Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Калина Екатерина, Зенкова Наталья, Балагуров Владимир

Тематическое моделирование

Конспект

Санкт-Петербург
2019

Оглавление

1. Введение

Тематическое моделирование (topic modeling) — одно из современных приложений машинного обучения к анализу текстов. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему.

Вероятностная тематическая модель (ВТМ) 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 ВТМ 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.

Некоторые приложения тематического моделирования:

- разведочный поиск в электронных библиотеках;
- поиск тематического контента в соцсетях;
- детектирование и трекинг новостных сюжетов;
- мультимодальный поиск текстов и изображений;
- анализ банковских транзакционных данных.

2. Вероятностная модель коллекции документов

2.1. Постановка задачи

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 $(w_1, ..., w_{nd})$ из словаря W. Термин может повторяться в документе много раз.

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 — это i.i.d выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$, заданная 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.*

Таким образом,

- Тема — условное распределение на множестве терминов, $p(w|t)$ — вероятность термина w в теме t
- Тематический профиль документа — условное распределение $p(t|d)$ — вероятность темы t в документе d

Задача тематического моделирования. Построить тематическую модель коллекции документов D — значит найти множество тем T, распределения $p(w|t)$ для всех тем $t \in T$ и распределения $p(t|d)$ для всех документов $d \in D$.

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

2.2. Гипотезы и предположения

Гипотеза независимости. 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.

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

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

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

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

2.3. Предварительная обработка документов

Мотивация: К предварительной обработке текстов прибегают для упрощения модели.

- **Лемматизация** — это приведение каждого слова в документе к его нормальной форме. 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.

  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.

- **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

- **Уменьшение словаря**:

  - $1000,\ 5,\ 23 \to \$number$; $(5+3),\ \frac{1}{2}ww^T + C \to \$formula$
  - **Отбрасывание стоп-слов**. Удаление слов (предлогов, союзов, числительных, местоимений, вводных слов, некоторых глаголов, прилагательных и наречий.) 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**. Удаление слов, встречающихся в длинном документе слишком редко, например, только один раз (они не характеризуют тематику документа). *Для коллекций коротких новостных сообщений 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 *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**. Понятие «термина» может изменяться в зависимости от целей построения тематической модели и таких особенностей задачи, как язык документов, средняя длина документов, тематика коллекции.

2.4. Вероятностная модель порождения данных

Согласно определению условной вероятности, формуле полной вероятности и гипотезе условной независимости:

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

Если распределения тем в каждом документе $p(t|d)$ и терминов в каждой теме $p(w|t)$ известны, то тематическая модель описывает процесс порождения коллекции D.

Ниже представлены Алгоритм (1) и иллюстрация (Рис.1).

Алгоритм 1. Вероятностная модель порождения коллекции документов.

Input: распределения $p(w|t)$, $p(t|d)$;
Output: выборка пар $(d_i, w_i), i = 1, ..., n$;

for all $d \in D$ **do**
  задать длину n_d документа d;
  for all $i = 1, ..., n_d$ **do**
    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»;
  end for
end for
=0

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

Дано: коллекция текстовых документов

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

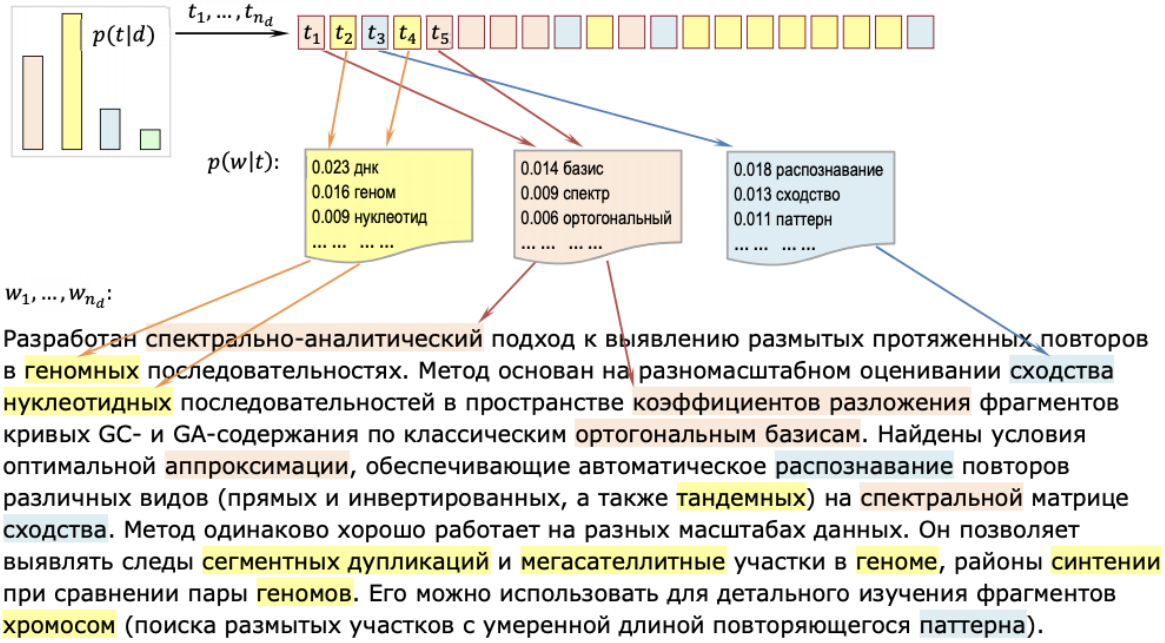

Рис. 1. Процесс порождения текстового документа вероятностной тематической моделью

Найти: параметры тематической модели $p(w|d) = \sum_{t \in T} wttd$

- $_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
- $_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:

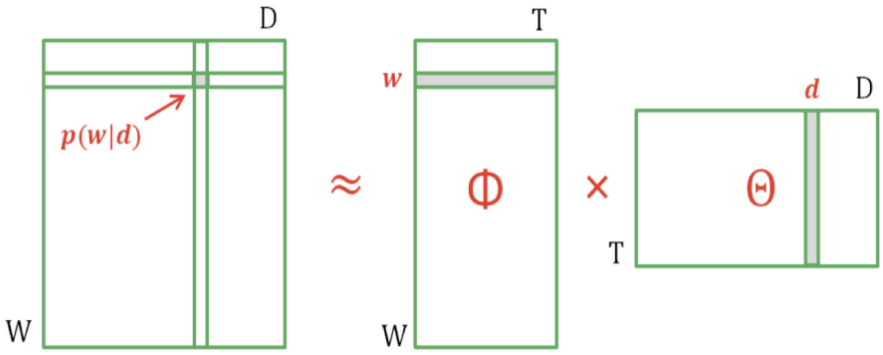

Рис. 2. Задача стохастического матричного разложения

Если $\mathbf{\Phi}$ и $\mathbf{\Theta}$ — решение, то существует матрица $\mathbf{S}$ ранга $|T|$ такая, что $\mathbf{\Phi}'\mathbf{\Theta}' = (\mathbf{\Phi S})(\mathbf{S}^{-1}\mathbf{\Theta})$, но $\mathbf{\Phi}'$ и $\mathbf{\Theta}'$ не обязательно стохастические, т.е. метод главных компонент не подходит для тематического моделирования.

2.5. 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 $\hat{p}$):

$$\hat{p}(d,w) = \frac{n_{dw}}{n}, \quad \hat{p}(d) = \frac{n_d}{n}, \quad \hat{p}(w) = \frac{n_w}{n}, \quad \hat{p}(w,d) = \frac{n_{dw}}{n_d}, \quad \text{где}$$

- n_{dw} — число вхождений термина w в документ d;
- $n_d = \sum_{w \in W} n_{dw}$ — длина документа d в терминах;
- $n_w = \sum_{d \in D} n_{dw}$ — число вхождений термина w во все документы коллекции;
- $n = \sum_{d \in D} \sum_{w \in W} n_{dw}$ — длина коллекции d в терминах.

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

$$\hat{p}(t) = \frac{n_t}{n}, \quad \hat{p}(w|t) = \frac{n_{wt}}{n_t} \quad \hat{p}(t|d) = \frac{n_{dt}}{n_d} \quad \hat{p}(t|d,w) = \frac{n_{dwt}}{n_{dw}}, \quad \text{где}$$

- n_{dwt} — число троек, в которых термин w в документе d связан с темой t;
- $n_{dt} = \sum_{w \in W} n_{dwt}$ — число троек, в которых термин в документе d связан с темой t;
- $n_{wt} = \sum_{d \in D} n_{dwt}$ — число троек, в которых термин w связан с темой t;
- $n_t = \sum_{d \in D} \sum_{w \in W} n_{dwt}$ — число троек, связанных с темой t.

В пределе $n \to \infty$ частотные оценки $\hat{p}(\cdot)$ стремятся к соответствующим вероятностям $p(\cdot)$, согласно закону больших чисел.

Частотная интерпретация даёт ясное понимание всех условных вероятностей, которые будут использоваться в дальнейшем.

3. Вероятностный латентный семантический анализ PLSA

3.1. Стохастическое матричное разложение

Если число тем $|T|$ много меньше числа документов $|D|$ и числа терминов $|W|$, то равенство $p(w|d) = \sum_{t \in T} p(t|d)p(w|t)$ можно понимать как задачу приближённого представления заданной матрицы частот:

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

в виде произведения $\mathbf{F} \approx \mathbf{\Phi\Theta}$ двух неизвестных матриц меньшего размера — матрицы терминов тем $\mathbf{\Phi}$ и матрицы тем документов $\mathbf{\Theta}$, где

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

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

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

3.2. Принцип максимума правдоподобия

Для оценивания параметров $\mathbf{\Phi}$, $\mathbf{\Theta}$ тематической модели по коллекции документов D будем максимизировать плотность распределения выборки:

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

где C — нормировочный множитель, зависящий только от чисел n_{dw}. Отбросим множители Cи $p(d)$, не влияющие на положение точки максимума, подставим выражение для $p(w|d)$. Прологарифмируем $p(D; \mathbf{\Phi}, \mathbf{\Theta})$, чтобы превратить произведения в суммы. Получим задачу максимизации логарифма правдоподобия при ограничениях неотрицательности и нормированности столбцов матриц $\mathbf{\Phi}$ и $\mathbf{\Theta}$:

$$\begin{cases} \mathcal{L}_{\log}(\mathbf{\Phi}, \mathbf{\Theta}) = \sum\limits_{d \in D} \sum\limits_{w \in d} n_{dw} \ln \sum\limits_{t \in T} \phi_{wt} \theta_{td} \to \max\limits_{\mathbf{\Phi}\mathbf{\Theta}} \\ \phi_{wt} \geq 0, \quad \sum\limits_{w \in W} \phi_{wt} = 1 \\ \theta_{td} \geq 0, \quad \sum\limits_{t \in T} \theta_{td} = 1. \end{cases}$$

Для решения задачи применяется EM-алгоритм

3.3. EM-алгоритм

Для решения задачи в PLSA применяется итерационный процесс, в котором каждая итерация состоит из двух шагов — E (expectation) и M (maximization).

Перед первой итерацией выбирается начальное приближение параметров ϕ_{wt}, θ_{td}.

E-шаг.

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

$$H_{dwt} = p(t|d, w) = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}.$$

M-шаг.

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

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

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 — частотные оценки условных вероятностей ϕ_{wt}, θ_{td}:

$$\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \quad \hat{n}_{wt} = \sum_{d \in D} n_{dw} H_{dwt}, \quad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt};$$
$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \quad \hat{n}_{dt} = \sum_{w \in d} n_{dw} H_{dwt}, \quad \hat{n}_d = \sum_{t \in T} \hat{n}_{dt}.$$

Покажем теперь, что эти оценки действительно являются решением задачи максимизации правдоподобия при фиксированных H_{dwt}:

Запишем лагранжиан задачи максимизации логарифма правдоподобия при ограничениях нормировки, проигнорировав ограничения неотрицательности:

$$\mathcal{L}(\mathbf{\Phi}, \mathbf{\Theta}) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \underbrace{\sum_{t \in T} \phi_{wt} \theta_{td}}_{p(w|d)} - \sum_{t \in T} \lambda_t (\sum_{w \in W} \phi_{wt} - 1) - \sum_{d \in D} \mu_d (\sum_{t \in T} \theta_{td} - 1).$$

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

$$\lambda_t = \sum_{d \in D} n_{dw} \frac{\theta_{td}}{p(w|d)}$$

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 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

$$\phi_{wt} = \frac{\sum_{d \in D} n_{dw} H_{dwt}}{\sum_{w' \in W} \sum_{d \in D} n_{dw'} H_{dw't}}$$

Обозначив числитель через $\hat{n}_{wt}$, получим

$$\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \quad \hat{n}_{wt} = \sum_{d \in D} n_{dw} H_{dwt}, \quad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt}$$

Проделав аналогичные действия с производной лагранжиана по θ_{td},

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

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.

3.4. Начальное приближение ϕ_{wt} и θ_{td}

Начальное приближение можно задать нормированными случайными векторами из равномерного распределения.

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

Инициализация с частичным обучением (некоторые t известны заранее и имеются дополнительные данные о привязке некоторых d или w к t):

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

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

- 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:

$$\phi_{wt}^0 = \frac{1}{|W_t|}[w \in W_t].$$

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

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

- Если нет никакой априорной информации о связи документов с темами, то последнюю формулу можно применить к случайным подмножествам документов D_t.

3.5. Недостатки PLSA

У PLSA существуют некоторые недостатки, рассмотрим их:

1. Медленно сходится на больших коллекциях, так как $\mathbf{\Phi}$ и $\mathbf{\Theta}$ обновляются после каждого прохода коллекции.
2. Не разреживает распределение $H_{dwt} = p(t|d, w)$.
3. Вынуждены хранить матрицу $\mathbf{H} = (H_{dwt})_{D\times W\times T}$.
4. Слишком много параметров ϕ_{wt} и θ_{td} ($|W||T| + |T||D|$).
5. Неверно оценивает вероятность новых слов ($\hat{p}(w|t) = 0$ для слова, которого не было в обучающейся коллекции, но оно встретилось в каком-нибудь документе).
6. Не позволяет управлять разреженностью $\mathbf{\Phi}$ и $\mathbf{\Theta}$:

$$(\text{в начале } \phi_{wt} = 0) \Leftrightarrow (\text{в конце } \phi_{wt} = 0),$$
$$(\text{в начале } \theta_{td} = 0) \Leftrightarrow (\text{в конце } \theta_{td} = 0).$$

3.6. Модификации

Как мы увидели, у алгоритма существует достаточно много недостатков, поэтому чаще используют модификации метода, которые помогают устранить некоторые из них.

Рациональный EM–алгоритм

Проблема: Вынуждены хранить матрицу $\mathbf{H} = (H_{dwt})_{D\times W\times T}$.
Решение: Вычислять H_{dwt} по мере необходимости.
Обоснование:

Вычисление переменных $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на М-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, Е-шаг встраивается внутрь М-шага без дополнительных вычислительных затрат, отпадает необходимость хранения трёхмерной матрицы H_{dwt}. Заметим также, что переменную $\hat{n}_d$ можно не вычислять, поскольку $\hat{n}_d = n_d$.

Алгоритм 2. Рациональный EM-алгоритм

Input: Коллекция D, число тем T, начальные $\mathbf{\Phi}$ и $\mathbf{\Theta}$
Output: Распределения $\mathbf{\Phi}$ и $\mathbf{\Theta}$

repeat
  обнулить $\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t$ для всех $d \in D$, $w \in W$, $t \in T$;
  for all $d \in D$, $w \in d$ **do**
    $Z := \sum_{t \in T} \phi_{wt}\theta_{td}$;
    for all $t \in T$ таких, что $\phi_{wt}\theta_{td} > 0$ **do**
      увеличить $\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t$ на $\frac{n_{dw}}{Z}\phi_{wt}\theta_{td}$;
    end for
  end for
  $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D$, $t \in T$;
until $\mathbf{\Phi}$ и $\mathbf{\Theta}$ не стабилизируются; $=0$

Обобщенный EM-алгоритм

Проблема:PLSA медленно сходится на больших коллекциях.

Решение: Обновлять значения $\mathbf{\Phi}$ и $\mathbf{\Theta}$ чаще.

Обоснование:

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. 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 переменных ϕ_{wt}, θ_{td}, которые зависят от изменившихся скрытых переменных.

В случае PLSA сокращение M-шага сводится к более частому обновлению параметров ϕ_{wt} и θ_{td} по значениям счётчиков $\hat{n}_{wt}$ и $\hat{n}_{dt}$. Частота обновления выбирается — после каждого документа, термина, и т.д.

На больших коллекциях частые обновления повышают скорость сходимости. Частота обновления влияет 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 $\mathbf{\Phi}$ и $\mathbf{\Theta}$. Обновления после каждого вхождения термина являются избыточно частыми, в этом случае каждый термин документа приходится обрабатывать n_{dw} раз.

Необходимость хранения трёхмерной матрицы n_{dwt} делает алгоритм неприменимым к большим коллекциям. Этот недостаток можно устранить, например, сэмплированием.

Алгоритм 3. Обобщенный EM-алгоритм

Input: Коллекция D, число тем T, начальные $\mathbf{\Phi}$ и $\mathbf{\Theta}$
Output: Распределения $\mathbf{\Phi}$ и $\mathbf{\Theta}$

Обнулить $\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t, \hat{n}_d, \hat{n}_{dwt}$ для всех $d \in D$, $w \in W$, $t \in T$;
repeat
  for all $d \in D$, $w \in d$ **do**
    $Z := \sum_{t \in T} \phi_{wt}\theta_{td}$;
    for all $t \in T$ таких, что $n_{dwt} > 0$ или $\phi_{wt}\theta_{td} > 0$ **do**
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$ на $\frac{n_{dw}}{Z}\phi_{wt}\theta_{td} - n_{dwt}$;
      $n_{dwt} := \frac{n_{dw}}{Z}\phi_{wt}\theta_{td} - n_{dwt}$;
    end for
    if не первая итерация и пора обновить параметры $\mathbf{\Phi}$ и $\mathbf{\Theta}$ **then**
      $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$ таких, что $\hat{n}_{wt}$ изменился;
      $\theta_{td} := \hat{n_{dt}}/\hat{n}_d$ для всех $d \in D$, $t \in T$ таких, что $\hat{n}_{td}$ изменился;
    end if
  end for
until $\mathbf{\Phi}$ и $\mathbf{\Theta}$ не стабилизируются; $=0$

Модификация обобщенного EM-алгоритма

Проблема: Необходимо хранить массив $n_{dwt} = n_{dw}H_{dwt}$, который занимает $O(n|T|)$ памяти.

Решение: На М-шаге вместо распределения $H_{dwt} \equiv p(t|d,w)$ взять его несмещенную эмпирическую оценку по очень маленькой выборке длины s:

$$\hat{H}_{dwt} = \hat{p}(t|d,w) = \frac{1}{s}\sum_{i=1}^{s} \mathbb{I}_{t_{dwi}=t}.$$

В ряде публикаций 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.