Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Лозицкий Иван Павлович

Вероятностный латентный семантический анализ

Конспект

Санкт-Петербург
2016

Оглавление

Введение

С развитием вычислительной техники появилась необходимость в автоматической обработке текстов, созданных человеком и написанных на естественном языке.В силу большого количества естественных языков, сложности языковых конструкций, многие из задач, связанные с анализом текстов, являются весьма нетривиальными. Одним из направлений является работа с множествами текстовых документов. Задачей, которая относится к этому направлению и будет рассматриваться далее, является выявления скрытых тем в некоторой коллекции документов.

Глава 1

Модель коллекции текстовых документов

Задача выявления тем в коллекции документов начала изучаться относительно недавно. Основным подходом и инструментарием к решению этой задачи является тематическое моделирование. Тематическое моделирование(topic modeling) — одно из приложений машинного обучения к анализу текстов, активно развивающееся с конца 90–х годов. Основной предмет тематического моделирования — тематическая модель коллекции документов, определяет, к каким темам относится каждый документ и какие слова образуют каждую тему. Более частный случай — вероятностная тематическая модель, описывает каждую тему дискретным распределением на множестве терминов и каждый документ — дискретным распределением на множестве тем. Если рассматривать задачу определения тем коллекции документов как задачу кластеризации, то здесь речь пойдет о «мягкой кластеризации», так как документ или термин может одновременно относиться ко многим темам с различными вероятностями. Тем самым решаются проблемы синонимии и омонимии терминов, возникающие при «жесткой кластеризации». Вероятностные тематические модели применяются для категоризации и классификации текстовых документов, для информационного поиска, тегирования веб–страниц, обнаружения спама, для рекомендательных систем. Основным инструментарием тематического моделирования являются латентное размещение Дирихле (LDA) и вероятностный латентный семантический анализ (PLSA). Остановимся подробнее на модели PLSA. Далее будет рассмотрена формальная постановка задачи, основные этапы решения задачи выявления тем с помощью PLSA, недостатки и достоинства метода.

1.1. Постановка задачи

Пусть D — множество(коллекция) текстовых документов, W — множество(словарь) всех употребляемых в них терминов. Документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, \ldots, w_{n_d})$ из словаря W. Условимся, что термин может повторяться в документе несколько раз. Изначально мы полагаем, что существует конечное множество тем T, причем каждый термин $w \in W$ в каждом документе $d \in D$ связан с

некоторой темой $t \in T$[1]. Таким образов, коллекция документов может рассматриваться как множество троек вида (d, w, t), выбранных случайно и независимо из дискретного распределения $p(d, w, t)$ (которое заданно на конечном множестве $D \times W \times T$). При этом, тема $t \in T$ является скрытой переменной.

Задача тематического моделирования. Таким образом, мы определили пространство «документы–термины–темы»: $D \times W \times T$. Построить тематическую модель для коллекции документов D — значит найти множество тем T, распределение $p(w|t)$ для всех тем $t \in T$ и распределение $p(t|d)$ для всех документов. Далее, найденные распределения могут использоваться для решения прикладных задач.

1.2. Гипотезы и предположения

Введем некоторые предположения и гипотезы, являющиеся частью тематической модели.

Гипотеза независимости. Как отмечалось раннее, коллекция документов может рассматриваться как множество троек вида (d, w, t), выбранных случайно и независимо из дискретного распределения $p(d, w, t)$. Это эквивалентно тому, что порядок терминов в документах (как и документов в коллекции документов) не важен для выявления тематики. Таким образом, мы можем определить тематику документа. даже если переставили все слова в документе[2].

Гипотеза условной независимости. Если мы под темой понимаем распределение на множестве слов, то это распределение не зависит от документа. В этом и состоит гипотеза условной независимости. Более формально:

$$p(w|d, t) = p(w|t) \tag{1.1}$$

Таким образом, каждый документ можно рассматривать как смесь таких вероятностных распределений $p(w|t)$ по нескольким темам. В противном случае, нам бы пришлось строить более сложную модель.

Гипотеза разреженности. Естественно предполагать, что каждый документ d

[1] Это можно понимать так: человек, когда писал тексты, думал то об одной теме, то о другой

[2] На самом деле это позволяет перейти к более простому представлению документа как подмножества словаря $d \in W$, в котором каждому элементу $w \in d$, соответствует числов вхождений термина w в документ d

и каждый термин w связан с небольшим количеством тем t. В этом случае большая часть вероятностей $p(t|d)$ и $p(w|t)$ должна обращаться в нуль. Возможны два варианта:
1. Документ относится к большому количеству тем. В некоторых случаях имеет смысл разбить его на части, более однородные по тематике.
2. Термин относится к большому числу тем. Можем полагать, что в этом случае термин является общеупотребительным словом и несет мало полезной информации с точки зрения определения тематики.

1.3. Предварительная обработка документов

Зачастую, чтобы не усложнять модель, прибегают к предварительной обработке текстов. Рассмотрим основные из них. **Лемматизация**. Лемматизация — это приведение каждого слова в документе к его нормальной (начальной) форме. В русском языке это именительный падеж, единственное число, инфинитив — в зависимости от части речи. Для создания лемматизатора требуется составление словаря, либо формализация правил языка со всеми исключениями, что является трудоемким процессом.

Стемминг. Стемминг — это более простой способ упрощения. Состоит он в отбрасывании изменяемых частей слова. Недостаток — большое число ошибок.

Удаление лишних слов. В каждом языке встречаются слова, которые никак не характеризуют тему каким–либо образом. Это предлоги, союзы, частицы, местоимения и т.д.

Выделение ключевых слов. Во многих специальных текстах (математические, медицинские и т.д.) можно выделить ключевые слова — словосочетания, характеризующие предметную область. Это отдельная довольно сложная задача.
Далее будем считать, что словарь W получен в результате предварительной обработки всех текстовых документов коллекции D.

1.4. Модель порождения документа

Вернемся к нашей задаче — определения тематики документов по известной коллекции документов. Для наилучшего понимания ситуации, следует рассмотреть задачу в некотором смысле обратную — а как мы можем «породить» коллекцию документов?

Запишем формулу полной вероятности, учитывая гипотезу (1.1):

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

Если распределение $p(w|t)$ — распределение терминов в каждой теме и $p(t|d)$ — распределение тем в каждом документе, известны, то говорят, что вероятностная модель (1.2) описывает процесс порождения коллекции документов. Ниже представлены алгоритм (1) и иллюстрация (1.1).

Data: распределения $p(w|t), p(t|d)$

Result: выборка $(d_i, w_i), i = 1, \ldots, n$

forall $d \in D$ **do**
  задать длину n_d документа d;
  forall $i = 1, \ldots, n_d$ **do**
    выбрать случайную тему t из распределения $p(t|d)$;
    выбрать случайный термин из распределения $p(w|t)$;
    добавить в выборку пару (d, w), при этом тема t «забывается»;
  end
end

Algorithm 1: Вероятностная модель порождения коллекции документов

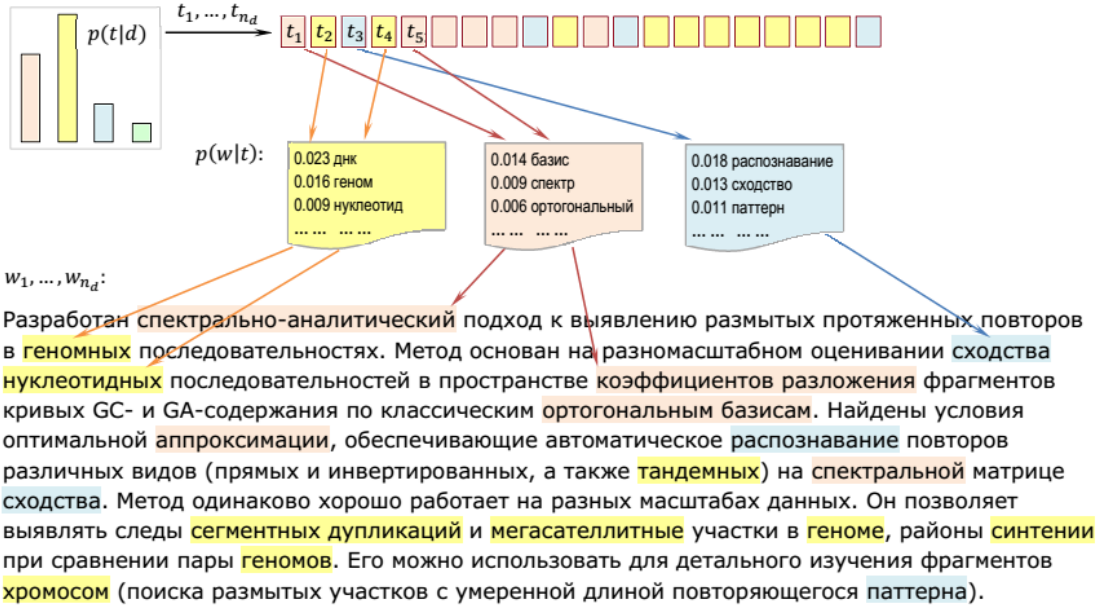

Рис. 1.1. Процесс порождения текстового документа вероятностной тематической моделью

Процесс построения тематической модели, как упоминалось выше, является обратной задачей: по известной коллекции документов требуется восстановить породившие

эту коллекцию распределения $p(t|d)$ и $p(w|t)$. Далее рассмотрим одно из решений поставленной задачи — вероятностный латентный семантический анализ.

Глава 2

Вероятностный латентный семантический анализ

Вероятностный латентный семантический анализ (PLSA) один из способов описания тематической модели представления коллекции текстовых документов. Рассмотрим основные обозначения, оценки и этапы данного метода.

2.1. Частотные оценки условных вероятностей

В дальнейшем будем использовать некоторые условные вероятности, часть из которых уже встречалась в формуле (1.2). Рассмотрим оценки вероятностей, частотная интерпретация которых даст понимание всех условных вероятностей.
Вероятности, связанные с наблюдаемыми переменными d и w, можно оценить по выборке как частоты:

$$\hat{p}(d,w) = \frac{n_{dw}}{n}, \quad \hat{p}(d) = \frac{n_d}{n}, \quad \hat{p}(w) = \frac{n_w}{n}, \quad \hat{p}(w,d) = \frac{n_{dw}}{n_d} \tag{2.1}$$

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, которые связаны с со скрытой переменной t, так же можно оценивать как частот, если рассматривать коллекцию документов как выборку троек (d, w, t):

$$\hat{p}(t) = \frac{n_t}{n}, \quad \hat{p}(w|t) = \frac{n_{wt}}{n_t}, \quad \hat{p}(t|d) = \frac{n_{dt}}{n_d}, \quad \hat{p}(t|d,w) = \frac{n_{dwt}}{n_{dw}} \tag{2.2}$$

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 частотные оценки (2.1)–(2.2) стремятся к соответствующим вероятностям по закону больших чисел.

2.2. Матричное разложение

Если число тем $|T|$ в коллекции документов много меньше числа документов $|D|$ и числа терминов $|W|$, то равенство (1.2) можно рассматривать как задачу приближенного представления матрицы частот:

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

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

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

Столбцы матриц неотрицательны и нормированы.
В вероятностном тематическом моделировании используется принцип максимума правдоподобия.

2.3. Принцип максимума правдоподобия

Для оценивания параметров Φ и Θ тематической модели, будем максимизировать плотность распределения выборки:

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

где C — нормировочный множитель. Выражение $Cp(d)^{n_{dw}}$ не влияет на положение максимума. Используя формулу (1.2) и введенные выше обозначения, прологарифмируем выражение выше и получим задачу максимума правдоподобия с ограничениями(неотрицательность и нормированность столбцов матриц Φ и Θ):

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}; \; : \phi_{wt} \geq 0; \quad \sum_{w \in W} \phi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1. \tag{2.3}$$

Для решения задачи (2.3) в PLSA применяется EM–алгоритм. Это итерационный процесс, состоящий из двух шагов. Перед первой итерацией выбираются начальные приближения параметров φ_{wt} и θ_{td} (об этом дополнительно будет сказано далее).

E–шаг.

На первом шаге вычисляем распределение вероятности $p(t|d,w)$ для скрытой переменной (темы) используя формулу Байеса для всех пар (d,w):

$$H_{dwt} = p(t|d,w) = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\varphi_{wt}\theta_{td}}{\sum_{s\in T}\varphi_{ws}\theta_{sd}}. \tag{2.4}$$

Пояснение: если у нас есть пара (d,w) — («документ», «слово») и хотим понять, из какой темы это слово.Если мы знаем вероятности слов в темах и тем в документах, то можем выразить вероятность того, что слово в документе относится к темам $t \in T$.

M–шаг.

На этом шаге по условным вероятностям H_{dwt} вычисляется новое значением параметров φ_{wt} и θ_{td}. Опишем процес вычисления φ_{wt} (для θ_{td} он будет аналогичен). Запишем лагранжиан задачи (2.3) при ограничениях нормировки:

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

Продифференцировав лагранжиан по φ_{wt}:

$$\frac{\partial\mathfrak{L}}{\partial\phi_{wt}} = \sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} - \lambda_t = 0;$$

и приравняв производную к нулю, получим:

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

Домножив это выражение на переменную, по которой дифференцировали и, просуммировав по w и используя формулу (2.4), получим:

$$\lambda_t = \sum_{d\in D}\sum_{w\in W} n_{dw}H_{dwt} \tag{2.6}$$

Теперь, использовав снова изначальное выражение (2.5) для λ_t, подставив в него (2.6), выражаем значение искомой переменной:

$$\phi_{wt} = \frac{\sum_{d\in D} n_{dw}H_{dwt}}{\sum_{d\in D}\sum_{w'\in W} n_{dw'}H_{dw't}} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \forall w\in W, t\in T$$

Аналогично для θ_{td}.

В итоге получаем:

$$\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \quad \hat{n}_{wt} = \sum_{d\in D} n_{dw}H_{dwt}, \quad \hat{n}_t = \sum_{w\in W}\hat{n}_{wt}; \tag{2.7}$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \quad \hat{n}_{dt} = \sum_{w \in d} n_{dw} H_{dwt}, \quad \hat{n}_d = \sum_{t \in T} \hat{n}_{dt} \tag{2.8}$$

Пояснение: заметим, что величина

$$\hat{n}_{dwt} = n_{dw} p(t|d,w) = n_{dw} H_{dwt} \tag{2.9}$$

оценивает число вхождений термина w в документ d, связанных с темой t. Просуммировав это значения по всем документам, получим сколько раз слово w было связано с темой t по всей коллекции и т.д. В H_{dwt} у нас возникает такой подсчет, где каждое слово в каждом документе учтено в итоге не через количество вхождений (0, 1,...), а скажем так в вероятностном смысле, в какой-то доле. То есть мы не можем однозначно сказать, что какое-то слово в каком-то документе принадлежит теме t, можем только слово в документе «распределить» по всем темам.

2.4. Выбор начальных параметров

Есть несколько способов задать начальные приближения φ_{wt} и θ_{td}.

Начальные приближения можно задавать нормированными случайными векторами из равномерного распределения.

Другой способ — пройти по всей коллекции, выбрать для каждой пары (d,w) случайную тему t и вычислить частотные оценки вероятностей φ_{wt} и θ_{td} для всех d, w, t.

Частичное обучение. Этот способ применяется, когда некоторые темы известны заранее и имеются некоторые дополнительные данные.

Если известно, что документ d относится к множеству тем $T_d \subset T$, то в качестве θ^0_{td} можно взять равномерное распределение:

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

Если некоторое множество терминов $W_t \subset W$ относится к теме t, то в качестве начального приближения можно взять равномерное распределение на этом множестве:

$$\varphi^0_{td} = \frac{1}{W_t}[w \in W_t].$$

Если известно, что некоторое множество документов $D_t \subset D$ относится к теме t, то можно взять распределение слов в «объединенном» документе:

$$\varphi^0_{td} = \frac{\sum_{d \in D_t} n_{dw}}{\sum_{d \in D_t} n_d}$$

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

2.5. Недостатки PLSA и модификации метода

Основными недостатками классического PLSA считают:

- Медленно сходится на больших коллекциях, так как Φ и Θ обновляются после каждого прохода коллекции.
- Не разреживает распределение $H_{dwt} = p(t|d,w)$ (Скорее всего, каждая пара (d,w) характеризуется малым количеством тем, но приходится для каждой пары строить все распределение тем).
- Вынуждены хранить матрицу $H = (H_{dwt})_{D\times W\times T}$.
- PLSA переобучается, так как параметров φ_{wt} и $theta_{td}$ слишком много ($|D|\cdot|T| + |W|\cdot|T|$) и на них не накладывается никаких ограничений.
- Неверно оценивает вероятность новых слов (если слова не было в обучающей коллекции и попался документ, в котором встретилось новое слово — для него частотная оценка вероятности $\hat{p}(w|t) = 0$).
- Не позволяет управлять разреженностью Φ и Θ[1], т.к.
  (в начале $\phi_{wt} = 0$) $\Leftrightarrow$ (в конце $\phi_{wt} = 0$)
  (в начале $\theta_{td} = 0$) $\Leftrightarrow$ (в конце $\theta_{td} = 0$)

Далее рассмотрим модификации и преобразования алгоритма, позволяющие устранить некоторые проблемы.

Рациональный EM–алгоритм. Данная модификация позволяет значения параметров тем же самым способом, но добиться некоторого выигрыша в памяти.
Идея — не хранить H_{dwt}, а вычислять по мере необходимости.

[1] то есть алгоритм структуру разжеренности матриц не дает сам — какую структуру мы заложим в начале, такую и получим. Мы хотим, чтобы алгоритм сказал, что большинства тем нет в конкретном документе или каких-то слов нет в конкретной теме

Data: Коллекция D, число тем T, начальные Φ и Θ;

Result: распределения Φ и Θ;

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

Algorithm 2: Рациональный EM–алгоритм

В данном варианте M–шаг берется за основу, а внутрь встраивается E–шаг. В то время, когда пробегаемся по всем парам (d, w) ко всем «счетчикам» (т.е. параметрам, из которых получим потом требуемые оценки) прибавляем величину $n_{dw}H_{dwt}$. Нормировочный множитель Z считается один раз до «пробежки» по всем темам. Это эквивалент старого алгоритма, но просто по-другому реализованный, без хранения дополнительной матрицы.

Обобщенный EM–алгоритм Представим, что коллекция очень большая. Поэтому хотим обновлять значения вычисляемых параметров чаще.

Идея: не обязательно точно решать задачу M–шага, достаточно сместиться в направлении максимума и снова сделать E–шаг.

В PLSA это приводит к частым обновлениям параметров Φ и Θ.

Data: Коллекция D, число тем T, начальные Φ и Θ;

Result: распределения Φ и Θ;

обнулить $\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t, \hat{n}_d, n_{dwt}$ для всех $d \in D$, $w \in W$, $t \in T$;

repeat

  forall $d \in D$, $w \in d$ **do**

    $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$;

    forall $t \in T$ *таких, что* $n_{dwt} > 0$ *или* $\varphi_{wt}\theta_{td} > 0$ **do**

      увеличить $\hat{n}_{wt}, \hat{n}_{dt}, \hat{n}_t, \hat{n}_d$ на $\frac{n_{dw}}{Z}\varphi_{wt}\theta_{td} - n_{dwt}$;

      $n_{dwt} := \frac{n_{dw}}{Z}\varphi_{wt}\theta_{td} - n_{dwt}$;

    end

    if *не первая итерация и пора обновить параметры* Φ *и* Θ **then**

      $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$ таких, что $\hat{n}_{wt}$ изменился;

      $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D$, $t \in T$ таких, что $\hat{n}_{td}$ изменился;

    end

  end

until Φ *и* Θ *не стабилизируются*;

Algorithm 3: Обобщенный EM–алгоритм

В EM–алгоритме нет необходимости сверхточно решать задачу максимизации правдоподобия на M–шаге. Достаточно немного приблизиться к точке максимума и снова выполнить E-шаг. Это связано с тем, что сам функционал правдоподобия известен не точно — он зависит от приближённых значений H_{dwt}, которые были получены на E–шаге. Еще одно обобщение состоит в том, что E–шаг выполняется только для части скрытых переменных. После этого M–шаг выполняется только для тех переменных $\varphi_{wt}, \theta_{td}$, которые зависят от изменившихся скрытых переменных. Частота обновления выбирается — после каждого документа, термина, и т.д. На больших коллекциях частые обновления повышают скорость сходимости.

Здесь мы жертвуем памятью и вводим n_{dwt}. Как это работает? Если на прошлой итерации посчитали n_{dwt} и добавили его ко всем счетчикам, то на текущей итерации мы должны это значение вычесть, посчитать новое значение n_{dwt} и добавить новое значение. Тогда мы будем гарантировать, что все счетчики содержат ровно один проход по коллекции.

Что значит «если не первая итерация»: одну итерацию по коллекции при начальных значениях параметров мы все–таки должны сделать, иначе параметры посчитаются по

«сырым» счетчикам, которые были посчитаны по маленькой части коллекции[2].

Сэмплирование. В ходе построения обобщенного EM–алгоритма, решили одну проблему, но получили другую — массив $n_{dwt} = n_{dw}H_{dwt}$.

Идея: зная распределение H_{dwt}(раз его уже посчитали), сгенерируем из него некоторое количество тем. То есть для каждой пары (d, w) генерируется s случайных тем $t_{dwi}, i = 1, \ldots, s$ из распределения $p(t|d, w)$.

Что в нашем случае значит сэмплирование? То, что истинное распределение $H_{dwt} \equiv p(t|d, w)$ по всем темам заменили эмпирическое оценкой по очень маленькой выборке длины s:

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

Так как все новые значения $\hat{H}_{dwt}$, так же как и старые, усредняются в ходе алгоритма, то получаем хорошие оценки и, учитывая что сэмплирование проходит по маленькой выборке, добиваемся разреженности матрицы.

[2] Иначе значения параметров могут получиться хуже, чем начальные приближения

Заключение

Таким образом рассмотрели задачи тематического моделирования, построения вероятностных тематических моделей и генерации коллекции текстовых документов. Так же познакомились с особенностями вероятностного латентного семантического анализа, его недостатках и модификациями.