллингера, но уменьшает число ошибок при определении структуры разреженности матриц Φ и Θ.

§13.3 Сравнение PLSA, LDA и SWB (*Потапенко А. А.*)

Цель эксперимента: определить, какая из эвристик важнее — сглаживание, сэмплирование или робастность. Эти три эвристики могут комбинироваться в любых сочетаниях, поэтому сравниваются 8 алгоритмов тематического моделирования.

Исходные данные и условия эксперимента. Эксперимент проводился на коллекциях RuDis и NIPS. Качество алгоритмов оценивалось перплексией обучающей и контрольной коллекции.

При вычислении перплексии на документах d контрольной коллекции D' параметры φ_{wt} и фон π_w оценивались по обучающей коллекции D, параметры θ_{td} и ν_d оценивались по первой половине документа d', параметры шума π_{dw} оценивались для каждой пары (d, w) согласно (4.7). Перплексия вычислялась по вторым половинам d'' контрольных документов.

Сэмплирование (2.6) применялось с параметром $s = n_{dw}$.

Сглаживание (3.3)–(3.4) применялось с параметрами $\alpha_t = 0.5,\ \beta_w = 0.01$.

Робастность применялась с параметрами $\gamma = 0.3,\ \varepsilon = 0.01$.

Число тем $|T| = 100$.

Результаты представлены на рис. 10.

Выводы. Для обеих задач нет существенного различия перплексии между сглаженными моделями (LDA) и несглаженными (PLSA).

Робастные алгоритмы существенно превосходят неробастные и гораздо меньше переобучаются.

Сэмплирование (2.6) сходится быстрее, но в итоге оказывается немного хуже пропорционального распределения (2.2).

Сэмплирование без сглаживания может приводить к увеличению перплексии.

Величина переобучения (разность перплексии на обучающей и контрольной выборке) больше зависит от задачи, чем от алгоритма. Сравнение алгоритмов по перплексии на обучающей выборке приводит к тем же качественным выводам, что и их сравнение по перплексии на контрольной выборке. По всей видимости, для сравнения алгоритмов не нужна столь сложная методика разделения контрольных документов для вычисления перплексии; вполне достаточно вычислять перплексию только на обучающей выборке.

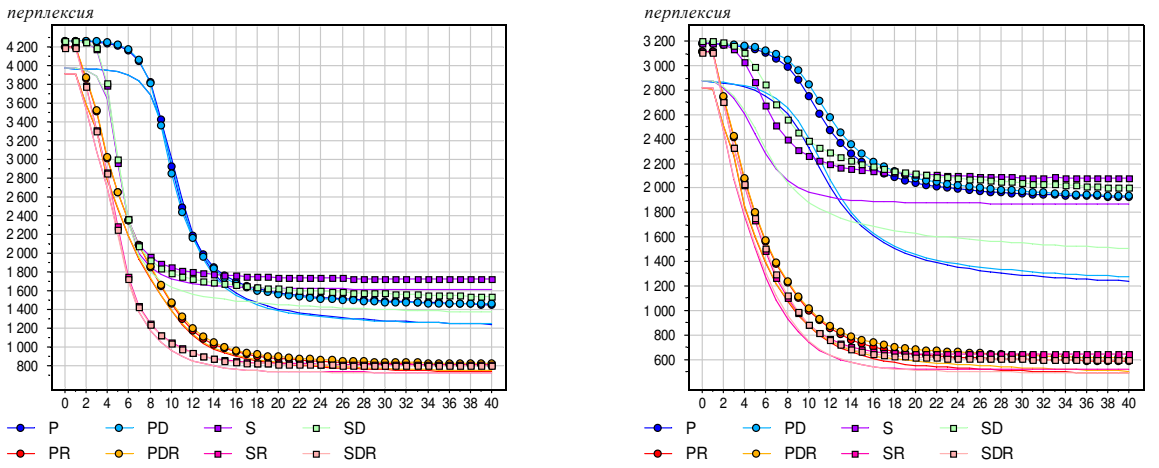


Рис. 10. Зависимость контрольной перплексии от числа итераций для всевозможных сочетаний эвристик: D — сглаживание Дирихле ($\alpha_t = 0.5,\ \beta_w = 0.01$); R — робастность ($\gamma = 0.3,\ \varepsilon = 0.01$); S — сэмплирование ($s = n_{dw}$), P — пропорциональное распределение (2.2); $|T| = 100$. Тонкие кривые без точек — перплексия обучающей выборки.

§13.4 Разреживание матриц Φ и Θ (*Потапенко А. А.*)

Согласно гипотезе разреженности, подавляющее большинство вероятностей $\varphi_{wt} = p(w \mid t)$ и $\theta_{td} = p(t \mid d)$ равны нулю. Однако стандартный EM-алгоритм не позволяет оптимизировать структуру разреженности моделей PLSA и LDA, то есть узнать, какие именно вероятности равны нулю. В PLSA структура разреженности фиксируется начальным приближением. В LDA априорные распределения Дирихле запрещают вероятностям φ_{wt} и θ_{td} и гиперпараметрам β_w и α_t принимать нулевые значения.

Парадокс LDA заключается в том, что при $\beta_w \to 0$ и $\alpha_t \to 0$ распределение Дирихле порождает сильно разреженные распределения φ_t, θ_d, в пределе стремящиеся к вырожденным. В литературе часто встречается утверждение, что модель LDA более разрежена, чем PLSA. Однако при оценивании параметров модели по выборке модель LDA, наоборот, оказывается менее разреженной, чем PLSA. Это с очевидностью следует из сравнения несмещённых частотных оценок PLSA (2.3)–(2.4) со сглаженными оценками LDA (3.4)–(3.3).

Известные подходы к разреживанию LDA требуют введения дополнительных параметров и усложнения EM-алгоритма. В [17] предлагается хранить не сами значения φ_{wt} и θ_{td}, а только их разности с фоновыми распределениями. В [61] предполагается, что каждая тема описывается распределением Дирихле на подмножестве слов, заданном бинарными переменными b_{wt} из распределения Бернулли. Сглаженность и разреженность регулируется независимо параметрами распределения Дирихле и распределения Бернулли. Недостатком данной модели является большое число дополнительных скрытых переменных, которые усложняют обучение. В [27] вводится распределение псевдо-Дирихле, которое строится путём расширения области определения распределения Дирихле и имеет ограниченную плотность, в то время как распределение Дирихле не ограничено в случае $\alpha < 1$, что и приводит к запрету нулевых значений φ_{wt} и θ_{td}.

Целью эксперимента является исследование более простых стратегий *принудительного разреживания*, см. §5.3, стр. 38. Идея разреживания заключается в том, чтобы в конце каждой итерации (полного прохода всей коллекции D) обнулять некоторое количество наименьших значений φ_{wt} и θ_{td}. Теоретические обоснования разреживания приводятся в §5.3.

Считается, что сглаживание в моделях языка необходимо для описания новых и редких слов. Мы предполагаем, что вместо сглаживания можно использовать противоположную эвристику — разреживание, а описание новых и редких слов возложить на робастную модель. Целью эксперимента является проверка этой гипотезы. Для этого сравнивается разреживание робастных и неробастных моделей.

Исходные данные и условия эксперимента. Эксперимент проводился на коллекциях RuDis и NIPS. Качество моделей оценивалось по контрольной перплексии.

Предварительные эксперименты показали, что одновременное обнуление большого числа параметров слишком резко изменяет структуру модели, может приводить к снижению её качества и неравномерному разреживанию, когда некоторые распределения оказываются сильно разреженными, тогда как другие почти не разреживаются. Поэтому предлагается разреживать матрицы Φ и Θ постепенно, придерживаясь одной из следующих стратегий.

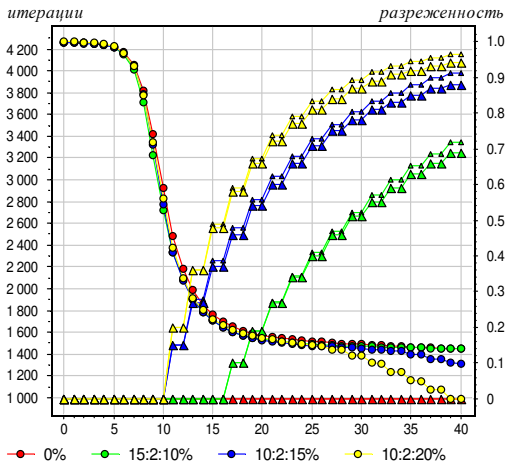
Простая стратегия: в каждом из распределений φ_t, θ_d обнуляется заданная доля r наименьших *ненулевых* значений. После обнуления производится перенормировка распределений. Число обнуляемых значений сокращается от итерации к итерации, поскольку доля берётся от числа ненулевых значений. Обнуления прекращаются, когда в распределении остаётся $\lfloor r^{-1} \rfloor$ ненулевых значений. Недостатком этой стратегии является стремление к выравниванию доли ненулевых значений во всех распределениях, что представляется довольно странным ограничением.

Сложная стратегия устраняет этот недостаток. В каждом из распределений φ_t, θ_d обнуляется максимальное число наименьших значений, так, чтобы оно не превышало $r|W|$ и $r|T|$ соответственно, и сумма обнуляемых значений не превышала заданного порога S_φ или S_θ для распределений φ_t или θ_d соответственно. В экспериментах эта стратегия показала лучшие результаты. Задание параметров S_φ, S_θ эквивалентно заданию коэффициентов регуляризации τ_φ, τ_θ в (5.5).

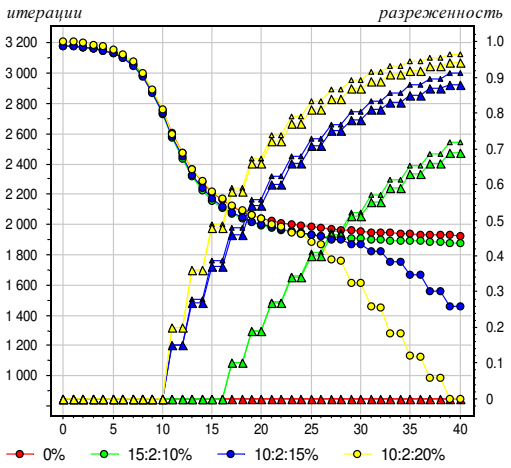
Начинать обнуления малых вероятностей можно и не дожидаясь сходимости. После нескольких первых проходов коллекции становится ясно, что самые малые вероятности останутся малыми на всех последующих итерациях. Поэтому разреживания включаются, начиная с итерации i_0, и производятся не на каждой итерации, чтобы модель успевала восстановить адекватность. В экспериментах [2] разреживания включались на итерациях с номерами $i = i_0 + k\delta$, $k = 1, 2, \ldots$, где i_0 и δ — параметры стратегии разреживания.

Под *агрессивным разреживанием* понимается уменьшение δ до 1 или уменьшение i_0 до 1 или применение сложной стратегии, когда число обнуляемых значений не уменьшается с итерациями.

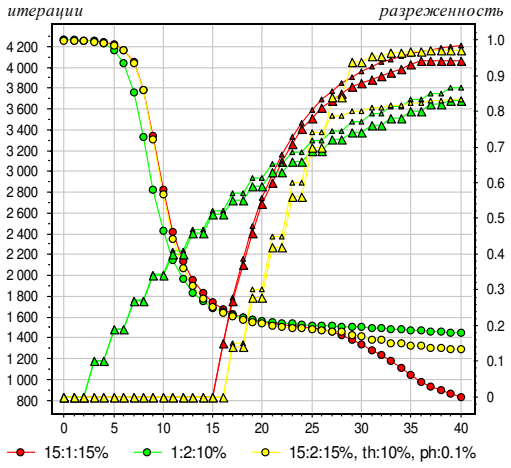
В каждом распределении $p(w \mid t)$ и $p(t \mid d)$ должно оставаться хотя бы одно ненулевое значение. Если в результате разреживания все вероятности $p(t \mid d, w)$ оцениваются как нулевые, то термин w считается нетематическим в документе d. Поэтому разреживание применяется совместно с робастной моделью (4.1), либо с упрощённой робастной моделью (4.8).



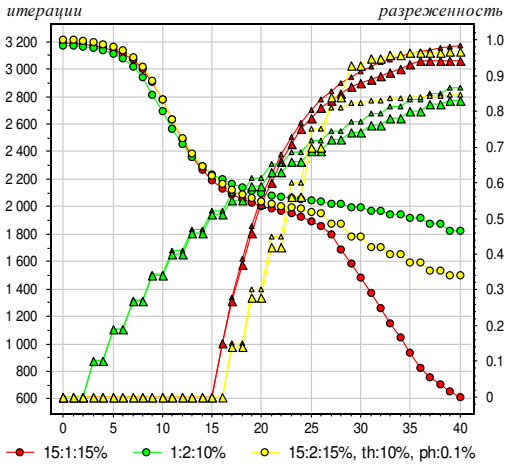
RuDis, разреживание через 2 итерации



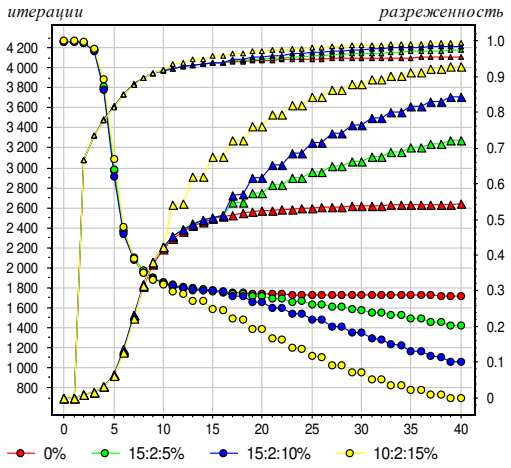
NIPS, разреживание через 2 итерации



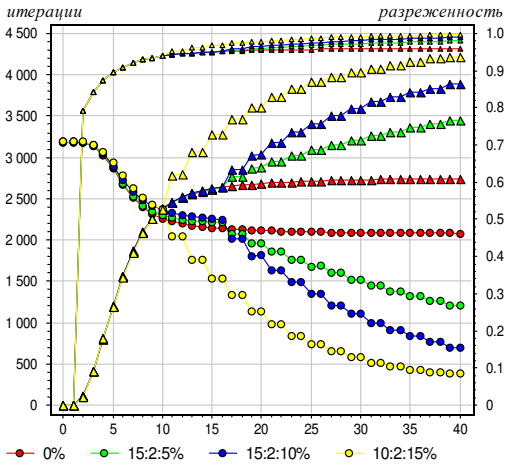
RuDis, агрессивное разреживание



NIPS, агрессивное разреживание

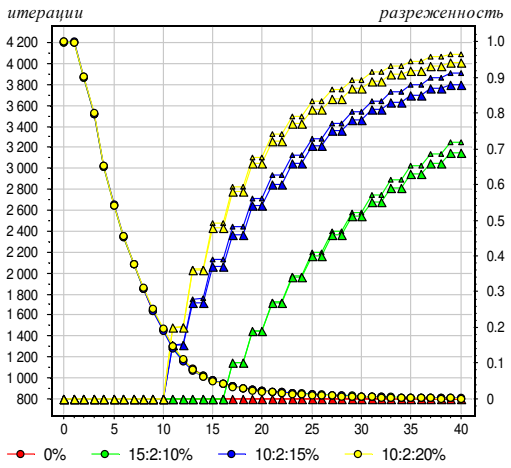


RuDis, SEM, через 2 итерации

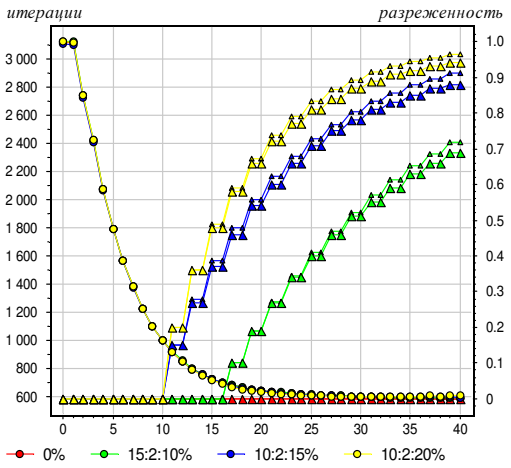


NIPS, SEM, через 2 итерации

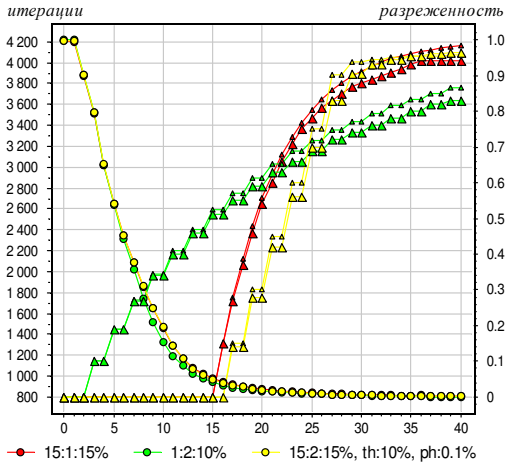
Рис. 11. Зависимость перплексии (○) и разреженности матриц Φ (△) и Θ (△) от числа итераций для рационального и стохастического EM-алгоритма при различных параметрах разреживания, обозначаемых $i_0{:}\delta{:}r$, th:S_θ, ph:S_φ. Число тем $|T| = 100$.



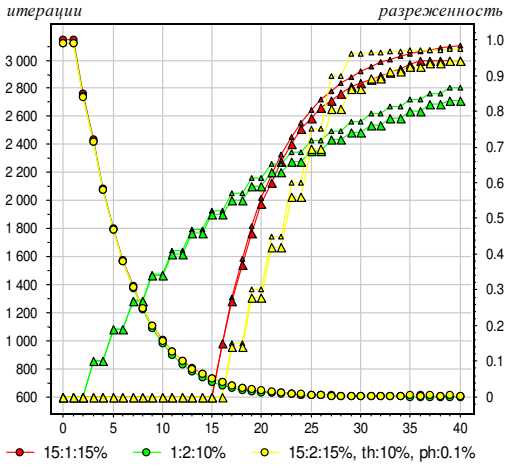
RuDis, разреживание через 2 итерации



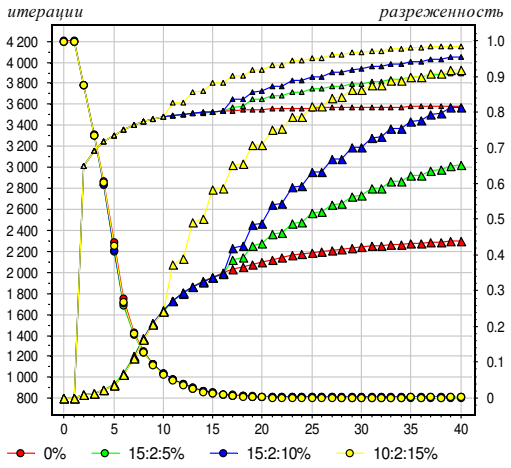
NIPS, разреживание через 2 итерации



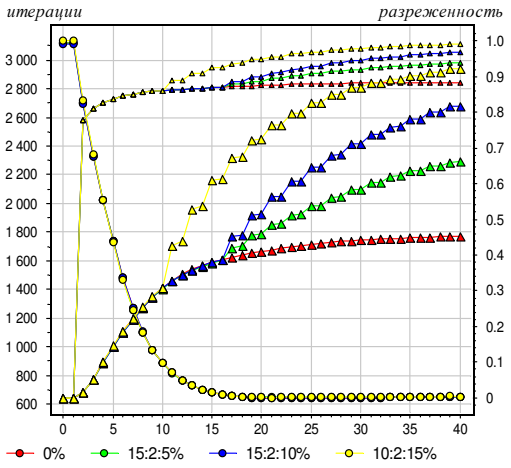
RuDis, агрессивное разреживание



NIPS, агрессивное разреживание

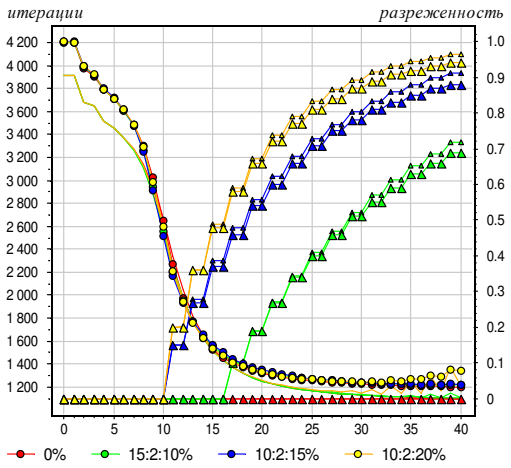


RuDis, SEM, через 2 итерации

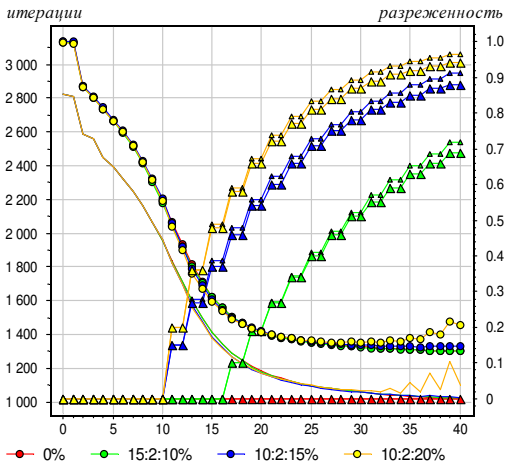


NIPS, SEM, через 2 итерации

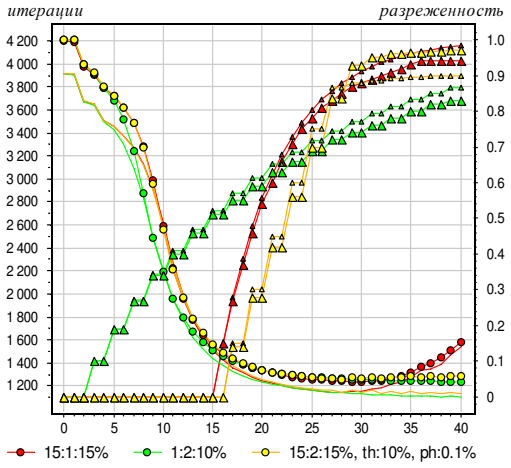
Рис. 12. Зависимость перплексии (○) и разреженности матриц Φ (△) и Θ (△) от числа итераций для рационального и стохастического робастного EM-алгоритма с параметрами робастности $\gamma = 0.3$, $\varepsilon = 0.01$ и параметрами разреживания $i_0{:}\delta{:}r$, $\mathsf{th}{:}S_\theta$, $\mathsf{ph}{:}S_\varphi$. Число тем $|T| = 100$.



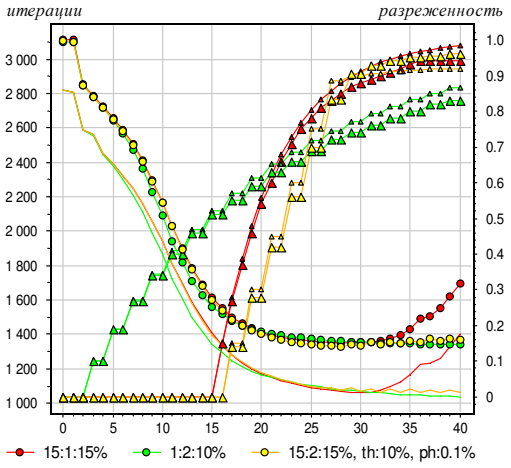
RuDis, разреживание через 2 итерации



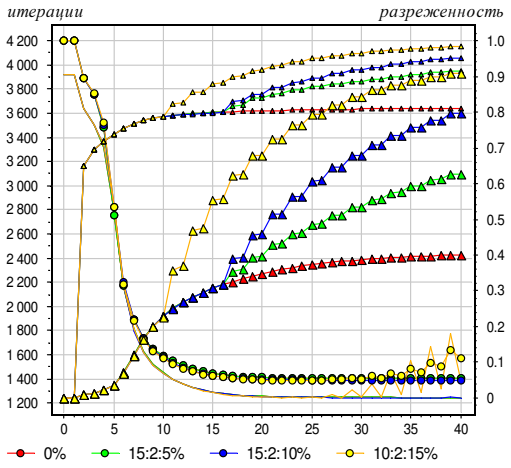
NIPS, разреживание через 2 итерации



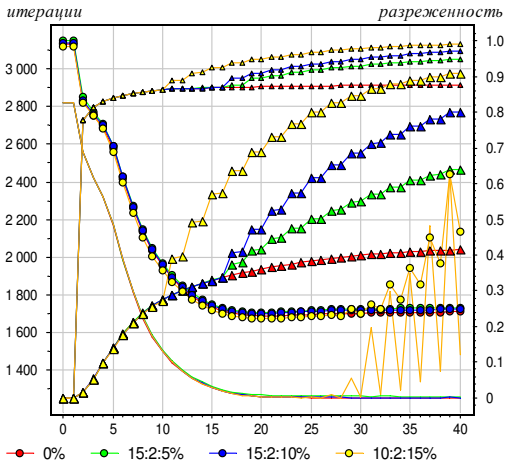
RuDis, агрессивное разреживание



NIPS, агрессивное разреживание



RuDis, SEM, через 2 итерации



NIPS, SEM, через 2 итерации

Рис. 13. Зависимость перплексии (○) и разреженности матриц Φ (△) и Θ (△) от числа итераций для рационального и стохастического робастного EM-алгоритма при малой априорной вероятности шума $\gamma = 0.01$, $\varepsilon = 0.01$, с параметрами разреживания $i_0{:}\delta{:}r$, th:S_θ, ph:S_φ. Число тем $|T| = 100$.

Результаты экспериментов [2]. На рис. 11 показаны зависимости контрольной перплексии и разреженности матриц Φ и Θ от числа итераций при различных стратегиях разреживания. На рис. 12 показаны аналогичные зависимости для робастных моделей с априорной долей шума $\gamma = 0.3$, на рис. 13 — с априорной долей шума $\gamma = 0.01$.

Наименьшая перплексия и одновременно наибольшая разреженность матрицы Φ до 99.4% для RuDis и 99.6% для NIPS достигается при использовании упрощённого робастного стохастического EM-алгоритма с параметрами $r = 0.15$, $S_\theta = 0.1$, $S_\varphi = 0.001$, i_0 от 15 до 20, $\delta = 1$ или 2, рис. 11. Разреживание распределений φ_t свыше 99% при числе тем $T = 100$ означает, что каждый термин в среднем относится только к одной теме.

В робастных алгоритмах с шумом и фоном (SWB) разреживание почти не влияет на перплексию и позволяет достигать сопоставимой разреженности, рис. 12.

В неробастных алгоритмах агрессивное разреживание может приводить к обратному результату — снижению разреженности φ_t до 80%.

При недостаточном априорном уровне шума $\gamma = 0.01$ агрессивное разреживание может приводить к расходимости EM-алгоритма, рис. 13. Тонкие кривые без точек, проходящие чуть ниже кривых контрольной перплексии, соответствуют перплексии на обучающей выборке. Они показывают, что расходимость возникает синхронно на контроле и обучении. Поэтому её легко обнаружить на стадии обучения, непосредственно во время итераций EM-алгоритма.

Для упрощённой робастной модели расходимость ни разу не наблюдалась.

Эвристика разреживания плохо совместима со сглаживанием и применяется только к PLSA, то есть при $\beta_w = 0$, $\alpha_t = 0$.

Выводы. Робастные модели с разреживанием не нуждаются в сглаживании. Для практического применения рекомендуется упрощённый робастный EM-алгоритм с достаточно высоким априорным уровнем шума γ и агрессивным разреживанием.

§13.5 Разреживание распределений тем $p(t \mid d, w)$ (*Потапенко А. А.*)

Цель эксперимента. Недостатком Алгоритма 2.2 является необходимость хранить массив значений $n_{dwt} = n_{dw} p(t \mid d, w)$, $t \in T$, для каждого термина (d, w). Расход памяти объёма $O(n|T|)$ может оказаться неприемлемым даже при небольшом числе тем. С другой стороны, согласно гипотезе разреженности, этот массив должен состоять преимущественно из нулей. Сэмплирование решает данную проблему, однако остаётся вопрос, не вносит ли генератор случайных чисел дополнительный шум, и не существует ли более простого способа разреживания распределений $p(t \mid d, w)$.

Стратегия *максимального разреживания* распределений $p(t \mid d, w)$ представляется наиболее естественной. Для каждого термина (d, w) остаются только s тем с наибольшими значениями n_{dwt}, вероятности остальных темы обнуляются.

Целью эксперимента является сравнение различных стратегий разреживания.

Исходные данные и условия эксперимента. Эксперимент проводился на коллекциях RuDis и NIPS. Качество моделей оценивалось по контрольной перплексии.

Результаты. Стратегия максимального разреживания приводит к накоплению систематической ошибки и расходимости ЕМ-алгоритма, рис. 14. На первых же итерациях возникает сильная (свыше 90%) разреженность распределений $\varphi_{wt} = p(w \mid t)$, которые к этому моменту ещё не сошлись. Значения φ_{wt}, оказавшиеся равными нулю, далее так и остаются нулевыми.

Включение разреживания с 10-й итерации даёт лучшие результаты, но также может приводит к расходимости (средний ряд графиков на рис. 14). При этом наблюдается интересный положительный эффект: сразу после разреживания перплексия улучшается скачком, даже при малых $s = 1, 2$.

Эвристика сглаживания даёт ещё лучшие результаты (нижний ряд графиков на рис. 14). Расходимость не возникает, но качество получаемой модели всё же хуже, чем при $s = |T|$, то есть когда разреживание не применяется. Снова наблюдается эффект скачкообразного падения перплексии сразу после разреживания.

Максимальное разреживание при $s = 1$ можно интерпретировать как применение оптимального байесовского решающего правила в задаче классификации терминов в документах (d, w) на $|T|$ классов-тем. Оно даёт адекватный результат только после достижения сходимости в ЕМ-алгоритме, рис. 15. При этом перплексия незначительно ухудшается, разреженность φ_t и θ_d достигает 0.8–0.9.

Менее требовательной к настройке параметров оказалась стратегия *постепенного разреживания*, когда в каждом распределении $p(t \mid d, w)$ обнуляется заданная доля r наименьших ненулевых значений и производится перенормировка. При $r \leqslant 0.2$ расходимость не возникает, и финальная перплексия мало отличается от случая $r = 0$, см. рис. 16. При включении разреживаний, начиная с 10-й итерации, темп разреживания r можно увеличить (средний ряд графиков на рис. 16). При этом постепенно увеличивается разреженность распределений θ_{td} (до 0.5 и выше) и распределений φ_{wt} (немного ниже 0.5).

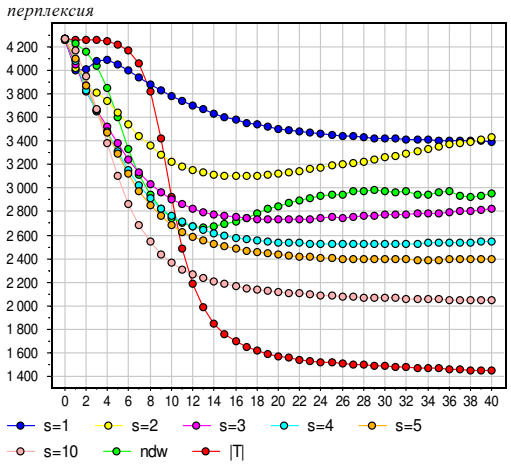
Робастные алгоритмы более устойчивы к постепенному разреживанию распределений $p(t \mid d, w)$. У них расходимость не наблюдалась, темп разреживания можно увеличивать до $r = 0.7$, (нижний ряд графиков на рис. 16). При этом разреженность φ_{wt} достигает почти 0.9, разреженность θ_{td} достигает 0.7.

Выводы. Стратегии максимального и постепенного разреживания требуют хранения слабо разреженного массива n_{dwt} или $p(t \mid d, w)$, хотя бы на ранних итерациях. Этим они уступают стохастическому ЕМ-алгоритму.

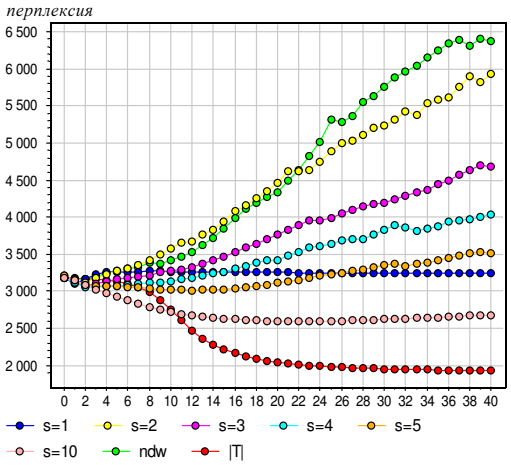
Максимальное разреживание может использоваться по окончании итераций для жёсткого присвоения тем каждому термину в каждом документе (d, w), практически без ухудшения перплексии.

Максимальное разреживание даёт резкое улучшение перплексии непосредственно после его применения. Возникает гипотеза, что максимальное разреживание на отдельных промежуточных итерациях ЕМ-алгоритма может улучшать его сходимость. Возможно также, что непосредственно после итераций разреживания придётся применять сглаживание.

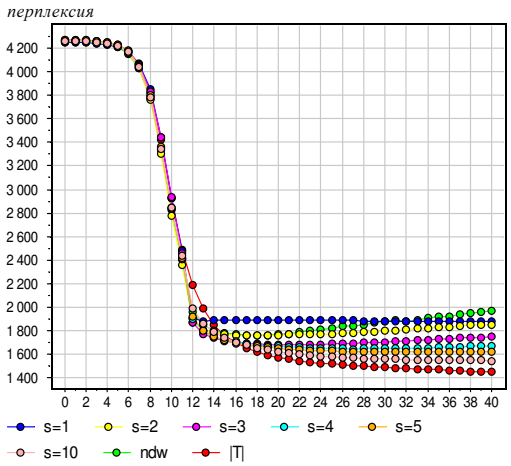
§13.6 Экономное сэмплирование (*Потапенко А. А.*)



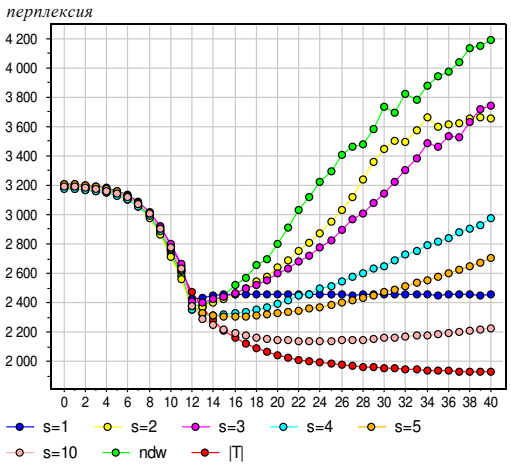
RuDis, разреживание с 1-й итерации



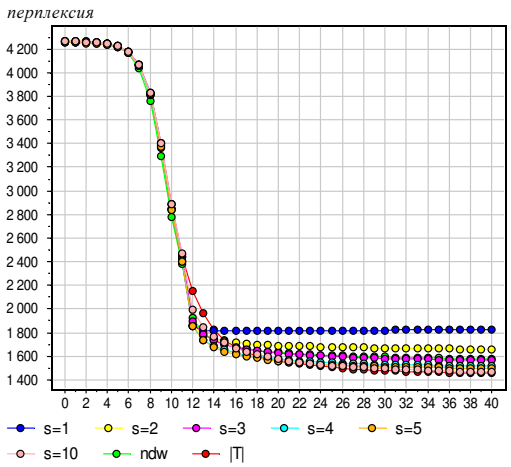
NIPS, разреживание с 1-й итерации



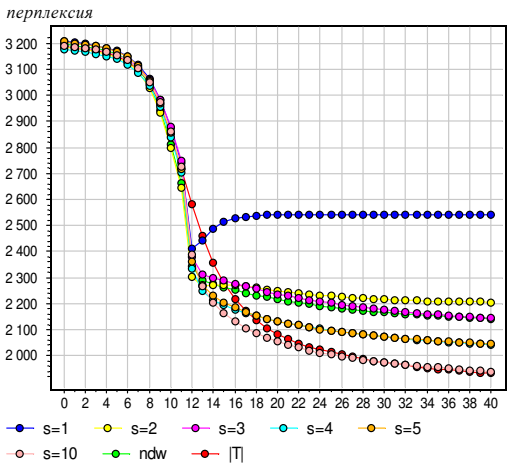
RuDis, разреживание с 10-й итерации



NIPS, разреживание с 10-й итерации

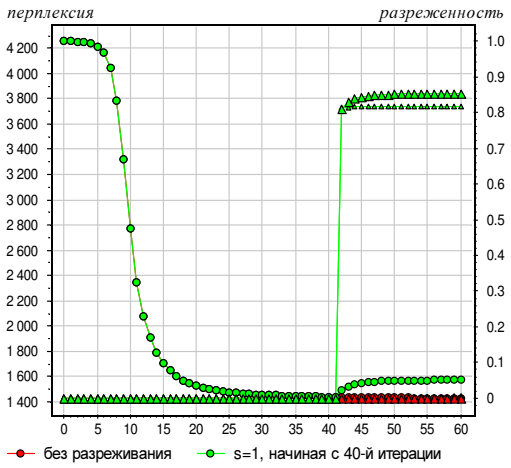


RuDis, с 10-й итерации, сглаживание LDA

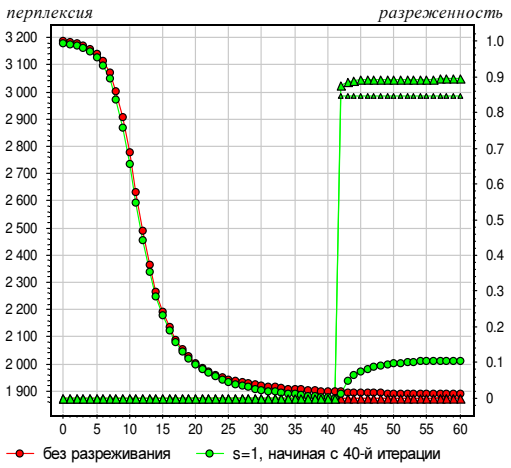


NIPS, с 10-й итерации, сглаживание LDA

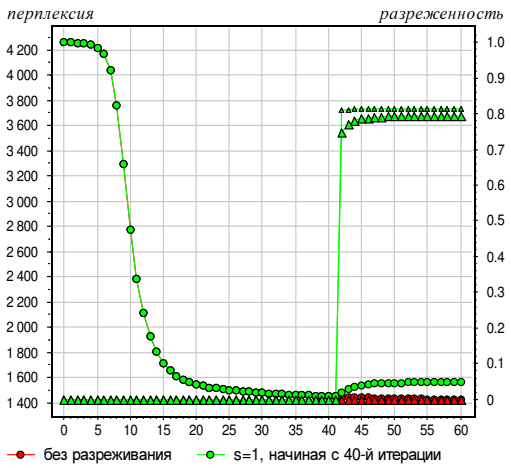
Рис. 14. Зависимость перплексии от числа итераций в рациональном EM-алгоритме при максимальном разреживании $p(t\,|\,d,w)$. Параметр разреживания: $s = 1, 2, 3, 4, 5, 10, n_{dw}$, при $s = |T|$ разреживания нет. Параметры сглаживания: $\alpha_t = 0.5,\ \beta_w = 0.01$. Число тем $|T| = 100$.



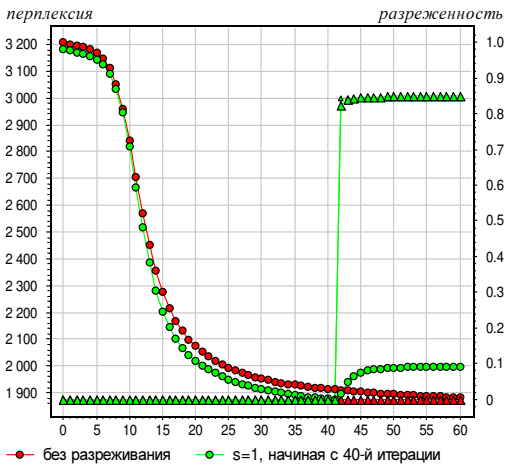
RuDis, разреживание с 40-й итерации



NIPS, разреживание с 40-й итерации

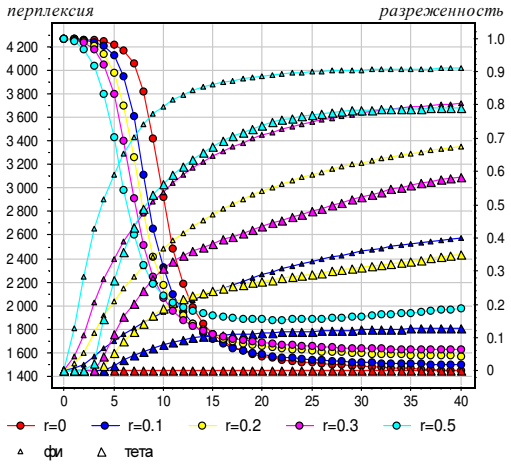


RuDis, с 40-й итерации, сглаживание LDA

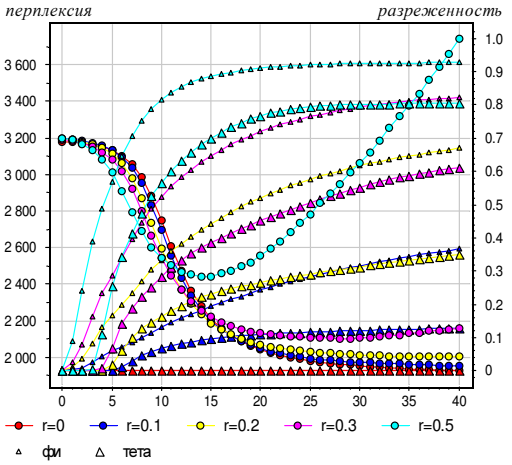


NIPS, с 40-й итерации, сглаживание LDA

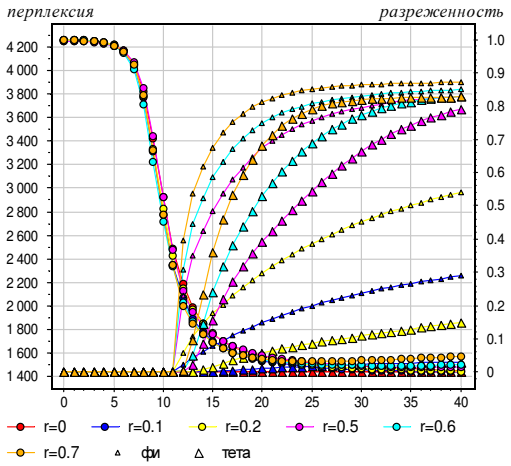
Рис. 15. Зависимость перплексии (○) и разреженности матриц Φ (△) и Θ (△) от числа итераций при максимальном разреживании $s = 1$ после достижения сходимости в рациональном EM-алгоритме. Параметры сглаживания: $\alpha_t = 0.5$, $\beta_w = 0.01$. Число тем $|T| = 100$.



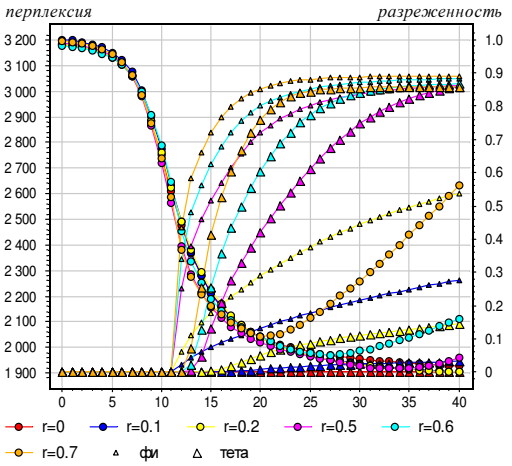
RuDis, разреживание с 1-й итерации



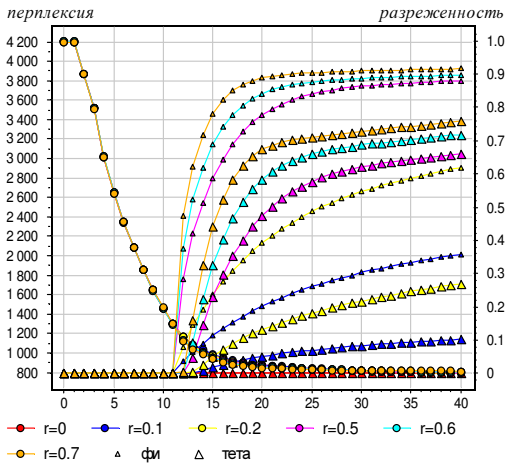
NIPS, разреживание с 1-й итерации



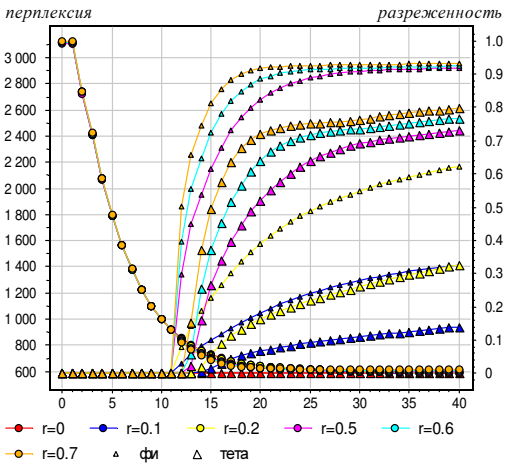
RuDis, разреживание с 10-й итерации



NIPS, разреживание с 10-й итерации



RuDis, с 10-й итерации, робастный PLSA



NIPS, с 10-й итерации, робастный PLSA

Рис. 16. Зависимость перплексии (○) и разреженности матриц Φ (△) и Θ (△) от числа итераций в рациональном EM-алгоритме при постепенном разреживании $p(t \mid d, w)$ с параметром r от 0 до 0.7. Параметры робастности: $\varepsilon = 0.01$, $\gamma = 0.3$. Число тем $|T| = 100$.

Список литературы

[1] *Воронцов К. В., Потапенко А. А.* Регуляризация, робастность и разреженность вероятностных тематических моделей // *Компьютерные исследования и моделирование.* — 2012. — Т. 4, № 4. — С. 693–706.

[2] *Воронцов К. В., Потапенко А. А.* Модификации EM-алгоритма для вероятностного тематического моделирования // *Машинное обучение и анализ данных.* — 2013 (в печати).

[3] *Гиляревский Р. С., Шапкин А. В., Белоозеров В. Н.* Рубрикатор как инструмент информационной навигации. — СПб.: Профессия, 2008. — 352 с.

[4] *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска. — Издательство МГУ имени М. В. Ломоносова, 2011.

[5] *Маннинг К. Д., Рагхаван П., Шютце Х.* Введение в информационный поиск. — Вильямс, 2011.

[6] *Павлов А. С., Добров Б. В.* Метод обнаружения массово порожденных неестественных текстов на основе анализа тематической структуры // *Вычислительные методы и программирование: новые вычислительные технологии.* — 2011. — Т. 12. — С. 58–72.

[7] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1986.

[8] *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. — 2009.

[9] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // *Journal of Machine Learning Research.* — 2003. — Vol. 3. — Pp. 993–1022.

[10] *Buntine W. L.* Estimating likelihoods for topic models // 1st Asian Conference on Machine Learning: Advances in Machine Learning. — 2009. — Pp. 51–64.

[11] *Celeux G., Chauveau D., Diebolt J.* On stochastic versions of the EM algorithm: Tech. Rep. RR-2514: INRIA, 1995.

[12] *Chemudugunta C., Smyth P., Steyvers M.* Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems. — MIT Press, 2007. — Vol. 19. — Pp. 241–248.

[13] *Cressie N., Read T. R. C.* Multinomial goodness-of-fit tests // *Journal of the Royal Statistical Society, Series B.* — 1984. — Vol. 46, no. 3. — Pp. 440–464.

[14] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // *Frontiers of Computer Science in China.* — 2010. — Vol. 4, no. 2. — Pp. 280–301.

[15] *Dempster A. P., Laird N. M., Rubin D. B.* Maximum likelihood from incomplete data via the EM algorithm // *J. of the Royal Statistical Society, Series B.* — 1977. — no. 34. — Pp. 1–38.

[16] *Dietz L., Bickel S., Scheffer T.* Unsupervised prediction of citation influences // Proceedings of the 24th international conference on Machine learning. — ICML '07. — New York, NY, USA: ACM, 2007. — Pp. 233–240.

[17] *Eisenstein J., Ahmed A., Xing E. P.* Sparse additive generative models of text // ICML'11. — 2011. — Pp. 1041–1048.

[18] *Feng Y., Lapata M.* Topic models for image annotation and text illustration // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Association for Computational Linguistics, 2010. — Pp. 831–839.

[19] *Friedman J. H., Hastie T., Tibshirani R.* Regularization paths for generalized linear models via coordinate descent // *Journal of Statistical Software.* — 2010. — Vol. 33, no. 1. — Pp. 1–22.

[20] *Grün B., Hornik K.* Topicmodels: An R package for fitting topic models // *Journal of Statistical Software.* — 2011. — Vol. 40, no. 13. — Pp. 1–30.

[21] *Hoffman M. D., Blei D. M., Bach F. R.* Online learning for latent dirichlet allocation // NIPS. — Curran Associates, Inc., 2010. — Pp. 856–864.

[22] *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. — Pp. 50–57.

[23] *Ito K., Jin B., Takeuchi T.* Multi-parameter Tikhonov regularization // *The Computing Research Repository (CoRR).* — 2011. — Vol. abs/1102.1173.

[24] *Kataria S., Mitra P., Caragea C., Giles C. L.* Context sensitive topic models for author influence in document networks // Proceedings of the Twenty-Second international joint conference on Artificial Intelligence — Volume 3. — IJCAI'11. — AAAI Press, 2011. — Pp. 2274–2280.

[25] *Kim S.-H., Choi H., Lee S.* Estimate-based goodness-of-fit test for large sparse multinomial distributions // *Computational Statistics and Data Analysis.* — 2009. — Vol. 53, no. 4. — Pp. 1122 – 1131.

[26] *Krestel R., Fankhauser P., Nejdl W.* Latent dirichlet allocation for tag recommendation // Proceedings of the third ACM conference on Recommender systems. — ACM, 2009. — Pp. 61–68.

[27] *Larsson M. O., Ugander J.* A concave regularization technique for sparse mixture models // Advances in Neural Information Processing Systems 24 / Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K. Weinberger. — 2011. — Pp. 1890–1898.

[28] *Lau J. H., Baldwin T., Newman D.* On collocations and topic models // *ACM - Transactions on Speech and Language Processing.* — 2013. — Vol. 10, no. 3. — Pp. 10:1–10:14.

[29] *LeCun Y., Denker J., Solla S., Howard R. E., Jackel L. D.* Optimal brain damage // Advances in Neural Information Processing Systems II / Ed. by D. S. Touretzky. — San Mateo, CA: Morgan Kauffman, 1990.

[30] *Li X.-X., Sun C.-B., Lu P., Wang X.-J., Zhong Y.-X.* Simultaneous image classification and annotation based on probabilistic model // *The Journal of China Universities of Posts and Telecommunications.* — 2012. — Vol. 19, no. 2. — Pp. 107–115.

[31] *Lu Y., Mei Q., Zhai C.* Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA // *Information Retrieval.* — 2011. — Vol. 14, no. 2. — Pp. 178–203.

[32] *Mann G. S., McCallum A.* Simple, robust, scalable semi-supervised learning via expectation regularization // Proceedings of the 24th international conference on Machine learning. — ICML '07. — New York, NY, USA: ACM, 2007. — Pp. 593–600.

[33] *Masada T., Kiyasu S., Miyahara S.* Comparing LDA with pLSI as a dimensionality reduction method in document clustering // Proceedings of the 3rd International Conference on Large-scale knowledge resources: construction and application. — LKR'08. — Springer-Verlag, 2008. — Pp. 13–26.

[34] *Mimno D., Blei D.* Bayesian checking for topic models // 11th Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics, 2011. — Pp. 227–237.

[35] *Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A.* Optimizing semantic coherence in topic models // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — EMNLP '11. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. — Pp. 262–272.

[36] *Minka T. P.* Estimating a dirichlet distribution: Tech. rep.: 2000 (revised 2003, 2009, 2012).

[37] *Munkres J.* Algorithms for the assignment and transportation problems // *Journal of the Society for Industrial and Applied Mathematics.* — 1957. — Vol. 5, no. 1. — Pp. 32–38.

[38] *Neal R. M., Hinton G. E.* A view of the EM algorithm that justifies incremental, sparse, and other variants // Learning in graphical models / Ed. by M. I. Jordan. — Cambridge, MA, USA: MIT Press, 1999. — Pp. 355–368.

[39] *Newman D., Bonilla E. V., Buntine W. L.* Improving topic coherence with regularized topic models // Advances in Neural Information Processing Systems 24 / Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K. Weinberger. — 2011. — Pp. 496–504.

[40] *Newman D., Karimi S., Cavedon L.* External evaluation of topic models // Australasian Document Computing Symposium. — December 2009. — Pp. 11–18.

[41] *Newman D., Lau J. H., Grieser K., Baldwin T.* Automatic evaluation of topic coherence // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — HLT '10. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. — Pp. 100–108.

[42] *Newman D., Noh Y., Talley E., Karimi S., Baldwin T.* Evaluating topic models for digital libraries // Proceedings of the 10th annual Joint Conference on Digital libraries. — JCDL '10. — New York, NY, USA: ACM, 2010. — Pp. 215–224.

[43] *Nigam K., McCallum A. K., Thrun S., Mitchell T.* Text classification from labeled and unlabeled documents using EM // *Machine Learning.* — 2000. — Vol. 39, no. 2-3. — Pp. 103–134.

[44] *Pecina P., Schlesinger P.* Combining association measures for collocation extraction // Proceedings of the COLING/ACL on Main conference poster sessions. — Association for Computational Linguistics, 2006. — Pp. 651–658.

[45] *Ramage D., Hall D., Nallapati R., Manning C. D.* Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1. — EMNLP '09. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. — Pp. 248–256.

[46] *Read T., Cressie N.* Goodness-of-Fit Statistics for Discrete Mutivariate Data. — Springer, New York, 1988.

[47] *Rosen-Zvi M., Griffiths T., Steyvers M., Smyth P.* The author-topic model for authors and documents // Proceedings of the 20th conference on Uncertainty in artificial intelligence. — UAI '04. — Arlington, Virginia, United States: AUAI Press, 2004. — Pp. 487–494.

[48] *Rubin T. N., Chambers A., Smyth P., Steyvers M.* Statistical topic models for multi-label document classification // *Machine Learning.* — 2012. — Vol. 88, no. 1-2. — Pp. 157–208.

[49] *Sebastiani F.* Machine learning in automated text categorization // *ACM Computing Surveys.* — 2002. — Vol. 34, no. 1. — Pp. 1–47.

[50] *Steyvers M., Griffiths T.* Finding scientific topics // *Proceedings of the National Academy of Sciences.* — 2004. — Vol. 101, no. Suppl. 1. — Pp. 5228–5235.

[51] *Steyvers M., Griffiths T.* Probabilistic Topic Models // Handbook of Latent Semantic Analysis / Ed. by T. Landauer, D. Mcnamara, S. Dennis, W. Kintsch. — Lawrence Erlbaum Associates, 2007.

[52] *Tan Y., Ou Z.* Topic-weak-correlated latent dirichlet allocation // 7th International Symposium Chinese Spoken Language Processing (ISCSLP). — 2010. — Pp. 224–228.

[53] *Taneichi N., Sekiya Y., Imai H.* Improvements of goodness-of-fit statistics for sparse multinomials based on normalizing transformations // *Annals of the Institute of Statistical Mathematics.* — 2003. — Vol. 55. — Pp. 831–848.

[54] *Teh Y. W., Newman D., Welling M.* A collapsed variational bayesian inference algorithm for latent dirichlet allocation // NIPS. — 2006. — Pp. 1353–1360.

[55] TextFlow: Towards better understanding of evolving topics in text. / W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, X. Tong // *IEEE transactions on visualization and computer graphics.* — 2011. — Vol. 17, no. 12. — Pp. 2412–2421.

[56] *Varadarajan J., Emonet R., Odobez J.-M.* A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions. — 2010.

[57] *Vulić I., Smet W., Moens M.-F.* Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora // *Information Retrieval.* — 2012. — Pp. 1–38.

[58] *Wallach H.* Structured Topic Models for Language: Ph.D. thesis / Newnham College, University of Cambridge. — 2008.

[59] *Wallach H., Mimno D., McCallum A.* Rethinking LDA: Why priors matter // Advances in Neural Information Processing Systems 22 / Ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta. — 2009. — Pp. 1973–1981.

[60] *Wallach H., Murray I., Salakhutdinov R., Mimno D.* Evaluation methods for topic models // 26th International Conference on Machine Learning, Montreal, Canada. — 2009. — Pp. 1105–1112.

[61] *Wang C., Blei D. M.* Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process // NIPS. — Curran Associates, Inc., 2009. — Pp. 1982–1989.

[62] *Wang Y.* Distributed Gibbs sampling of latent dirichlet allocation: The gritty details. — 2008.

[63] *Wu Y., Ding Y., Wang X., Xu J.* A comparative study of topic models for topic clustering of chinese web news // Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on. — Vol. 5. — july 2010. — Pp. 236–240.

[64] *Yeh J.-h., Wu M.-l.* Recommendation based on latent topics and social network analysis // Proceedings of the 2010 Second International Conference on Computer Engineering and Applications. — Vol. 1. — IEEE Computer Society, 2010. — Pp. 209–213.

[65] *Yi X., Allan J.* A comparative study of utilizing topic models for information retrieval // Advances in Information Retrieval. — Springer Berlin Heidelberg, 2009. — Vol. 5478 of *Lecture Notes in Computer Science*. — Pp. 29–41.

[66] *Zavitsanos E., Paliouras G., Vouros G. A.* Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes // *Journal of Machine Learning Research*. — 2011. — Vol. 12. — Pp. 2749–2775.

[67] *Zelterman D.* Goodness-of-fit tests for large sparse multinomial distributions // *Journal of the American Statistical Association*. — 1987. — Vol. 398, no. 82. — Pp. 624–629.

[68] *Zhang J., Song Y., Zhang C., Liu S.* Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2010. — Pp. 1079–1088.

[69] *Zhang Z., Iria J., Brewster C., Ciravegna F.* A comparative evaluation of term recognition algorithms // Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08). — 2008.

[70] *Zou H., Hastie T.* Regularization and variable selection via the elastic net // *Journal of the Royal Statistical Society B*. — 2005. — Vol. 67. — Pp. 301–320.