Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Молчанова Юлия Юрьевна

Проверка адекватности тематических моделей в онлайновых алгоритмах

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:
д.ф.-м.н., доцент
К.В. Воронцов

Москва, 2016

Содержание

1 Введение

Тематическое моделирование — одно из современных приложений машинного обучения к анализу текстов. Тематическая модель коллекции текстовых документов определяет каждую тему как дискретное распределение на множестве терминов, а каждый документ — как дискретное распределение на множестве тем. Предполагается, что каждый документ — набор терминов, выбранных независимо и случайно из смеси распределений. Задача тематического моделирования состоит в восстановлении компонент смеси по выборке.

Основополагающей гипотезой вероятностного тематического моделирования является гипотеза условной независимости — предположение о том, что распределения слов в теме и различных документах не зависят от документа. Характерная особенность задачи — сильно разреженные дискретные распределения, что затрудняет применение стандартных асимптотик различных критериев.

В работе предложен способ оценивания тематических моделей, базирующийся на проверке гипотезы условной независимости . Предложенный способ допускает оценивание адекватности отдельных документов, тем, а также модели в целом. Также предложена и реализована адаптация указанного способа для онлайнового алгоритма обучения тематической модели.

2 Вероятностное тематическое моделирование текстовых коллекций

2.1 Задача матричного разложения и гипотеза условной независимости

Пусть D — коллекция текстовых документов, W — множество всех употребляемых в них терминов. Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, \ldots, w_{n_d})$ из словаря W. Через n_{dw} обозначается число вхождений термина w в документ d.

Следующие предположения являются ключевыми для вероятностных моделей:

1. Существует конечное множество тем T, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$. Коллекция документов рассматривается как множество троек (d, w, t), выбранных 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 скрытой переменной.

2. *Гипотеза «мешка слов»* (bag of words) заключается в предположении, что порядок терминов в документах не важен для выявления тематики, то есть тематику документа можно узнать даже после произвольной перестановки терминов. 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.

3. *Гипотеза условной независимости* эквивалентна предположению, 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 допускает три эквивалентных представления:

$$\begin{aligned} p(w \mid d, t) &= p(w \mid t); \\ p(d \mid w, t) &= p(d \mid t); \\ p(d, w \mid t) &= p(d \mid t) p(w \mid t). \end{aligned} \tag{1}$$

Согласно формуле полной вероятности и гипотезе условной независимости распределение слов в каждом документе d представляется в виде

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d). \tag{2}$$

Задача построения тематической модели коллекции состоит в том, чтобы по известной левой части этого равенства $p(w \mid d) = n_{dw}/n_d$ оценить неизвестные условные распределения в правой части: $p(w \mid t)$ для каждой темы $t \in T$ и $p(t \mid d)$ для каждого документа $d \in D$, а также определить оптимальное число тем $|T|$.

Вероятности, связанные с наблюдаемыми переменными d и w, можно оценивать по выборке как частоты (здесь и далее выборочные оценки вероятностей p будем обо-

значать через $\hat{p}$):

$$\hat{p}(d,w) = \frac{n_{dw}}{n}, \qquad \hat{p}(d) = \frac{n_d}{n}, \qquad \hat{p}(w) = \frac{n_w}{n}, \qquad \hat{p}(w \mid d) = \frac{n_{dw}}{n_d}, \tag{3}$$

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):

$$\hat{p}(t) = \frac{n_t}{n}, \qquad \hat{p}(w \mid t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(t \mid d) = \frac{n_{td}}{n_d}, \qquad \hat{p}(t \mid d, w) = \frac{n_{dwt}}{n_{dw}}, \tag{4}$$

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

$n_{td} = \sum\limits_{w \in W} n_{dwt}$ — 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.

Если число тем $|T|$ много меньше числа документов $|D|$ и числа терминов $|W|$, то равенство (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 стохастических матриц — *матрицы терминов тем* Φ и *матрицы тем документов* Θ:

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

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

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

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 (2) и воспользуемся обозначениями $\theta_{td} = p(t\,|\,d)$, $\phi_{wt} = p(w\,|\,t)$. Прологарифмируем $p(D;\Phi,\Theta)$. Получим задачу максимизации логарифма правдоподобия (log-likelihood) при ограничениях неотрицательности и нормированности столбцов матриц Φ и Θ:

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}; \tag{5}$$

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

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

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 за собой неустойчивость EM-алгоритма.

Задачи, решение которых неединственно или неустойчиво, называются *некорректно поставленными*. Общий подход к их решению называется *регуляризацией* [14]. Он заключается в том, чтобы некоторым разумным образом ввести дополнительные ограничения на Φ,Θ, сузив тем самым множество решений.

Допустим, что наряду с правдоподобием (5) требуется максимизировать n критериев $R_i(\Phi,\Theta)$, $i = 1,\ldots,n$, называемых *регуляризаторами*. Для решения задачи многокритериальной оптимизации применим метод *скаляризации*. Будем максимизировать линейную комбинацию логарифма правдоподобия и критериев R_i с неотрицательными *коэффициентами регуляризации* τ_i, при условии неотрицательности и нормировки столбцов матриц Φ и Θ:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \tag{6}$$

$$R(\Phi,\Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta);$$

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

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

Такой подход к решению задачи тематического моделирования задаёт модель *ARTM* (Additive Regularization Topic Modelling) [2]. Частным случаем *ARTM* при $R_i(\Phi,\Theta) = 0$, $i = 1,\ldots,n$ является модель *PLSA* (Probabilistic Latent Semantic Analysis) [3]. Для решения задачи (6) в ARTM 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}.

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:

$$p(t \mid d, w) = \frac{p(w \mid t)p(t \mid d)}{p(w \mid d)} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}}. \tag{7}$$

На M-шаге по условным вероятностям тем $p(t \mid d, w)$ вычисляется новое приближение параметров ϕ_{wt}, θ_{td}. Величина $n_{dwt} = n_{dw}p(t \mid d, w)$ оценивает число вхождений термина w в документ d, связанных с темой t.

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(\hat{n}_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \qquad \theta_{td} = \underset{t \in T}{\operatorname{norm}}\left(\hat{n}_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \tag{8}$$

где $\underset{s \in S}{\operatorname{norm}}(x) = \dfrac{\max\{0, x_s\}}{\sum\limits_{s \in S} \max\{0, x_s\}}$ — операция нормирования вектора.

2.2 Оффлайновый EM-алгоритм

Одна из возможных реализаций EM-алгоритма показана в Алгоритме 1. Этот алгоритм применим для решения задачи тематического моделирования только при небольших объёмах данных, поскольку в ходе алгоритма совершается несколько полных проходов по всей коллекции документов.

2.3 Онлайновый EM-алгоритм

Вычисление параметров модели ϕ_{wt}, θ_{td} на M-шаге требует однократного прохода по всей коллекции в цикле по всем документам $d \in D$ и всем словам каждого документа $w \in d$. Существует множество версий EM-алгоритма, отличающихся частотой обновления параметров модели ϕ_{wt}, θ_{td} по переменным n_{wt} и n_{td} [4].

Одной из таких модификаций оффлайнового алгоритма является онлайновый EM-алгоритм, основная идея которого заключается в том, что итерации для вычисления θ_{td} производятся при фиксированной матрице Φ для каждого документа d до сходимости. Переменные n_{wt} накапливают частоту слова w в теме t. На последней итерации документа производится накопительное обновление переменных n_{wt}. Переменные, по которым происходит обновление, обозначим $\tilde{n}_{wt}$. Обновления матрицы Φ по переменным n_{wt} происходят по окончании обработки документа или пакета документов (batch). При этом

Вход: коллекция документов D, число тем $|T|$;

Выход: Φ, Θ;

инициализировать вектор-столбцы ϕ_t, θ_d случайным образом

повторять

  обнулить n_{wt}, n_{td} для всех $d \in D,\ w \in W,\ t \in T$

  для всех $d \in D,\ w \in d$

    для всех $t \in T$

      $p(t \mid d, w) = \underset{t \in T}{\operatorname{norm}}(\phi_{wt}\theta_{td})$

      увеличить n_{wt}, n_{td} на $n_{dw}p(t \mid d, w)$

  $\phi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(\hat{n}_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$ для всех $w \in W,\ t \in T$

  $\theta_{td} = \underset{t \in T}{\operatorname{norm}}\left(\hat{n}_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$ для всех $d \in D,\ t \in T$

пока Θ и Φ не сойдутся

Алгоритм 1. Оффлайновый EM-алгоритм для модели ARTM.

обычно для сглаживания изменения матрицы Φ используется усреднение с помощью эспоненциального скользящего среднего с коэффициентом дисконтирования $\rho \in (0, 1]$.

На больших коллекциях матрица Φ обычно сходится после обработки относительно небольшой части документов. В результате даже одного прохода по коллекции бывает достаточно для построения модели, поэтому онлайновый алгоритм является предпочтительным для обработки больших коллекций. Одна из возможных реализаций онлайнового EM-алгоритма показана в Алгоритмах 2, 3.

2.4 Оценки адекватности тематических моделей

Существует много различных подходов к оцениванию качества вероятностных тематических моделей. Наиболее распространённый метод основан на вычислении логарифма правдоподобия для отложенной выборки: часть документов разбиваются на две половины: на первой половине настраивается вектор θ_d, на второй половине вычисляется логарифм правдоподобия. Этот метод является неточным. Более точные, но и более ресурсоёмкие способы предложены в статье [5].

Было предложено еще много различных автоматических способов оценивания адекватности тематических моделей: в [6] используется тест на слишком большую дисперсию остатков модели, что позволяет автоматически определять число тем в модели. Д.Мимно и Д.Блэй [1] предложили способ проверки гипотезы условной независимости,

Вход: коллекция D, число тем $|T|$, коэффициент дисконтирования $\rho \in (0, 1]$;

Выход: матрица Φ;

инициализировать вектор-столбцы ϕ_t случайным образом для всех $t \in T$

n_{wt}:=0, $\tilde{n}_{wt}$:=0 для всех $w \in W$, $t \in T$

для всех пакетов D_b, $b = 1, ..., B$

  $(\tilde{n}_{wt})$:=$(\tilde{n}_{wt})$ + ProcessBatch(D_b, Φ)

  $n_{wt} := \rho n_{wt} + \tilde{n}_{wt}$ для всех $w \in W$, $t \in T$

  $\phi_{wt} = \underset{w \in W}{\text{norm}} \left(\hat{n}_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ для всех $w \in W$, $t \in T$

  $\tilde{n}_{wt}$:=0 для всех $w \in W$, $t \in T$

Алгоритм 2. Онлайновый EM-алгоритм для модели ARTM.

основанный на постериорных проверках.

Ещё одной популярной метрикой является когерентность (семантическая согласованность), которая показывает, насколько часто слова, встречающиеся рядом, оказываются в одной теме [7]. В работе М. Робертс [8] предлагается вычислять семантическую согласованность и коэффициенты корреляции тем и выбирать модель, для которой семантическая согласованность высокая, а коэффициенты корреляции тем низкие.

Такие автоматические метрики удобны тем, что они сравнительно просто вычисляются и позволяют получить объективный результат, но эксперименты над интерпретируемостью моделей показывают, что автоматические метрики качества и оценки экспертов далеко не всегда согласованы [9]. Разумеется, наилучший метод оценки модели — тщательное чтение экспертами каждого текста и слов, характерных для каждой темы.

Несмотря на большую освещённость в литературе проблемы оценивания качества тематических моделей, был предложен всего один способ проверки гипотезы условной независимости [1]. Характеристика независимости распределений $p(w \mid d, t)$ и $p(w \mid t)$ измерялась посредством вычисления взаимной информации между W_t (словами, отнесёнными к теме t) и документами D при фиксированной теме t. Для оценки степени независимости этих распределений для каждой темы был предложен следующий способ: взаимная информация вычислялась для реальных данных и для 100 сэмплированных наборов данных, в которых дополнительно для каждой пары (d, w) сэмплировалась тема t. В качестве характеристики степени независимости использовалась выраженная в стандартных отклонениях разница между значением взаимной информации для ре-

Вход: пакет D_b, матрица $\Phi = (\phi_{wt})$;

Выход: матрица $(\tilde{n}_{wt})$;

$\tilde{n}_{wt}$:=0 для всех $w \in W$, $t \in T$

для всех $d \in D_b$

  инициализировать $\theta_{td} := \frac{1}{T}$ для всех $t \in T$

  повторять

    $p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$ для всех $w \in W$, $t \in T$

    $n_{td} := \sum_{w \in d} n_{dw} p_{tdw}$ для всех $t \in T$

    $\theta_{td} = \underset{t \in T}{\text{norm}}\left(\hat{n}_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$ для всех $t \in T$

  пока Θ и Φ не сойдутся

  $\tilde{n}_{wt} := \tilde{n}_{wt} + n_{dw} p_{tdw}$ для всех $w \in W$, $t \in T$

Алгоритм 3. ProcessBatch(D_b, Φ)

альных данных и математическим ожиданием эмпирического распределения значения взаимной информации, вычисленного на основе сэмплированных данных. Соответствующие значения получаются очень высокими: 16 стандартных отклонений для темы с выполненной гипотезой условной независимости и 34 стандартных отклонения — с невыполненной. Такие значения слабо подчиняются вероятностной интерпретации.

Метод позволяет сравнивать степень выполнения гипотезы для нескольких тем и предоставляет новый способ визуализации тематических моделей. Недостатком этого подхода является отсутствие интерпретируемого результата и, как следствие, фактическая невозможность непосредственной проверки гипотезы условной независимости для темы. К тому же, этот способ предполагает сэмплирование 100 наборов данных с последующим сэмплированием темы для каждой пары (d, w) для каждого набора данных. Это очень ресурсоёмкие операции при больших объёмах данных.

2.5 Постановка задачи

Задача состоит в разработке вычислительно эффективного статистического теста для проверки гипотезы условной независимости сильно разреженных распределений и его адаптации для использования в онлайновом алгоритме. Способ должен быть достаточно быстрым для практического применения в тематическом моделировании для оценки адекватности модели в процессе её построения — после обработки каждого пакета документов.

3 Статистический тест для проверки гипотезы условной независимости

3.1 Семейство дивергенций Кресси-Рида

Пусть имеется выборка n независимых наблюдений $X_n = \{x_1, \ldots, x_n\}$ случайной величины, принимающей значения из конечного множества Ω. Её эмпирическое распределение определяется как доля наблюдений x_i, равных x:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} [x_i = x], \quad x \in \Omega.$$

Критерий хи-квадрат проверяет нулевую гипотезу о том, что случайная величина имеет заданное распределение $p(x)$, $x \in \Omega$. Для этого вычисляется статистика хи-квадрат X^2, G^2, связанная с дивергенцией Кульбака–Лейблера, или H^2, связанная с метрикой Хеллингера в пространстве распределений:

$$X^2 = n \sum_{x \in \Omega} \frac{(\hat{p}(x) - p(x))^2}{p(x)},$$
$$G^2 = 2n \sum_{x \in \Omega} \hat{p}(x) \ln \frac{\hat{p}(x)}{p(x)},$$
$$H^2 = 4n \sum_{x \in \Omega} \left(\sqrt{\hat{p}(x)} - \sqrt{p(x)}\right)^2.$$

Все эти статистики являются частными случаями дивергенции Кресси–Рида [11] при значениях параметра $\lambda = 1$, $\lambda \to 0$ и $\lambda = -\frac{1}{2}$ соответственно:

$$\mathrm{CR}_\lambda(\hat{p} : p) = \frac{2n}{\lambda(\lambda + 1)} \sum_{x \in \Omega} \hat{p}(x) \left(\left(\frac{\hat{p}(x)}{p(x)} \right)^\lambda - 1 \right).$$

При условии истинности нулевой гипотезы распределение каждой из этих статистик стремится к распределению хи-квадрат с $k = |\Omega| - 1$ степенями свободы: $\mathrm{CR}_\lambda \sim \chi^2(k)$. Нулевая гипотеза отвергается на уровне значимости α, если значение статистики превышает $(1 - \alpha)$-квантиль этого распределения: $\mathrm{CR}_\lambda > \chi^2_{1-\alpha}(k)$.

Считается, что асимптотика хи-квадрат применима, если объём выборки $n \geqslant 50$ и ожидаемое число наблюдений $np(x) \geqslant 5$ для каждого $x \in \Omega$. Однако для разреженных распределений $p(x)$, когда вероятности $p(x)$ малы для многих $x \in \Omega$ или когда $|\Omega| \gg n$, второе условие может не выполняться даже на больших выборках [12]. Стандартная рекомендация — объединять значения $x \in \Omega$ в группы — для разреженных распределений оказывается неприемлемой, так как результат теста может зависеть от способа группирования, выбираемого произвольно.

Для разреженных распределений $p(x)$ предлагается вместо асимптотического распределения $\chi^2(k)$ использовать эмпирическое распределение.

Для оценивания квантили распределения статистики используется тест на основе сэмплирования. Генерируется N выборок $X_{jn} = \{x_{j1}, \ldots, x_{jn}\}$ независимых наблюдений из распределения $p(x)$; для каждой из них вычисляется эмпирическое распределение $\hat{p}_{jn}(x)$ и значение статистики $S_{jn} = \mathrm{CR}_\lambda(\hat{p}_{jn} : p)$, $j = 1, \ldots, N$. По значениям $S_{1n}, \ldots, S_{Nn}$ строится эмпирическая функция распределения статистики

$$\hat{F}_n(S) = \frac{1}{N} \sum_{j=1}^{N} \big[S > S_{jn}\big].$$

и вычисляется её $(1-\alpha)$-квантиль $\hat{F}_{n,1-\alpha}$. Число N рекомендуется брать не менее 1000, если необходимо оценивать всю функцию распределения. Однако если оценивается только одна квантиль, N можно брать порядка нескольких десятков [13].

3.2 Зависимость статистики Кресси-Рида от длины документа

Ввиду отсутствия приемлемых асимптотик для проверки гипотезы условной независимости $p(w \mid d, t) = p(w \mid t)$ предлагается сравнить значение статистики Кресси-Рида с квантилью её эмпирического распределения. Значение статистики Кресси-Рида $CR_\lambda(d, t)$ вычисляется по счётчикам $n_{dwt}, n_{td}, n_{wt}, n_t$:

$$\begin{aligned}\mathrm{CR}_\lambda(d,t) &= \frac{2n_{td}}{\lambda(\lambda+1)} \sum_{w \in d} p(w \mid d, t) \left(\left(\frac{p(w \mid d, t)}{p(w \mid t)} \right)^\lambda - 1 \right) = \\ &= \frac{2}{\lambda(\lambda+1)} \sum_{w \in d} n_{dwt} \left(\left(\frac{n_{dwt} n_t}{n_{td} n_{wt}} \right)^\lambda - 1 \right). \end{aligned} \tag{9}$$

Экспериментально показано, что распределение статистики $CR_\lambda(t, d)$ существенно зависит от числа n_d вхождений темы t в документ d при условии выполнения гипотезы условной независимости [10]. Соответственно, основная задача состоит в том, чтобы эффективно восстановить эту зависимость.

3.3 Эмпирическое распределение статистики Кресси-Рида

Для получения эмпирического распределения статистики $CR_\lambda(t, d)$ необходимо многократно сгенерировать выборку, заведомо удовлетворяющую гипотезе условной независимости. В тематической модели тест гипотезы условной независимости строится для каждой темы в каждом документе, поэтому число искусственных выборок может достигать $|D| \times |T|$. Тем не менее, оказывается, что для восстановления зависимости

квантили эмпирического распределения достаточно сгенерировать только одну искусственную выборку, совпадающую по размерам с исходной коллекцией.

Для получения искусственной выборки в документе d для каждого вхождения слова w сэмплируем тему t из распределения $p(t \mid d)$ и новое слово w' из распределения $p(w \mid t)$. Таким образом, получаем искусственную коллекцию документов D' того же объёма, что D.

Искусственная коллекция обладает следующими свойствами: гипотеза условной независимости выполняется по построению, а длины документов совпадают с длинами документов исходной коллекции. Для этой коллекции обозначим счётчики числа слов $n'_{dwt}, n'_{dt}, n'_{wt}, n'_t$.

Вычислиется статистика $CR_\lambda(t,d)$ для каждой пары тема–документ (t,d) в этих двух коллекциях. Обозначим $CR_\lambda(t,d)$ статистику Кресси–Рида, вычисленную по исходной коллекции, и $CR'_\lambda(t,d)$ — по искусственной. Вычисление этих значений легко встраивается в EM-алгоритм. Реализация для модели PLSA показана в Алгоритме 4. Аналогичным образом вычисление может быть произведено в оффлайновом EM-алгоритме для модели ARTM. Заметим, что такая схема вычисления не меняет асимптотической сложности работы алгоритма.

На основе совокупности значений $\{CR'_\lambda(t,d) | d \in D', t \in T\}$ статистики Кресси-Рида, вычисленной для искусственной коллекции, можно восстановить зависимость эмпирического распределения статистики от числа вхождений темы в документ для каждой темы. При этом количество наблюдений — количество документов в коллекции — обычно достаточно велико для качественного восстановления зависимости.

Для восстановления зависимости $\tilde{F}_{1-\alpha}(n_{td})$ статистики Кресси-Рида от числа вхождений темы t в документ d предлагается упорядочить документы по неубыванию числа вхождений темы в них и вычислить эмпирическую квантиль $\hat{F}_{1-\alpha}$ для каждых 100 подряд идущих документов.

Полученную таким образом выборку пар $(n_{td}, \hat{F}_{1-\alpha})$ предлагается использовать для восстановления зависимости $\tilde{F}_{1-\alpha}(n_{td})$ с помощью полиномиальной регрессии. В экспериментах все вычисления производились для квантили порядка $1-\alpha = 0.95$, где $alpha$ — уровень значимости статистического теста.

Вход: коллекция документов D, число тем $|T|$;

Выход: Φ, Θ, $CR_\lambda(d,t)$;

инициализировать вектор-столбцы ϕ_t, θ_d случайным образом

повторять

  обнулить n_{wt}, n_{td}, n_t, n'_{wt}, n'_{td}, n'_t для всех $d \in D,\ w \in W,\ t \in T$

  обнулить n'_{wd}

  для всех $d \in D,\ w \in d$

    для всех $t \in T$

      $p(t \mid d, w) = \underset{t \in T}{\operatorname{norm}}(\phi_{wt}\theta_{td})$, увеличить n_{wt}, n_{td}, n_t на $n_{dw}p(t \mid d, w)$

    сэмплировать тему t' из распределения $p(t \mid d)$, сэмплировать слово w' из распределения $p(w \mid t')$, увеличить $n'_{w'd}$ на единицу

    для всех $t \in T$

      увеличить $n'_{w't}$, n'_{td}, n'_t на $n'_{dw't} = p(t \mid d, w')$

  обнулить $CR_\lambda(d,t)$, $CR'_\lambda(d,t)$ для всех $d \in D,\ t \in T$

  для $d \in D, w \in W, t \in T$

    если $n_{dwt} > 0$ **то**

      $CR_\lambda(d,t) := CR_\lambda(d,t) + \frac{2}{\lambda(\lambda+1)} n_{dwt} \left(\left(\frac{n_{dwt} n_t}{n_{td} n_{wt}} \right)^\lambda - 1 \right)$

    если $n'_{dwt} > 0$ **то**

      $CR'_\lambda(d,t) := CR'_\lambda(d,t) + \frac{2}{\lambda(\lambda+1)} n'_{dwt} \left(\left(\frac{n'_{dwt} n'_t}{n'_{dt} n'_{wt}} \right)^\lambda - 1 \right)$

  $\phi_{wt} = \underset{w \in W}{\operatorname{norm}}(n_{wt})$ для всех $w \in W,\ t \in T$

  $\theta_{td} = \underset{t \in T}{\operatorname{norm}}(n_{td})$ для всех $d \in D,\ t \in T$

пока Θ и Φ не сойдутся

Алгоритм 4. Оффлайновый EM-алгоритм для модели PLSA с вычислением $CR_\lambda(t,d)$, $CR'_\lambda(t,d)$ для каждой пары тема–документ (t,d) на каждой итерации.

3.4 Сэмплирование из разреженных дискретных распределений

Для получения искусственной коллекции необходимо сэмплировать большое число слов из дискретного распределения $p(w \mid t)$. У этого распределения большая мощность множества носителя — весь словарь коллекции, — но обычно оно является сильно разреженным, что позволяет эффективно применить метод сэмплирования сложностью $O(1)$ с подготовкой данных сложностью $O(N \log N)$ для каждой темы, где N — число слов с ненулевой вероятностью для фиксированной темы t [15].

Для эффективного сэмплирования из дискретных распределений обычно используется метод алиасов, мы же рассмотрим одну из его модификаций, особенно эффективных в условиях разреженных распределений: метод квадратной гистограммы [16].

Алгоритм, подготавливающий данные для метода квадратной гистограммы, принимает на вход вектор вероятностей дискретного распределения p_i, $i = 1, ..., N$. Для эффективного сэмплирования требуется разбить отрезок $[0, 1]$ на участки, соответствующие разным элементам носителя, таким образом, чтобы определение участка, в который попала реализация случайной величины $\xi \sim R[0, 1]$, происходило максимально быстро. Для этого формируется квадратная матрица размера 1×1, в которой каждый столбец и каждая строка имеют ширину $\frac{1}{N}$.

Изначально столбец с номером i соответствует одному элементу носителя распределения i и имеет высоту p_i. Выбираются самый высокий столбец i и самый низкий столбец j, самый низкий столбец заполняется до высоты $\frac{1}{N}$ элементами j, в отдельный массив сохраняется элемент верхней части i-го столбца $K[i] := j$ и точка перехода от элементов i к элементам j: $V[j] := p_i + \frac{i}{N}$. Теперь столбец i имеет высоту $\frac{1}{N}$ и в дальнейших преобразованиях не участвует. Столбец j имеет высоту $p_j - \left(\frac{1}{N} - p_i\right)$. Процесс повторяется, пока каждый столбец не окажется высотой $\frac{1}{N}$. Искомая матрица сформирована.

Заметим, что в итоге каждый столбец включает в себя элементы не более, чем двух разных типов, а расход памяти составляет $2N$. После такой подготовки возможно сэмплирование сложностью $O(1)$:

1. сэмплировать реализацию U равномерной случайной величины $\xi \sim R[0, 1]$;
2. вычислить номер столбца i, которому она соответствует: $i := \lfloor UN \rfloor$;
3. если U выше сохранённой точки перехода $V[i]$, вернуть элемент верхней части столбца $K[i]$, иначе вернуть элемент нижней части столбца i.

Метод квадратных гистограмм позволяет особенно эффективно сэмплировать из разреженных дискретных распределений, при этом дополнительные затраты памяти при встраивании в EM-алгоритм оказываются пренебежимо малыми.

4 Адаптация теста условной независимости для онлайнового EM-алгоритма

4.1 Особенности пакетной обработки текстовой коллекции

Для пакетной обработки данных характерны некоторые особенности, которые делают невозможным вычисление статистики по предложенной ранее схеме. В онлайновом EM-алгоритме тест применяется к каждому пакету так, как он выше применялся ко всей коллекции:

1. Обнулить n'_{wt} для всех тем и всех документов.

2. Перед обработкой пакета D_b провести сэмплирование эталонного пакета аналогично уже описанному: в каждом документе пакета d для каждого вхождения любого слова w сэмплировать тему t из распределения $p(t \mid d)$ и новое слово w' из распределения $p(w \mid t)$. Таким образом, получить новый пакет D'_b, счётчики для которого обозначим $\tilde{n}'_{wt}, \tilde{n}'_t, \tilde{n}'_{td}$ и $\tilde{n}'_{dwt}$.

3. По сэмплированному пакету вычислить статистику Кресси-Рида для всех тем и документов пакета, используя в качестве n_{wt} и n_t не локальные счётчики сэмплированного пакета $\tilde{n}'_{wt}$ и $\tilde{n}'_t$, а глобальные счётчики сэмплированной коллекции n'_{wt} и n'_t:
$$CR'_\lambda(d,t) = \frac{2}{\lambda(\lambda+1)} \sum_{w \in d} \tilde{n}'_{dwt} \left(\left(\frac{\tilde{n}'_{dwt} n'_t}{\tilde{n}'_{dt} n'_{wt}} \right)^\lambda - 1 \right). \tag{10}$$

4. Для каждой темы восстановить зависимость квантили эмпирического распределения статистики Кресси-Рида от длины документа по документам сэмплированного пакета.

5. Начать проход по пакету документов. После каждой итерации обработки каждого документа вычисление статистики Кресси-Рида произвести аналогичным изложенному ранее способом за тем исключением, что в качестве n_{wt} и n_t использовать не локальные счётчики пакета, а глобальные счётчики обработанной части коллекции, усреднённые методом экспоненциального скользящего среднего: вычислить
$$CR_\lambda(d,t) = \frac{2}{\lambda(\lambda+1)} \sum_{w \in d} \tilde{n}_{dwt} \left(\left(\frac{\tilde{n}_{dwt} n_t}{\tilde{n}_{dt} n_{wt}} \right)^\lambda - 1 \right). \tag{11}$$
Таким образом, мы получим характеристику адекватности документа модели в процессе его анализа.

6. Обновить счётчики n'_{wt} с учётом $\tilde{n}'_{wt}$ с коэффициентами метода экспоненциального скользящего среднего.

Отметим, что сэмплирование перед каждым пакетом необходимо, так как перед пакетом происходит обновление счётчиков n_{wt} и матрицы Φ, соответственно. Следовательно, темы при обработке нового пакета могут иметь другие распределения, в частности, другую разреженность, поэтому зависимость необходимо восстанавливать заново, она может существенно измениться.

4.2 Проверка адекватности темы в пакете документов

Для проверки адекватности темы в пакете документов предлагается вычислить долю документов, для которых значение статистики Кресси-Рида не превысило значение квантили распределения статистики в точке, соответствующей числу вхождений данной темы в данный документ. При этом значение квантили в точке определяется согласно полиномиальной интерполяции. Доля таких документов характеризует адекватность модели для этой темы.

Темы, для которых указанная доля документов значительно ниже $1 - \alpha$, где α — уровень значимости теста, считаются не отвечающими гипотезе условной независимости. Параметром является порог разделения тем по данному критерию, он может быть заменён на произвольную долю документов. В экспериментах использовался порог, в точности равный $1 - q$.

В качестве другого критерия адекватности модели для отдельно взятой темы предлагается рассмотреть долю слов данной темы, которые приходятся на документы, для которых значение статистики Кресси-Рида не превысило значение квантили распределения статистики в точке, соответствующей числу вхождений данной темы в данный документ. Обозначим такие документы адекватными для фиксированной темы.
Тогда взвешенная адекватность темы:

$y_t = \sum\limits_{d \in D_t} \frac{n_{td}}{\sum\limits_{d \in D_t} n_{td}} \left[CR_\lambda(d,t) < \hat{F}_{1-\alpha} \right]$, где D_t — множество адекватных документов для темы t,

Адекватность темы:

$y'_t = \sum\limits_{d \in D_t} \frac{\left[CR_\lambda(d,t) < \hat{F}_{1-\alpha} \right]}{D_t}$, где D_t — множество адекватных документов для темы t.

Доля пар (d, t), прошедших статистический тест, может служить легко интерпретируемой метрикой качества всей тематической модели.

4.3 Проверка адекватности тематической модели документа

Аналогично, для каждого документа пакета можно определить адекватные темы: темы, для которых значение статистики Кресси-Рида не превысило значение квантили распределения статистики в точке, соответствующей числу вхождений даннной темы в данный документ. Обозначим адекватные для документа d темы T_d

В качестве критерия адекватности модели конкретного документа можно рассмотреть долю адекватных для него тем и долю слов документа, соответствующим адекватным для него темам.

Взвешенная адекватность документа:

$$y_d = \sum_{t \in T_d} \frac{n_{td}}{\sum\limits_{t \in T_d} n_{td}} \left[CR_\lambda(d,t) < \hat{F}_{1-\alpha} \right].$$

Адекватность документа:

$$y'_d = \sum_{t \in T_d} \frac{\left[CR_\lambda(d,t) < \hat{F}_{1-\alpha} \right]}{T_d}.$$

Документы с аномально низким значением этих характеристик могут считаться шумовыми, их можно исключить из пакета и повторить обработку. Соответственно, такой метод оценивания тематических моделей дополнительно предоставляет способ автоматического отбора качественных документов.

4.4 Критерий остановки итераций по документу

Заметим, что после каждой итерации по документу d мы вычисляем значение статистики Кресси-Рида $CR_\lambda(d,t)$ для всех тем t. Следовательно, на каждой итерации можно вычислить введённые выше характеристики адекватности документа. Значения этих характеристик могут выступать в качестве критерия остановки итераций. Обычно совершается фиксированное число итераций по каждому документу, но по достижении хорошей тематизации документа, которое сопровождается выполнением гипотезы условной независимости для многих тем, можно прекратить итерации раньше, тем самым сократив время работы алгоритма.

4.5 Обнаружение эффектов перерегуляризации

Резкое ухудшение критериев качества для большого числа тем может свидетельствовать о чрезмерной регуляризации. Если после обработки одного пакета документов тест стал гораздо чаще отвергать гипотезу условной независимости во всей модели,

скорее всего, после обработки пакета была применена слишком сильная регуляризация. Особенно критично слишком высокое значение коэффициента при регуляризаторе разреживания матрицы Φ. Соответственно, сильное снижение характеристик адекватности тем может послужить сигналом к откату последних изменений и уменьшению коэффициентов регуляризации.

4.6 Обнаружение новых тем в пакете документов

Невыполнение гипотезы условной независимости для темы может свидетельствовать о том, что число тем задано неправильно, и может позволить получить начальное приближение для разбиения этой темы на несколько новых тем: для каждой темы можно определить документы, не прошедшие тест гипотезы условной независимости, затем вычислить суммарные вклады слов w в значения $CR_\lambda(d,t)$, не прошедшие тест; и для каждой темы построить списки слов с максимальными суммарными вкладами. Это и будет начальным приближением для расщепления тем.

Также признаком появления новой темы может служить отклонение гипотезы условной независимости для многих документов пакета. Если пакеты соответствуют интервалам времени в новостном потоке, плохая тематизация всего пакета может свидетельствовать о появлении совершенно новой темы.

5 Вычислительные эксперименты

5.1 Исследование значения адекватности тем и документов на различных итерациях оффлайнового EM-алгоритма

Цель эксперимента: исследовать зависимость значения адекватности тем и документов от номера шага EM-алгоритма при выполнении гипотезы условной независимости.

Ход эксперимента: эксперимент проводился на полумодельных данных. Полумодельные данные представляют собой коллекцию, сэмплированную из точного произведения матриц Φ_m и Θ_m, полученных в результате решения задачи тематического моделирования для реальной коллекции документов. Длины документов полагались равными случайным числам от 100 до 1500. В матрицах Φ_m и Θ_m 100 тем. Восстанавливались 100 тем в течение 25 итераций EM-алгоритма. Запуск оффлайнового EM-алгоритма производился на полумодельном наборе данных с $|D| = 1377$, $|W| = 6906$.

Использовался регуляризатор разреживания матрицы Θ с коэффициентом регуляризации -0.15.

Значение адекватности и взвешенной адекватности вычислялось для всех тем и всех документов на каждой итерации оффлайнового EM-алгоритма.

На графике рис. 3 приводится средняя адекватность тем и документов в зависимости от итерации. Из графика видно, что гипотеза условной независимости не отвергается для значительной доли тем, начиная с восьмой итерации EM-алгоритма, что говорит о достижении хорошей тематизации уже на данном этапе и согласуется с заведомым выполнением гипотезы условной независимости в полумодельной коллекции. Также из графика видно, что, начиная с этой же итерации, происходит перерегуляризация матрицы Θ: монотонное улучшение средней адекватности сменяется её резким убыванием.

На графике рис. 1 показана зависимость адекватности документа от длины на различных итерациях EM-алгоритма. Из графика видно, что на ранних итерациях EM-алгоритма короткие документы считаются хорошо тематизированными, а в длинных документах гипотеза условной независимости считается существенно нарушенной. По мере сходимости EM-алгоритма адекватность длинных документов принимает высокие значения, а адекватность коротких документов становится несколько ниже. Это связано с тем, что короткие документы сильнее подвержены эффектам сэмплирования: эмпирические распределения слов в документах и темах существенно отличаются от распределений слов в темах из-за небольшого числа наблюдений, поэтому при построении модельной коллекции в коротких документах заведомо сильнее нарушается гипотеза условной независимости. Также на этом графике заметен эффект перерегуляризации матрицы Θ: после десятой итерации EM-алгоритма значения адекватности документов заметно уменьшаются, что указывает на необходимость ослабления или выключения регуляризации.

На графике рис. 2 показана зависимость адекватности темы от частоты её слов n_t на различных итерациях EM-алгоритма. Из графика видно, что, во-первых, темы расслаиваются по мощностям не сразу, а только после 10-й итерации, а во-вторых, адекватность темы не зависит от её частоты в коллекции.

На графиках рис. 4, 5 показаны значения адекватности и взвешенной адекватности тем на различных итерациях EM-алгоритма. Темы упорядочены по возрастанию значения адекватности (взвешенной адекватности, соответственно). На первом графике заметно монотонное быстрое улучшение адекватности всех тем по мере обучения тематической модели. На втором графике заметен эффект перерегуляризации: взвешенная

адекватность тем сначала возрастает, а затем резко убывает для некоторых тем. Это говорит о том, что значения θ_{td}, непосредственно от которых зависит взвешенная адекватность, сильнее всего занижены из-за чрезмерного разреживания именно для этих тем.

Результат эксперимента:

- Тест адекватности тем является очень чувствительным к изменениям модели: даже на первых итерациях заметно значительное изменение распределения значений адекватности тем.
- Тест адекватности документов также является очень чувствительным к изменениям модели.
- При нарушении гипотезы условной независимости тест адекватности документа наиболее чувствителен к длинным документам, для слишком коротких документов он может быть неинформативен.
- Адекватность темы не зависит от частоты её вхождений в коллекцию.
- Адекватность документов и взвешенная адекватность тем позволяют обнаруживать эффект перерегуляризации матрицы Θ.

5.2 Исследование мощности введённых метрик качества для онлайнового и оффлайнового EM-алгоритмов

Цель эксперимента: определение чувствительности критериев к нарушениям гипотезы условной независимости.

Ход эксперимента: эксперимент проводился на полумодельных данных. Матрицы Φ_m и Θ_m были получены в результате решения задачи тематического моделирования для реальной коллекции документов. Генерация коллекций n_{dw}^b для каждого $b \in \{0, 0.5, 0.8, 1\}$ производилась из вероятностной смеси распределений

$$bp_0(w|d) + (1-b)p_1(w \mid d),$$

где $p_0(w \mid d) = \sum\limits_{t \in T} \phi_{wt}\theta_{td}$, а $p_1(w \mid d) = \sum\limits_{t \in T} p(w \mid d, t)p(t \mid d)$, $p(w \mid d, t)$ - специфичное для каждого документа d распределение, принимающее ненулевые значения на уникальном для этого документа множестве слов. При $b = 0$ гипотеза условной независимости

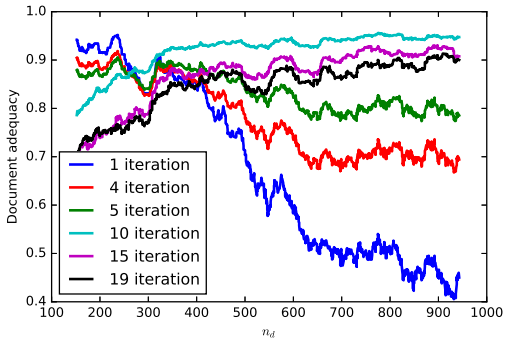

Рис. 1: Зависимость адекватности документа от его длины.

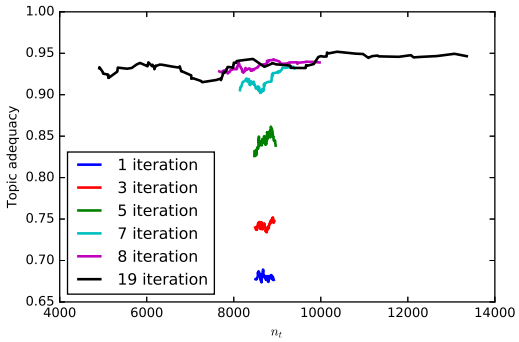

Рис. 2: Зависимость адекватности темы от её частоты.

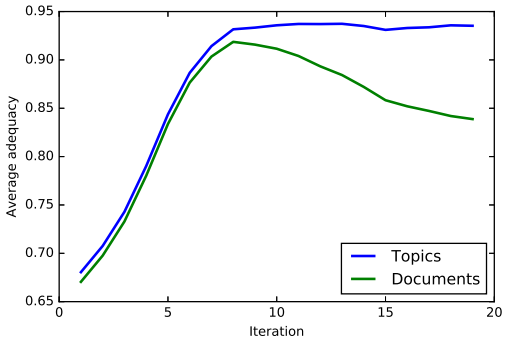

Рис. 3: Зависимость средней адекватности от номера итерации EM-алгоритма.

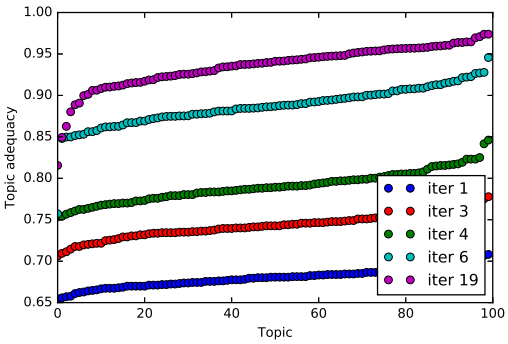

Рис. 4: Адекватность тем на различных итерациях.

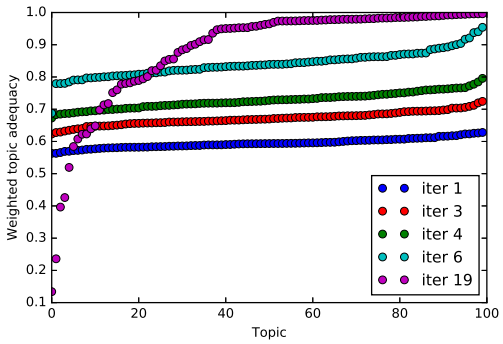

Рис. 5: Взвешенная адекватность тем на различных итерациях.

не выполнена для каждой темы, при $b = 1$ выполнена с точностью до погрешности из-за сэмплирования. Для онлайнового алгоритма аналогичные коллекции были получены как результат пятнадцатикратного дублирования каждой из описанных выше коллекций n_{dw}^b, что нарушило уникальность документов: гипотеза условной независимости в коллекции для онлайнового алгоритма заведомо выполнена в большей степени.

Для каждой такой коллекции проводилось 25 шагов EM-алгоритма. Для результирующих матриц Φ, Θ вычислялись статистики $CR(d,t)$, $CR'(d,t)$, восстанавливалась

зависимость 0.95 квантили от числа вхождений темы в документ и для каждой темы определялись документы, для которых значение статистики $CR(d,t)$ превосходило квантиль распределения статистики для данного числа n_{td} вхождений темы в документ.

Исследовалась зависимость характеристики адекватности темы от степени выполнения гипотезы условной независимости b.

На рис. 6 показаны изменения распределения значений адекватности тем в зависимости от степени выполнения гипотезы условной независимости. Темы упорядочены по возрастанию значения адекватности. На графиках заметно, что при нарушении гипотезы условной независимости тест адекватности темы отвергает гипотезу для всех или почти всех тем. По мере уменьшения «загрязнённости» коллекции адекватность тем возрастает, а при выполнении гипотезы условной независимости тест не отвергает гипотезу для всех тем на уровне значимости $0.05 - 0.06$.

Результат эксперимента:

- Тест адекватности тем является очень чувствительным к нарушениям гипотезы условной независимости: даже при небольшой «загрязнённости» коллекции заметно значительное ухудшение распределения значений адекватности тем.
- Тест адекватности тем является информативной характеристикой нарушения гипотезы условной независимости: результаты принимают значение от 0 до 1 и отлично интерпретируются.
- Введённые метрики качества обычно являются более мощными в оффлайновом алгоритме, что объясняется точной формулой вычисления статистики Кресси-Рида в оффлайновом алгоритме и приближённой в онлайновом, а также заведомо меньшим нарушением гипотезы условной независимости в коллекции для онлайнового алгоритма.

5.3 Проверка необходимости хранения счётчиков n'_{wt}

Цель эксперимента: проверить необходимость хранения счётчиков n'_{wt}, установить, что вычисление статистики $CR_\lambda(t,d)$ без них невозможно. Эксперимент имеет важное практическое значение, так как хранение счётчиков n'_{wt} для искусственной коллекции требует довольно большого объёма памяти.

Ход эксперимента: эксперимент проводился на реальных данных. Была использована коллекция текстовых аннотаций к статьям по вычислительной биологии и био-

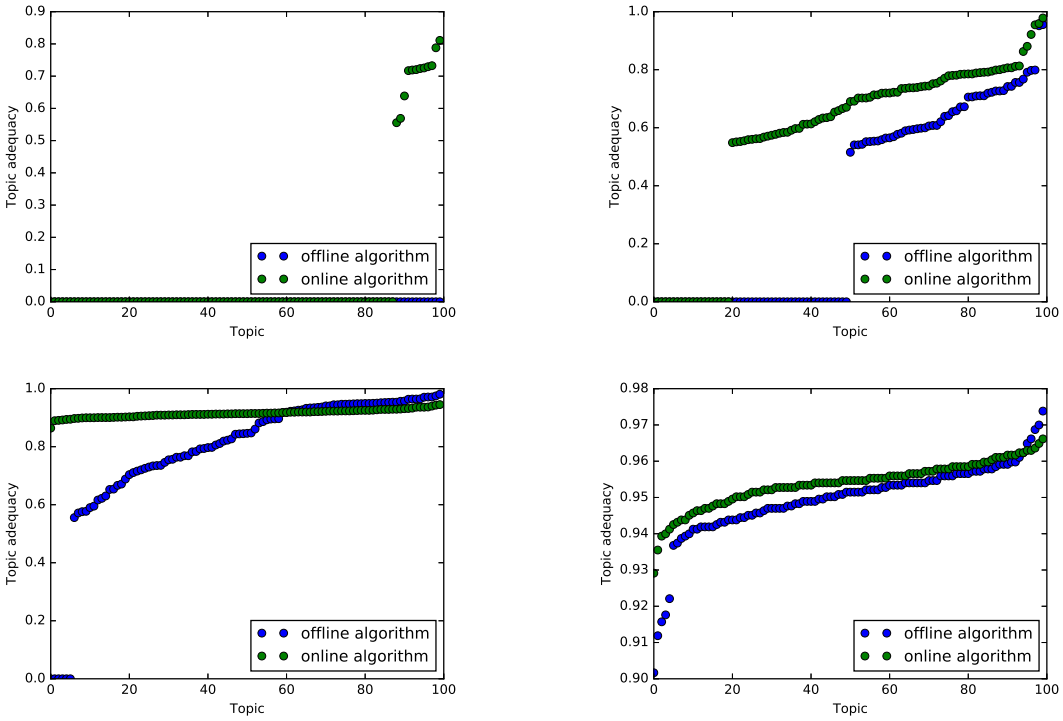

Рис. 6: Значения адекватности тем, упорядоченные по неубыванию, для $b = 0, 0.5, 0.8, 1$.

информатике Pubmed с $|D| = 8200000$, $|W| = 141043$, размер батча полагался равным 200, число тем $|T|$ полагалось равным 40. Использовался регуляризатор разреживания матрицы Θ с коэффициентом регуляризации -0.15.

После каждого второго батча для текущих матриц Φ, Θ вычислялись статистики $CR(d,t)$, $CR'(d,t)$.

Проверялось предположение, что в выражении (10) можно заменить n'_{wt} и n'_t на n_{wt} и n_t, соответственно. Для проверки этого предположения после каждого второго батча также вычислялось такое выражение:

$$\hat{CR}'_\lambda(d,t) = \frac{2}{\lambda(\lambda+1)} \sum_{w \in d} \tilde{n}'_{dwt} \left(\left(\frac{\tilde{n}'_{dwt} n_t}{\tilde{n}'_{dt} n_{wt}} \right)^\lambda - 1 \right). \tag{12}$$

На основе двух полученных наборов статистик $CR'(d,t)$ и $\hat{CR}'_\lambda(d,t)$ восстанавливалась зависимость 0.95 квантили от числа вхождений темы в документ и для каждой темы определялись документы, для которых значение статистики $CR(d,t)$ превосходило квантиль распределения статистики для данного числа n_{td} вхождений темы в документ. Затем вычислялись значения адекватности и взвешенной адекватности для каждой темы и документа для двух вариантов статистик, вычисленных по искусственной коллекции.

На рис. 7 показаны средние значения адекватности и взвешенной адекватности в зависимости от номера батча для двух вариантов вычисления статистики по искус-

ственной коллекции. Из графиков видно, что значения адекватности, вычисленные на основе статистики $\hat{CR}'_\lambda(d,t)$, оказываются значительно ниже соответствующих значений, вычисленных на основе статистики $CR'(d,t)$. Более того, характер зависимости средней адекватности от номера батча существенно различается. Следовательно, такой вариант приближённого вычисления статистики Кресси-Рида для искусственной коллекции является неприемлемым.

Результат эксперимента: предположение о возможности вычисления статистики Кресси-Рида для искусственной коллекции на основе счётчиков n_{wt} и n_t оказалось ошибочным; нельзя сократить расход памяти за счёт отказа от хранения счётчиков n'_{wt} и n'_t.

5.4 Проверка возможности определения количества тем на основе адекватности тем

Цель эксперимента: проверить возможность определения количества тем в коллекции документов на основе значений адекватности тем.

Ход эксперимента: эксперимент проводился на реальных данных. Была использована коллекция текстовых аннотаций к статьям по вычислительной биологии и биоинформатике Pubmed с $|D| = 8200000$, $|W| = 141043$, размер батча полагался равным 200. Использовался регуляризатор разреживания матрицы Θ с коэффициентом регуляризации -0.15. На этой коллекции обучались тематические модели с числом тем $|T|$, равным 10, 40 и 60. После обучения каждой тематической модели вычислялась взвешенная адекватность всех полученных тем.

На рис. 8 показаны значения взвешенной адекватности тем для моделей с разным количеством тем. Темы упорядочены по неубыванию значения взвешенной адекватности. Из графиков видно, что взвешенная адекватность тем в модели с 10 темами существенно ниже, нежели в модели с 40 темами, а при увеличении количества тем с 40 до 60 существенного увеличения взвешенной адекватности не происходит, что согласуется с результатами определения числа тем в данной коллекции, полученными на основе скорости изменения перплексии [17] и проверенными с помощью оценки интерпретируемости полученных тем.

Результат эксперимента: тест взвешенной адекватности модели позволяет выбирать среди нескольких моделей ту, число тем в которой ближе к числу тем в коллекции. Это открывает широкие возможности для его применения совместно с регуляризатором разреживания тем для автоматического определения числа тем в коллекции докумен-

тов.

6 Результаты, выносимые на защиту

В данной работе предложен простой и эффективный тест адекватности тематической модели, основанный на проверке гипотезы условной независимости с помощью эмпирического распределения статистики Кресси-Рида. Также в работе предложен способ встраивания этого метода в онлайновый EM-алгоритм с учётом его особенностей, не приводящий к существенным расходам памяти и времени. Было рассмотрено несколько метрик качества тематической модели на основе статистики Кресси-Рида. Экспериментально показана информативность полученных метрик.

Список литературы

[1] *D. Mimno and D. Blei*, Bayesian checking for topic models // *Proceedings of the Conference on Empirical Methods in Natural Language Proceedings.*– 2011.– Pp. 262–272.

[2] *Воронцов, К. В.*, Аддитивная регуляризация тематических моделей коллекций текстовых документов // *Доклады РАН, т.456, с.268–271.*– 2014.

[3] *Hofmann, Thomas*, Probabilistic latent semantic indexing // *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp50–57.*– 1999.

[4] *Воронцов, К. В. and Потапенко, А. А.*, Модификации EM-алгоритма для вероятностного тематического моделирования // *Машинное обучение и анализ данных, т.1, №6, с.657–686,.*– 2013.

[5] *Hanna Wallach and Iain Murrey and Ruslan Salakhutdinov and David Mimno*, Evaluation methods for topic models // *ICML.*– 2009.

[6] *M. A. Taddy*, On estimation and selection for topic models // *arXiv preprint.*– 2011.

[7] *D. Mimno and H. M. Wallach and E. Talley and M. Leenders and A. McCallum*, Optimizing semantic coherence in topic models // *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*– 2011.– Pp. 262–272.

[8] *M. E. Roberts and B. M. Stewart and D. Tingley and C. Lucas and J. Leder-Luis and S. Gadarian and B. Albertson and D. Rand*, Structural topic models for open-ended survey responces // *American Journal of Political Science.*– 2014.

[9] *J. Chang and J. Boyd-Graber and C. Wang and S. Gerrich and D. M. Blei*, Reading tea leaves: How human interpret topic models // *NIPS.*– 2009.

[10] *Целых В. Р., Воронцов К. В.*, Критерии согласия для разреженных распределений и их применение в тематическом моделировании // *Машинное обучение и анализ данных.*– 2012.– С. 437–447.

[11] *Cressie, N. and Read, T. R. C.*, Multinomial Goodness-of-fit Tests // *Journal of the Royal Statistical Society, Series B.*– Pp. 440–464, 1984

[12] *David Zelterman*, Goodness-of-fit tests for large sparse multinomial distributions // *Journal of the American Statistical Association.*– 1987.– Pp. 624–629.

[13] *von Davier, Matthias*, Bootstrapping Goodness-of-Fit Statistics for Sparse Categorical Data — Results of a Monte Carlo Study // *Methods of Psychological Research Online.*– 1997.

[14] *Тихонов, А. Н. и Арсенин, В. Я.* , Методы решения некорректных задач // 1986.

[15] *Walker, A. J.*, New fast method for generating discrete random numbers with arbitrary frequency distributions // *Electronics Letters.*– 1974.

[16] *Marsaglia, George; Tsang, Wai Wan; Wang, Jingbo*, Fast Generation of Discrete Random Variables// *Journal of Statistical Software.*– 2004.

[17] *Weizhong Zhao, James J Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding and Wen Zou*, A heuristic approach to determine an appropriate number of topics in topic modeling// *BMC Bioinformatics.*– 2015.

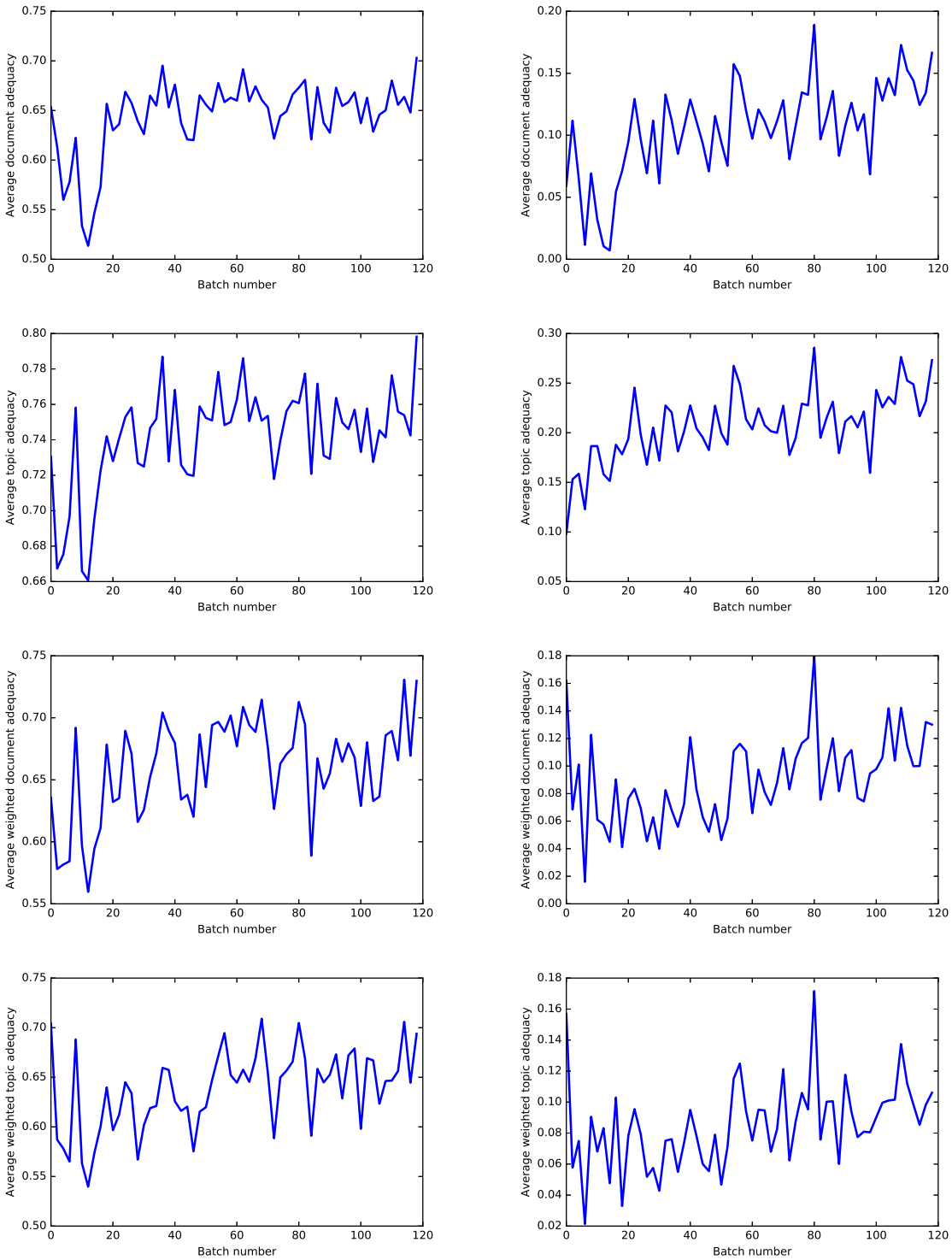

Рис. 7: Средняя адекватность и взвешенная адекватность тем и документов в зависимости от номера батча. В левом столбце значения, вычисленные на основе счётчиков n'_{wt}, в правом — на основе n_{wt}.

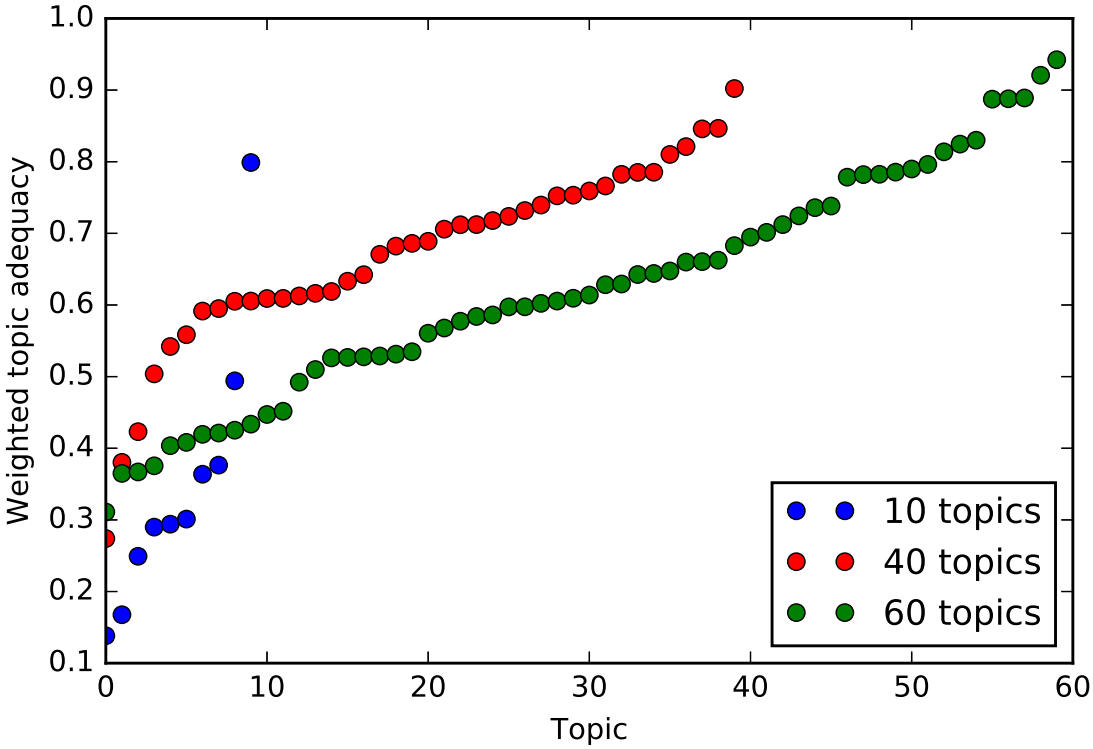

Рис. 8: Взвешенная адекватность тем для моделей с разным числом тем.