Вероятностное тематическое моделирование

Мурат Апишев *
great-mel@yandex.ru

25 сентября 2015

1 Введение

Тематическое моделирование — одно из основных направлений статистического анализа текстов, активно развивающееся последние два десятилетия. Цель построения вероятностной тематической модели заключается в автоматическом извлечении тем из коллекции текстовых документов. При этом происходит поиск информации о том, какими терминами описывается каждая тема, и каким темам принадлежит каждый из документов коллекции.

В рамках лекции будет описана вероятностная постановка задачи тематического моделирования, показан метод её решения на основе метода простых итераций (EM-алгоритм). Также будет рассказано о теории *аддитивной регуляризации тематических моделей* (АРТМ), которая предлагает гибкий и простой математический аппарат для обучения моделей с различными свойствами.

2 Вероятностное тематическое моделирование

Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) всех употребляемых в них терминов. В качестве терминов могут выступать как отдельные слова, так и словосочетания. Каждый документ $d \in D$ — последовательность n_d терминов $w_1, \cdots, w_{n_d}$ из словаря W.

Вероятностное пространство. Предполагается существование конечного множества тем T, и каждое вхождение термина w в документ d связано с некоторой темой $t \in T$. D рассматривается как случайная и независимая выборка троек (w_i, d_i, t_i), $i = \overline{1, n}$ из дискретного распределения $p(w, d, t)$ на конечном вероятностном пространстве $W \times D \times T$. Термины w и документы d являются наблюдаемыми переменными, тема $t \in T$ является *латентной* скрытой переменной.

Предположим две гипотезы:

1. **Гипотеза мешка слов**: предположим, что тематика документа не зависит от порядка терминов в документе, а также от того, каким по счёту этот документ

*Материал лекции полностью основан на работах К. В. Воронцова. Можно обращаться к ним при поиске ссылок на дополнительную литературу.

Алгоритм 2.1: Алгоритм генерации текстовой коллекции

Входные данные: распределения $p(w|t), p(t|d)$;
Выходные данные: выборка пар $(d_i, w_i), i = \overline{1, n}$;
для каждого $d \in D$ **выполнять**
  задать длину n_d документа d;
  для каждого $i = \overline{1, n_d}$ **выполнять**
    $d_i := d$;
    выбрать случайную тему t_i из распределения $p(t|d_i)$;
    выбрать случайный термин w_i из распределения $p(w|t_i)$;

был в коллекции. Гипотеза «мешка слов» позволяет перейти к компактному представлению коллекции документов, где каждому документу d ставится в соответствие словарь его уникальных слов, в котором каждому термину w ставится в соответствие счётчик числа его вхождений в документ n_{dw} [1].

2. **Гипотеза условной независимости:** предполагаем, что появление слов в документе d по теме t зависит от темы, но не зависит от документа, и описывается общим для всех документов распределением $p(w|t)$. Допускается три эквивалентных представления этой гипотезы:

$$p(w|d,t) = p(w|t); \quad p(d|w,t) = p(d|t); \quad p(d,w|t) = p(d|t)p(w|t). \tag{2.1}$$

Вероятностная порождающая модель выражает вероятности $p(w|d)$ появления терминов w в документах d через распределения $p(w|t)$ и $p(t|d)$. Из формулы полной вероятности и гипотезы условной независимости получаем:

$$p(w|d) = \sum_{t \in T} p(w|d,t)p(t|d) = \sum_{t \in T} p(w|t)p(t|d). \tag{2.2}$$

Данная генеративная модель описывает процесс создания коллекции при известных распределениях $p(w|t)$ и $p(t|d)$ (см. алгоритм 2.1).

Тематическое моделирование решает обратную задачу — по известной коллекции D произвести оценку параметров модели $\varphi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$.

Обычно число тем $|T| << |D|$ и $|W|$, и задача сводится к поиску приближенного представления заданной матрицы частот $F = (f_{wd})_{W \times D}, f_{wd} = \dfrac{n_{dw}}{n_d}$ в виде произведения $F \approx \Phi\Theta$ двух неизвестных матриц меньшего размера — *матрицы терминов тем* $\Phi = (\varphi_{wt})_{W \times T}$ и *матрицы тем документов* $\Theta = (\theta_{td})_{T \times D}$. Все три матрицы являются *стохастическими*, т.е. все их столбцы неотрицательны и нормированы (являются вероятностными распределениями).

[1] На текущий момент, во многих исследованиях, связанных с тематическими моделями, гипотеза «мешка слов» отвергается в пользу использования последовательного текста. Это связано с тем, что отбрасывание данных о порядке терминов в текстах на самом деле приводит к потере полезной информации.

Частотные оценки условных вероятностей. Вероятности, связанные с наблюдаемыми переменными d, w можно оценивать по выборке как частоты. Такие частотные оценки являются несмещёнными оценками максимального правдоподобия (здесь и далее все выборочные оценки вероятностей p будем обозначать $\hat{p}$):

$$\hat{p}(d,w) = \frac{n_{dw}}{n}, \quad \hat{p}(w) = \frac{n_w}{n}, \quad \hat{p}(d) = \frac{n_d}{n}, \quad \hat{p}(w|d) = \frac{n_{dw}}{n_d} \tag{2.3}$$

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 вхожденией документа w во все документы коллекции;

$n = \sum_{d \in D} \sum_{w \in d} n_{dw}$ — 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):

$$\hat{p}(t) = \frac{n_t}{n}, \quad \hat{p}(w|t) = \frac{n_{wt}}{n_t}, \quad \hat{p}(t|d) = \frac{n_{dt}}{n_d}, \quad \hat{p}(t|d,w) = \frac{n_{dwt}}{n_{dw}} \tag{2.4}$$

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

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

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

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

Эти оценки нельзя вычислить непосредственно по исходным данным, поскольку темы t неизвестны. Однако все эти оценки выражаются через $n_{tdw} = p(t|d,w) n_{dw}$. Таким образом, знание условных распределений $p(t|d,w)$ даёт возможность оценить искомые параметры тематической модели $\varphi_{wt} = \hat{p}(w|t)$ и $\theta_{td} = \hat{p}(t|d)$.

В пределе при $n \to \infty$ частотные оценки $\hat{p}(\cdot)$, определяемые формулами 2.3–2.4, стремятся к соответствующим вероятностям $p(\cdot)$, согласно закону больших чисел.

3 PLSA и EM-алгоритм

Принцип максимума правдоподобия. Для оценивания параметров Φ и Θ тематической модели по коллекции документов D будем максимизировать правдоподобие выборки:

$$p(D; \Phi, \Theta) = \prod_{i=1}^{n} p(d_i, w_i) = \prod_{i=1}^{n} p(w_i|d_i) p(d_i) = \prod_{d \in D} \prod_{w \in d} p(w|d)^{n_{dw}} p(d)^{n_{dw}} \to \max_{\Phi, \Theta}.$$

Прологарифмируем правдоподобие, чтобы превратить произведения в суммы, и отбросим константные слагаемые, не зависящие от параметров модели. Получим задачу максимизации логарифма правдоподобия при ограничениях неотрицательности и нормированности столбцов матриц Φ и Θ:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \to \max_{\Phi, \Theta} \tag{3.1}$$

$$\sum_{w \in W} \varphi_{wt} = 1, \; \varphi_{wt} \geq 0; \tag{3.2}$$

$$\sum_{t \in T} \theta_{td} = 1, \; \theta_{td} \geq 0. \tag{3.3}$$

Такая постановка задачи является основой *вероятностного латентного семантического анализа* (PLSA). Для её решения используется EM-алгоритм. Прежде всего получим алгоритм элементарным путём, который даёт интуитивное понимание, а после приведём его строгое обоснование.

EM-алгоритм. Искомые параметры модели выражаются через частотные оценки условных вероятностей: $\varphi_{wt} = \frac{n_{wt}}{n_t}$ и $\theta_{td} = \frac{n_{dt}}{n_d}$. Как было сказано ранее, их можно оценить, зная $n_{tdw} = n_{dw} p(t|d, w)$. Чтобы выразить условные вероятности $p(t|d, w)$ через параметры модели, воспользуемся формулой Байеса и гипотезой условной независимости:

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

Таким образом, получаем систему уравнений относительно параметров модели φ_{wt} и θ_{td} и вспомогательных переменных p_{tdw}, n_{wt}, n_{dt}:

$$p_{tdw} = \frac{\varphi_{wt}\theta_{td}}{\sum_{s \in T} \varphi_{ws}\theta_{sd}}; \tag{3.4}$$

$$\varphi_{wt} = \frac{n_{wt}}{\sum_{v \in W} n_{vt}}; \qquad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \tag{3.5}$$

$$\theta_{td} = \frac{n_{dt}}{\sum_{s \in T} n_{sd}}; \qquad n_{dt} = \sum_{w \in d} n_{dw} p_{tdw}. \tag{3.6}$$

Для решения данной системы нелинейных уравнений подходит метод простых итераций: сначала выбираются начальные приближения параметров φ_{wt} и θ_{td}, затем вычисления по формулам 3.4–3.6 продолжаются в цикле до сходимости. В терминах EM-алгоритма вычисление условных вероятностей по формуле 3.4 называется E-шагом (expectation), а вычисление оценок максимального правдоподобия по формулам 3.5–3.6 — M-шагом (maximiazation).

Строгое обоснование формул шагов EM-алгоритма. Покажем, что полученные оценки параметров модели 3.5–3.6 действительно являются решением задачи максимизации правдоподобия 3.1–3.3 при фиксированных p_{tdw}.

Запишем лагранжиан задачи 3.1 при ограничениях нормировки, проигнорировав ограничения неотрицательности (позже убедимся, что решение действительно неотрицательно):

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

Продифференцируем лагранжиан по φ_{wt}, приравняем производную к нулю и выразим λ_t:

$$\lambda_t = \sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)}. \tag{3.7}$$

Домножим обе части этого равенства на φ_{wt}, просуммируем по всем терминам $w \in W$:

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

Теперь в левой части применим условие нормировки вероятностей φ_{wt} и выделим перменную p_{tdw} в правой:

$$\lambda_t = \sum_{d\in D}\sum_{w\in W} n_{dw}p_{tdw}. \tag{3.8}$$

Снова домножим обе части 3.7 на φ_{wt}, выделим переменную p_{tdw} в правой части и выразим φ_{wt} из левой части, подставив уже известное выражение для λ_t. Получим:

$$\varphi_{wt} = \frac{\sum_{d\in D} n_{dw}p_{tdw}}{\sum_{v\in W}\sum_{d\in D} n_{dv}p_{tdv}}.$$

Обозначим числитель через n_{wt}, и сразу получаем формулы 3.5.

Проделав аналогичные действия с производной по θ_{td}, получим 3.6.

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

Алгоритм 3.1: PLSA-EM: рациональный EM-алгоритм для модели PLSA

Входные данные: коллекция D, число тем $|T|$, нач. приближения Φ, Θ;
Выходные данные: распределения Φ, Θ;

повторять
  обнулить n_{wt}, n_{dt}, n_t для всех $d \in D$, $w \in W$, $t \in T$;
  для каждого $d \in D$, $w \in d$ **выполнять**
    $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$; увеличить n_{wt}, n_{dt}, n_t на $\dfrac{n_{dw}\varphi_{wt}\theta_{td}}{Z}$ для всех тем $t \in T$;
  $\varphi_{wt} := \dfrac{n_{wt}}{n_t}$, для всех $w \in W, t \in T$;
  $\theta_{td} := \dfrac{n_{dt}}{n_t}$, для всех $d \in D, t \in T$;
до тех пор, пока Φ, Θ *не сойдутся*;

Рациональный EM-алгоритм является простейшей модификацией обычного, описанного выше. Вычисление переменных n_{wt}, n_{dt}, n_t на M-шаге требует однократного прохода по всей коллекции в цикле по всем документам $d \in D$ и по всем терминам $w \in d$. Внутри этого цикла переменные p_{tdw} можно считать только в тот момент, когда они нужны. От этого результат алгоритма не изменится, E-шаг встраивается внутрь M-шага без дополнительных вычислительных затрат, отпадает необходимость хранения трёхмерной матрицы p_{tdw}. Рациональный EM-алгоритм показан в листинге 3.1.

4 Аддитивная регуляризация

Поставленная задача стохастического матричного разложения является некорректно поставленной, так как множество её решений в общем случае бесконечно. Если $F = \Phi\Theta$ — решение, то $F = (\Phi S)(S^{-1}\Theta)$ тоже является решением для всех невырожденных S, при которых матрицы $\Phi' = \Phi S$ и $\Theta' = S'\Theta$ являются стохастическими.

Для решения некорректно поставленных задач существует общий подход, называемый *регуляризацией*. К недоопределенной оптимизационной задаче добавляется дополнительный критерий — регуляризатор, по возможности учитывающий специфические особенности данной задачи и знания предметной области.

Аддитивная регуляризация тематических моделей (АРТМ) основана на введении дополнительных критериев-регуляризаторов $R_i(\Phi, \Theta), i = \overline{1, r}$,и максимизации их линейной комбинации с логарифмом правдоподобия $L(\Phi, \Theta)$:

$$R(\Phi,\Theta)=\sum_{i=1}^{k}\tau_i R_i(\Phi,\Theta), \qquad \mathcal{L}(\Phi,\Theta)+R(\Phi,\Theta)\to\max_{\Phi,\Theta} \tag{4.1}$$

$$\sum_{w\in W}\varphi_{wt}\in\{0,1\},\ \varphi_{wt}\geq 0; \tag{4.2}$$

$$\sum_{t\in T}\theta_{td}\in\{0,1\},\ \theta_{td}\geq 0. \tag{4.3}$$

где τ_i — неотрицательные *коэффициенты регуляризации*. Оптимизация взвешенной суммы критериев является широко распространённым приёмом в многокритериальной оптимизации.

Модель PLSA соответствует частному случаю отсутствия регуляризатора ($R(\Phi,\Theta)=0$).

В модели PLSA рост числа тем будет приводить только к росту правдоподобия модели. Для регуляризованной модели это не обязательно. Поэтому ограничения-равенства 4.2, 4.3 записаны с вариативной правой частью, допускающей обнуление столбцов Φ и Θ.

Обнуление φ_t, как и t-й строки матрицы Θ, означает исключение темы из модели. Таким образом, в модель закладывается возможность определять оптимальное количество тем, при условии, что изначально было задано избыточное число тем.

Если $\theta_d=0$, то документ d фактически исключается из коллекции. Регуляризованная модель может отказаться определять тематику документа, если он слишком короткий или не релевантен тематике коллекции.

В байесовский методах обучения тематических моделей реуляризатор $R(\Phi,\Theta)$ интерпретируется как логарифм априорного распределения, а оптимизационная задача 4.1 соответствует принципу максимума апостериорной вероятности. В АРТМ регуляризатор не обязан иметь вероятностную интерпретацию.

Введём оператор неотрицательного нормирования, который преобразует произвольный вектор $(x_i)_{i\in I}$ в вектор вероятностей $(p_i)_{i\in I}$ дискретного распределения путём обнуления отрицательных элементов с последующей нормировкой:

$$p_i=\operatorname*{norm}_{i\in I} x_i=\frac{\max\{x_i,0\}}{\sum_{j\in I}\max\{x_j,0\}}, \forall i\in I.$$

Если $x\leq 0,\ \forall i\in I$, то результатом применения операции будет нулевой вектор.

Теорема 1. Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка (Φ,Θ) локального экстремума задачи 4.1–4.3 удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}=$ $=p(t|d,w)$:

$$p_{tdw}=\operatorname*{norm}_{t\in T}\left(\varphi_{wt}\theta_{td}\right); \tag{4.4}$$

$$\varphi_{wt}=\operatorname*{norm}_{w\in W}\left(n_{wt}+\varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right); \qquad n_{wt}=\sum_{d\in D}n_{dw}p_{tdw}; \tag{4.5}$$

$$\theta_{td}=\operatorname*{norm}_{t\in T}\left(n_{td}+\theta_{td}\frac{\partial R}{\partial\theta_{td}}\right); \qquad n_{td}=\sum_{w\in d}n_{dw}p_{tdw}. \tag{4.6}$$

Доказательство основано на теореме Каруша-Куна-Таккера и проводится по тем же принципам, что и в выкладках, сделанных в 3.

Решение системы уравнений 4.4–4.6 методом простых итераций приводит к регуляризованному EM-алгоритму. Вычисление параметров p_{tdw} по формуле 4.4 называется E-шагом, оценивание параметров φ_{wt}, θ_{td} по формулам 4.5–4.6 — M-шагом. Эти параметры можно инициализировать случайным образом.

Дивергенция Кульбака-Лейблера или *KL-дивергенция* будет активно использоваться при построении регуляризаторов. Это несимметричная функция расстояния между дискретными распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

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

Предполагается, что $p_i > 0$ и $q_i > 0$. Кроме того, распределения P и Q должны иметь общий носитель $\Omega = \{i: \ p_i > 0, q_i > 0\}$.

Основные свойства KL-дивергенции:

1. KL-дивергенция неотрицательна и равна нулю тогда, и только тогда, когда распределения совпадают.

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.

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

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

Максимизация правдоподобия 3.1 эквивалентна минимизации взвешенной суммы дивергенций Кульбака-Лейблера между эмпирическими распределениями $\hat{p}(w|d) = \dfrac{n_{dw}}{n_d}$ и модельными $p(w|d)$ по всем документам $d \in D$:

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

где весом документа d является его длина n_d.

§4.1 Простейшие регуляризаторы.

Сглаживание. Потребуем, чтобы столбцы φ_t и θ_d были близки к заданным распределениям $\beta_t = (\beta_{wt})_{w\in W}$ и $\alpha_d = (\alpha_{td})_{t\in T}$ в смысле KL-дивергенции:

$$\sum_{t\in T}\mathrm{KL}_w(\beta_{wt}\|\varphi_{wt}) \to \min_{\Phi}, \quad \sum_{d\in D}\mathrm{KL}_t(\alpha_{td}\|\theta_{td}) \to \min_{\Theta}.$$

Складывая два критерия с коэффициентами β_0, α_0 и удаляя из суммы константы, получим регуляризатор

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

Применение общих формул 4.5 и 4.6 даёт выражение для М-шага:

$$\varphi_{wt} = \underset{w\in W}{\mathrm{norm}}(n_{wt} + \beta_0\beta_{wt}); \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}(n_{td} + \alpha_0\alpha_{td}).$$

Сглаживающий регуляризатор эквивалентен предположению, что столбцы матриц Φ и Θ порождаются априорными распределениями Дирихле в гиперпараметрами $\beta_0\beta_t$ и $\alpha_0\alpha_d$. В модели *латентного размещения Дирихле* LDA гиперпараметры могут быть только положительными.

Разреживание. Недостаток сглаживающего регуляризатора — его противоречие с *гипотезой разреженности*. Естественно предполагать, что каждый документ и каждый термин связаны с небольшим числом тем t. В таком случае значительная часть вероятностей φ_{wt} и θ_{td} должна быть нулевой.

Чем сильнее разрежено распределение, тем ниже его энтропия. Максимальной энтропией обладает равномерное распределение. Идея разреживания состоит в том, чтобы максимизировать дивергенции $\mathrm{KL}_w\left(\frac{1}{|W|}\middle\|\varphi_{wt}\right)$ и $\mathrm{KL}_t\left(\frac{1}{|T|}\middle\|\theta_{td}\right)$ между искомыми распределениями и равномерными. Обобщая эту идею, зададим вместо равномерных распределений произвольные распределения $\beta_t = (\beta_{wt})_{w\in W}$ и $\alpha_d = (\alpha_{td})_{t\in T}$. В таком случае разреживание оказывается полной противоположностью сглаживанию:

$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in T}\sum_{w\in W}\beta_{wt}\ln\varphi_{wt} - \alpha_0\sum_{d\in D}\sum_{t\in T}\alpha_{td}\ln\theta_{td} \to \max.$$
$$\varphi_{wt} = \underset{w\in W}{\mathrm{norm}}(n_{wt} - \beta_0\beta_{wt}); \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}(n_{td} - \alpha_0\alpha_{td}).$$

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

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

Это критерий не зависит от Θ, поэтому формулы М-шага для θ_{td} не претерпят изменений.

Формула для φ_{wt}, согласно 4.5, примет вид

$$\varphi_{wt} = \underset{w \in W}{\mathrm{norm}} \left(n_{wt} - \tau \varphi_{wt} \sum_{s \in T \setminus t} \varphi_{ws} \right).$$

Смысл этой формулы в том, что условные вероятности $\varphi_{wt} = p(w|t)$ постепенно уменьшаются для тех терминов w, которые имеют большие значения вероятности φ_{ws} в других темах. Это регуляризатор также является разреживающим.

5 Применение тематического моделирования

Одним из основных приложений тематического моделирования является *информационный поиск*. Современные поисковые системы предназначены для поиска по коротким поисковым запросам. Они основаны на инвертированных индексах, в которых для каждого слова хранится список документов, в которых оно встречается. Система ищет документы, в которых встречаются все слова запросов, поэтому по длинному запросу, скорее всего, ничего не будет найдено. Тематический или *разведочный поиск* — это разновидность информационного поиска. Он подходит не для ответов на конкретные вопросы, а для расширения профессиональных знаний. Если пользователь плохо ориентируется в терминологии или слабо представляет себе структуру предметной области, то его потребностью будет получение «дорожной карты» предметной области, систематизация и визуализация релевантной информации по заданной теме. Тема запроса формулируется не словами, а текстовым фрагментом произвольной длины.

Тематические модели применяются также для выявления трендов в новостных потоках или научных публикациях, для многоязычного информационного поиска, для анализа данных социальных сетей, для классификации и категоризации документов, для тематической сегментации текстов, для построения рекомендательных систем.