Московский физико-технический институт
(Государственный университет)

Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 974 ГРУППЫ

«Построение тематической классификации коллекции документов с неизвестным числом тем»

Выполнил:
студент 4 курса 974 группы
Лобастов Степан Юрьевич

Научный руководитель:
д.ф.-м.н., н.с. ВЦ РАН
Воронцов Константин Вячеславович

Москва, 2013

Содержание

Аннотация

Данная работа посвящена решению задачи тематической кластеризации текстов, то есть разбиению текстов из некоторой коллекции документов на темы. Для классификации используются вероятностные модели, которые позволяют автоматически определить число тем в коллекции. Для их построения используются различные эвристики. В работе исследуется модель иерархического процесса Дирихле(HDP) и предлагаются методы ее улучшения.

Ключевые слова: иерархический процесс Дирихле, вероятностные тематические модели, байесовсвие сети, семплирование по Гиббсу

1 Введение

Данная работа посвящена решению задачи тематической кластеризации текстов, то есть разбиению текстов из некоторой коллекции документов на темы. Тема представляет собой множество документов, обладающих терминологическим сходством. Под тематической кластеризацией подразумевается, что в полученном разбиении документы в каждом кластере можно отнести к одной и той же теме. Предложенные методы, решающие эту задачу, имеет свои достоинства и недостатки. К достоинствам можно отнести, например, точность получаемой кластеризации(проверяемой на контрольной выборке), быстродействие, простоту модели, теоретическое обоснование используемых эвристик или автоматическое определение числа тем. В работе исследован алгоритм, построенный на основе существующих, который использует их сильные стороны и компенсирует недостатки.

Приведем краткий обзор направлений, в которых развиваются сегодня методы, используемые для построения тематической модели.

Первый подход к решению задачи основывается на использовании представлении матрицы слово-документ в виде произведения двух матриц с неотрицательными элементами. В работе [1] был пред-

ложен простой алгоритм для вычисления такого представления, также доказывается теорема о локальной оптимальности получаемого решения.

Второй подход использует сингулярное разложение матрицы слово-документ, благодаря чему появляется возможность кластеризации её строк, то есть описаний документов. Впервые он был оформлен в работе [2], и получил развитие в [3] и [4]. В [3] было доказано, что он нечувствителен к шумовым компонентам текстов, а в работе [4] обосновывается возможность учета им синонимичных слов.

Третий рассматриваемый в работе подход основывается на предположении о вероятностном характере коллекции документов [5]. Предполагается, что существует распределение слов в документах и темах, а данная коллекция документов - это реализация выборки из этого распределения. Соответствующая этому предположению постановка задачи и краткое описание метода ее решения, основанного на использовании скрытого распределения Дирихле(Latent Dirichlet Allocation), описываются в работе [6]. Тем же автором описываются возможные модификации этого алгоритма, использующие оптимизацию гиперпараметров распределения Дирихле, возможную корреляцию между темами коллекции (CTM) [7], а также онлайн-модификация метода LDA [8]. Кроме того, на модели LDA основывается динамической тематической модели (DTM) [9] [10]. Она используется при работе с коллекциями, полученными в течении длительного периода времени (несколько десятилетий) и учитывает изменение тем на протяжении этого времени.

В статье [5] также приводятся результаты работы алгоритма LDA — тематическая кластеризация 16000 научных статей. Общие выводы — методы, использующие подход, проще модифицировать, чем методы, основанные на неотрицательном матричном разложении [11] [1] или сингулярном разложении [2].

В [12] приводится сравнение эвристик, которые можно применять к модели LDA. Они применяются, например, для анализа зашум-

ленной коллекции документов путем добавления к модели шумовой и фоновой компонент [13], а также используют различные аппроксимации распределений слов [14] и оптимизацию гиперпараметров распределения Дирихле [15] и другие.

Однако, все эти подходы имеют существенный недостаток — они не позволяют автоматически определять число тем коллекции, оно считается известным априорно. В реальности же оно, как правило, не известно. Для преодоления этого недостатка можно использовать, например, модель иерархичесского процесса Дирихле (HDP), описанного в [16]. Эта модель является модификацией модели LDA, и так же использует распределение Дирихле [17] и процесс китайского ресторана [18] для описания процесса порождения документов. Число тем в коллекции документов здесь определяется автоматически также в процессе семплирования по Гиббсу. В [16] приводится реализация модели HDP, а также результаты проведенных экспериментов. В [19] приводится более эффективная реализация HDP.

Так как алгоритм HDP имеет вероятностную природу, результат его работы является неустойчивым [16]. Кроме того, как показано в настоящей работе, этот алгоритм существенно зависит от входных параметров. В работе исследована сходимость алгоритма и предприняты попытки улучшить его устойчивость.

2 Постановка задачи

В наиболее общем виде задача тематической классификации текстов выглядит следующим образом.

«Дан словарь(множество всех слов коллекции) W и коллекция документов $\mathcal{D}$. Необходимо разбить документы коллекции на кластеры, число которых неизвестно. Под тематической кластеризацией подразумевается, что в полученном разбиении документы в каждом кластере относятся к одной и той же теме. Документы внутри темы имеют тер-

минологическое сходство»

Такую постановку задачи можно формализовать по-разному, чем и отличаются описанные выше методы. Вот как это происходит при использовании вероятностного подхода к задаче.

Вводится матрица слово-документ D, такая, что $D = \|D_{dw}\|$, где D_{dw} — количество слов w в документе d. Считается, что в качестве входных данных используется эта матрица, а не вся коллекция документов $\mathcal{D}$.

Вероятностный подход опирается на ряд предположений о коллекции документов. Рассмотрим их:

1. Гипотеза о существовании распределения, из которого была порождена коллекция. Предполагается, что существует множество тем $\mathcal{T}$, а также распределение слов в документах $p(d, w, t)$ на $\mathcal{D} \times W \times \mathcal{T}$, а наблюдаемая коллекция документов - это реализация некоторой выборки из этого распределения. Введем, кроме того, обозначения некоторых условных вероятностей:

$$p(w|t) = \phi_{wt}, \quad p(t|d) = \theta_{td} \quad \forall t \in \mathcal{T}, d \in \mathcal{D}.$$

   Именно поэтому подход и получил название «вероятностный»

2. Гипотеза об условной независимости. Считается, что распределение слов в темах не зависит от документа, а зависит только от темы. Иными словами,

$$p(w|t, d) = p(w|t) = \phi_{wt}.$$

3. Гипотезы «мешка слов» и «мешка документов». Предполагается, что при построении модели не важен ни порядок слов в документах, ни порядок документов в коллекции. То распределения всех слов коллекции одинаковы, а сами слова независимо

распределены. Каждое слово коллекции получено из распределения

$$p(w|d) = \sum_{t \in \mathcal{T}} p(w|d,t)p(t|d) = \sum_{t \in \mathcal{T}} p(w|t)p(t|d) = \sum_{t \in \mathcal{T}} \phi_{wt}\theta_{td}.$$

Именно значения ϕ_{wt} и θ_{td} нам и предстоит найти, поскольку они описывают, с какой вероятностью той или иной документ относится к той или иной теме.

В условиях истинности этой гипотезы, как нетрудно видеть, в матрице D содержится вся информация о коллекции документов $\mathcal{D}$, поэтому в дальнейшем будем считать, что $D \equiv \mathcal{D}$, и называть матрицу D коллекцией документов.

Для левой части этого равенства можно построить эмпирическую оценку

Основываясь на этих предположениях, а также введя обозначения
$F = \|p(w|d)\|_{w \in W, d \in \mathcal{D}}$,
$\Phi = \|p(w|t)\|_{w \in W, t \in \mathcal{T}}$,
$\Theta = \|p(t|d)\|_{k \in \mathcal{K}, d \in \mathcal{D}}$,
можно рассматривать нашу задачу как задачу разложения матрицы F в произведение матриц Φ и Θ, таких, что все элементы в них неотрицательны, а сумма элементов в каждой строке равна единице (т.к. эти элементы являются вероятностями):

$$F = \Phi \cdot \Theta$$

Таким образом, под построением тематической модели понимается нахождение неизвестных распределений, стоящих в правой части равенства, то есть матриц Φ и Θ.

Для оценки этих матриц максимизируется функция правдоподобия коллекции документов:

$$(\Phi,\Theta) = \underset{\substack{\Phi\in Mat_{\|\mathcal{T}\|\times\|W\|}\\ \Theta\in Mat_{\|\mathcal{D}\|\times\|\mathcal{T}\|}}}{\operatorname{argmax}} \prod_{d\in\mathcal{D},w\in W} p(w|d)^{D_{wd}} =$$

$$= \underset{\substack{\Phi\in Mat_{\|\mathcal{T}\|\times\|W\|}\\ \Theta\in Mat_{\|\mathcal{D}\|\times\|\mathcal{T}\|}}}{\operatorname{argmax}} \prod_{d\in\mathcal{D},w\in W} \Big(\sum_{t\in\mathcal{T}} \phi_{wt}\theta_{td}\Big)^{D_{wd}} =$$

$$= \underset{\substack{\Phi\in Mat_{\|\mathcal{T}\|\times\|W\|}\\ \Theta\in Mat_{\|\mathcal{D}\|\times\|\mathcal{T}\|}}}{\operatorname{argmax}} \sum_{d\in\mathcal{D},w\in W} D_{wd}\log\Big(\sum_{k\in\mathcal{K}} \phi_{wt}\theta_{kt}\Big) \quad (1)$$

Для сравнения качества построенных вероятностных моделей используется перплексия. Для её оценки каждый документ коллекции D случайным образом разбивается на две части D_1 и D_2, первая часть используется для обучения модели (нахождения профилей тем $\overrightarrow{\phi_k}$), а для второй вычисляется значение $P(D)$:

$$P(D) = \exp\Big(-\frac{1}{n}\sum_{d\in\mathcal{D},w\in D_2} D_{wd}\log p(w|d)\Big).$$

Чем меньше перплексия, тем лучше модель описывает коллекцию документов.

3 Семплирование

При построении тематической классификации берется за основу предположение о том, что коллекция документов была получена из той или иной вероятностной порождающей модели. Построение классификации при этом сводится к поиску параметров этой модели. Один из способов нахождения этих параметров — алгоритм семплирования. Опишем его в общем виде.

Обозначим множество неизвестных параметров за A. В простейшем случае A — это темы документов коллекции.

- 1. Выбирается некоторое начальное приближение для неизвестных параметров A.
- 2. Фиксируются все параметры, кроме одного ($a \in A$).
- 3. Вычисляются вероятности всех возможных значений a_i параметра a при условии истинности всех остальных параметров:

$$P(a = a_i | A \backslash a) = \frac{P(A \backslash a, a = a_i)}{P(A \backslash a)} \propto P(A \backslash a, a = a_i),$$

  поскольку $P(A \backslash a)$ не зависит от a. Вероятность $P(A \backslash a, a = a_i)$ оценивается по коллекции документов.
- 4. Процесс повторяется для всех $a \in A$, пока значения всех элементов A не стабилизируются.

4 Распределение Дирихле

Распределение Дирихле, наряду с семплированием, также часто используется при построении вероятностных моделей. Например, удобно полагать, что параметры мультиномиальных распределения слов в темах были получены из симметричного распределения Дирихле. Это обусловлено следующими его свойствами:

- Плотность такого распределения равна

$$f(\overrightarrow{\phi}_t) = \frac{1}{B(\alpha)} \prod_{w \in W} \phi_{wt}^{\alpha - 1}$$

- $\sum_{w \in W} \phi_{wt} = 1$
- $\phi_{wt} \geqslant 0 \quad \forall w \in W$

Здесь $B(\alpha)$ — нормирующий множитель:

$$B(\alpha) = \frac{\Gamma(\alpha|W|)}{(\Gamma(\alpha))^{|W|}}$$

Именно поэтому $\overrightarrow{\phi}_t$ можно использовать в качестве параметров мультиномиального распределения.

5 Латентное размещение Дирихле (LDA)

Один из методов построения тематической модели носит название LDA(Latent Dirichlet Allocation). В качестве порождающей модели он использует следующие предположения. Модель предполагает, что каждое слово в коллекции относится ровно к одной теме. Кроме того, считается, что параметры мультиномиальных распределений $\overrightarrow{\phi}_t = \|\phi_{wt}\|_{w \in W}$ и $\overrightarrow{\theta}_d = \|\theta_{dt}\|_{t \in \mathcal{T}}$ слов в темах и тем в документах были получены из симметричного распределения Дирихле с некоторыми параметрами α и β соответственно:

$$\overrightarrow{\phi_t} \in Dir(\alpha), \quad \overrightarrow{\theta_d} \in Dir(\beta).$$

Введем обозначения: x_{di} — i-е слово в документе d
t_{di} — тема i-го слова в документе d
$T = \|t_{di}\|$
$X = \|x_{di}\|$
n_{dt} — количество слов в документе d, принадлежащих теме t
$n_{\cdot t}$ — общее число слов в коллекции, принадлежащих теме t
$n_{d\cdot}$ — общее число слов в документе d
c_{wt} — количество слов w теме t
$c_{\cdot t} = n_{\cdot t}$

В процессе семплирования мы будем находить темы для слов из коллекции. Для семплирования i-го слова в d-м документе необхо-

димо найти вероятности

$$P(t_{di} = t|T\backslash t_{di}, X) \quad \forall t \in \mathcal{T}.$$

Воспользовавшись формулой Байеса и отбросив не зависящие от t параметры, получим:

$$P(t_{di} = t|T\backslash t_{di}, X) \propto P(t_{di} = t|T\backslash t_{di}, X\backslash x_{di}) \cdot P(x_{di}|K, X\backslash x_{ji})$$

Пусть слово $x_{di} = w$ принадлежало теме $t_{di} = t$. Для оценки этих вероятностей пересчитаем значения параметров, описывающих коллекцию документов, с учетом исключения из нее слова x_{di}:

$n_{dt} := n_{dt} - 1$;
$n_{d\cdot} := n_{d\cdot} - 1$;
$n_{\cdot t} := n_{\cdot t} - 1$;
$c_{wt} := c_{wt} - 1$;
$c_{\cdot t} := c_{\cdot t} - 1$.
Согласно [6],

$$P(t_{di} = d|T\backslash t_{di}, X\backslash x_{di}) = \frac{n_{dt} + \beta}{n_{d\cdot} + \beta|\mathcal{T}|}$$

И, если $t_{di} = t$, а $x_{di} = w$,

$$P(x_{di}|T, X\backslash x_{ji}) = \frac{c_{wt} + \alpha}{c_{\cdot t} + \alpha|W|}$$

Таким образом, при семплировании темы для слова x_{di} с вероятностью

$$P(t_{di} = t|T\backslash t_{di}, X) \propto \frac{n_{dt} + \beta}{n_{d\cdot} + \beta|\mathcal{T}|} \cdot \frac{c_{wt} + \alpha}{c_{\cdot t} + \alpha|W|} \tag{2}$$

будет выбрана тема t. Реализация процесса построения модели LDA описана в 1. Мы видим, что модель LDA, несмотря на простоту

реализации, не дает возможности определять число тем в коллекции, а может лишь принимать это значение в качестве входного параметра. Конечно, возможно искать это значение путем полного перебора, но в этом случае сложность и время выполнения алгоритма существенно увеличатся.

Data: matrix D word-document, number of topics $|\mathcal{T}|$
Result: matrix T
Select first approximation of T;
Compute X, c, n according to definition;
while T *not stabilize* **do**
  foreach *document* $d \in \mathcal{D}$ **do**
    foreach *word* x_{di} *in document* d **do**
      $n_{dt} - -; n_{d\cdot} - -;$
      $c_{wt} - -; c_{\cdot t} - -;$
      compute priors $p(t)$ according to (2) ;
      select k from computed priors;
      $n_{dt} + +; n_{d\cdot} + +;$
      $c_{wt} + +; c_{\cdot t} + +;$
    end
  end
end

Algorithm 1: LDA sampling process

Как мы видим, модель LDA не может автоматически определять число тем в коллекции документов. Более того, даже перебор всех чисел тем не решит эту проблему, поскольку это очень ресурсоемкий (вычислительно) процесс. Кроме того, как показывают исследования [16] зависимости перплексии от числа тем, заданных на вход LDA, при различных значениях числа тем(больших, чем реальное) перплексия остается практически неизменной, и ее оказывается недостаточно для того, чтобы выбрать правильное число тем.

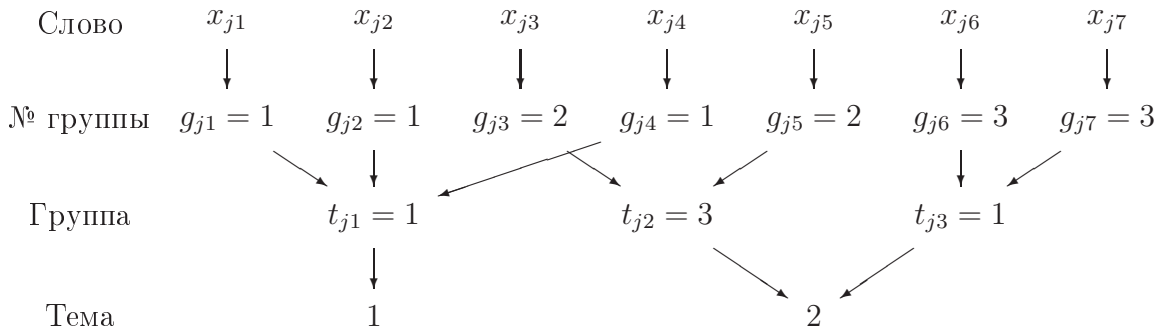

Рис. 1: Представление i-го документа коллекции

6 Иерархический процесс Дирихле (HDP)

Модель HDP (Hierarchical Dirichlet Process) является усложнением модели LDA. Эта модель способна автоматически определять число тем в коллекции. Опишем ее.

Как и в модели LDA, в HDP предполагается, что параметры мультиномиальных распределений слов в темах были получены из распределения Дирихле с параметром α. Кроме того, согласно этой модели, слова внутри документов разбиты на группы. При этом все слова внутри одной группы относятся к одной теме, а при генерации документов для слова сначала выбирается группа, а его тема соответствует теме, к которой относятся все слова в выбранной группе. Такая модель позволяет определять число тем автоматически, и, кроме того, позволяет добиться существенной разреженности распределений слов в темах и тем в документах. Для того, чтобы описать эту модель подробнее, введем обозначения.

Обозначим g_{di} номер группы i-го слова в документе d, а за $G = \bigcup_{d\in\mathcal{D}} g_{di}$ — множество всех групп всех слов. Пример представления документа коллекции в указанном виде приведен на **??**. Введем также обозначения:

$N = |\mathcal{T}|$ — общее количество тем в документах.

n_{dgt} — число слов в группе g документа d в теме t. $n_{dgt} \neq 0 \Leftrightarrow t = t_{dg}$

$n_{dg\cdot}$ — число слов в d-м документе в g-й группе слов

$n_{dg\cdot}$ — число слов в d-м документе, принадлежащих группе g

m_{dt} - число групп слов в документе d с темой t

$m_{d\cdot}$ - число групп слов в документе d

$m_{\cdot t}$ - число групп слов с темой t

$m_{\cdot\cdot}$ - общее число групп слов во всех документах

Опишем саму модель. Слова поступают в документы по очереди. Рассмотрим момент поступления i-го слова в документ d. Сначала для этого слова определяется группа слов следующим образом:

С вероятностью, пропорциональной $n_{dg\cdot}$ слово относится к группе g (для всех g от 1 до $m_{d\cdot}$). В этом случае тема слова — это тема группы слов, к которой слово отнесено. Само слово генерируется из распределения слов в этой теме.

С вероятностью, пропорциональной β для слова создается новая, $m_{d\cdot}+1$-я группа слов. Теперь для этой группы слов необходимо определить тему. Это осуществляется следующим образом:

С вероятностью, пропорциональной $m_{\cdot t}$ тема новой группы — t (для всех t от 1 до N).

С вероятностью, пропорциональной γ, тема новой группы новая для всей коллекции документов. В этом случае распределение слов в ней семплируется из распределения Дирихле с параметром α:

После того, как тема новой группы слов (пока состоящей из одного слова) определена, само слово семплиуется из распределения слов в этой теме.

Суть этой модели в том, что при появлении нового слова в документе, это слово с большей вероятностью будет относиться к темам, уже представленным в документе большим числом слов, однако существует ненулевая вероятность появления слова из темы, представленной малым числом слов, или вообще не представленной

Входные параметры:α, β, γ

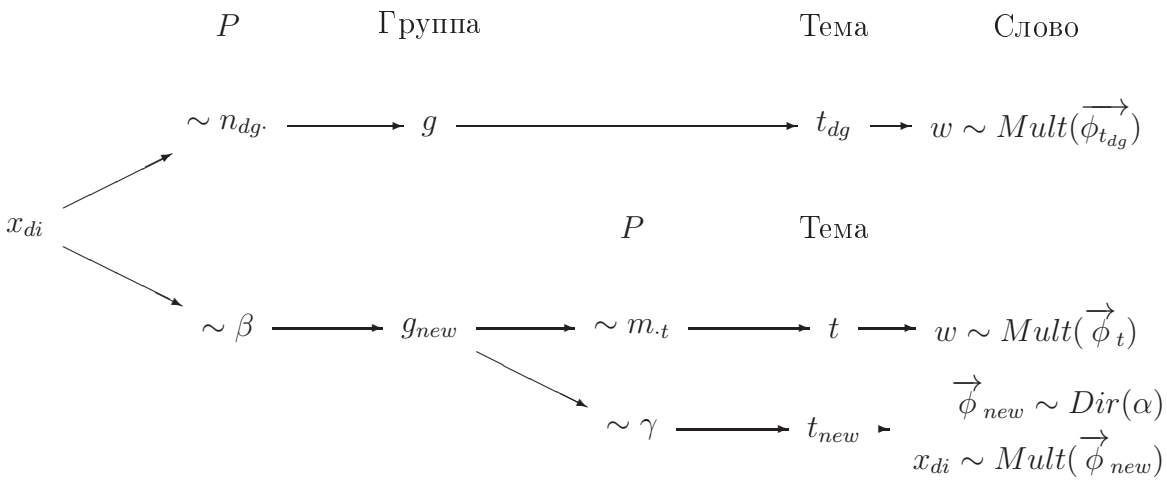

Рис. 2: Схема порождающей модели

в документе.

Эта порождающая модель описывает процесс генерации коллекции документов с неизвестным заранее числом тем. При этом матрица документ-тема Θ оказывается разреженной, в отличии от модели LDA, что лучше описывает реальные коллекции документов, когда один документ принадлежит, как правило, не к всем темам, а лишь к небольшому их подмножеству.

7 Семплирование

В предположении описанной порождающей модели, задача тематической классификации коллекции документов сводится к нахождению групп для всех слов (параметров g_{di}) и тем для всех групп слов (параметров t_{gj}).

В данном случае $A = G \cup T$. Для получения алгоритма семплирования необходимо вычислить $P(A, a = a_i)$. Для этого докажем две леммы.

Лемма 1. Пусть имеется выборка A из мультиномиального распределения, параметры которого получены из распределения Дирихле с известным параметром α. Кроме того, имеется выборка B, состоящая из c одинаковых элементов. Пусть мультиномиальная случайная величина принимает значения $1, \ldots, k$, а в выборке A значение 1 реализовалось n_1 раз, 2 — n_2 раз, ..., k — n_k раз. Выборка B состоит из одинаковых значений t. Тогда вероятность того, что она получена из того же мультиномиального распределения, равна

$$\frac{(\alpha + n_t)_c}{(k\alpha + \sum_{i=1}^{k} n_i)_c},$$

где

$$(\alpha + n_t)_c = (\alpha + n_t)(\alpha + n_t + 1) \cdot \cdots \cdot (\alpha + n_t + c - 1),$$

$$(k\alpha + \sum_{i=1}^{k} n_i)_c = (\alpha + \sum_{i=1}^{k} n_i)(\alpha + \sum_{i=1}^{k} n_i + 1) \cdot \dots \cdot (\alpha + \sum_{i=1}^{k} n_i + c - 1)$$

Доказательство: Необходимо найти вероятность $P(B|A, \alpha)$.

$$P(B|A, \alpha) = \frac{P(B, n_1, n_2, \dots, n_k, |\alpha)}{P(n_1, n_2, \dots, n_k, |\alpha)}$$

Так как параметр мультиномиального распределения $\overrightarrow{\phi}$ получен из распределения Дирихле с параметром α, то

$$P(n_1, n_2, \dots, n_k, \alpha) = \int P(n_1, n_2, \dots, n_k, |\alpha, \overrightarrow{\phi}) p(\overrightarrow{\phi}|\alpha) d\overrightarrow{\phi} =$$

$$= \int (\phi_1^{n_1} \phi_2^{n_2} \cdot \dots \cdot \phi_k^{n_k}) \Big(\frac{1}{B(\alpha, \dots, \alpha)} \prod_{i=1}^{k} \phi_i^{\alpha - 1} d\phi_i \Big) =$$

$$= \frac{1}{B(\alpha, \dots, \alpha)} \int \prod_{i=1}^{k} (\phi_i^{n_i + \alpha - 1} d\phi_i) =$$

$$= \frac{B(\alpha + n_1, \alpha + n_2, \dots, \alpha + n_k)}{B(\alpha, \alpha, \dots, \alpha)} \quad (3)$$

Значит,

$$P(B|n_1, n_2, \dots, n_k, \alpha) = \frac{P(B, n_1, n_2, \dots, n_k, |\alpha)}{P(n_1, n_2, \dots, n_k, |\alpha)} =$$

$$\frac{B(\alpha + n_1, \dots, \alpha + n_t + c, \dots, \alpha + n_k)}{B(\alpha, \alpha, \dots, \alpha)} \cdot \frac{B(\alpha, \alpha, \dots, \alpha)}{B(\alpha + n_1, \dots, \dots, \alpha + n_k)}$$

Так как $B(x_1, \dots, x_k) = \frac{\prod_{i=1}^{k} \Gamma(x_i)}{\Gamma(\sum_{i=1}^{k} x_i)}$, где Γ — гамма-функция,

и $\Gamma(z+1) = z\Gamma(z)$ то

$$\frac{P(t, n_1, n_2, \ldots, n_k, |\alpha)}{P(n_1, n_2, \ldots, n_k, |\alpha)} = \frac{\Gamma(\alpha + n_t + c)}{\Gamma(\alpha + n_t)} \cdot \frac{\Gamma(\alpha k + \sum_{i=1}^{k} n_i)}{\Gamma(\alpha k + \sum_{i=1}^{k} n_i + c)} = \\ = \frac{(\alpha + n_t)_c}{(k\alpha + \sum_{i=1}^{k} n_i)_c}$$

Лемма доказана.

Если не принимать во внимание, что параметры мультиномиального распределения были получены из распределения Дирихле, то оценка параметров мультиномиального распределения была бы $\frac{n_t}{\sum_{i=1}^{k} n_i}$, а оценка искомой вероятности была бы

$$\left(\frac{n_t}{\sum_{i=1}^{k} n_i}\right)^c \approx \frac{(\alpha + n_t)_c}{(k\alpha + \sum_{i=1}^{k} n_i)_c} \quad \text{при} \quad c \ll n_t, \alpha \ll n_t.$$

Таким образом, использование распределения Дирихле обосновывает регуляризацию оценок параметров мультиномиального распределения.

Лемма 2. Пусть имеется выборка из мультиномиального распределения, параметры которого получены из распределения Дирихле с известным параметром α. Кроме того, имеется вторая выборка. Пусть мультиномиальная случайная величина принимает значения $1, \ldots, k$, а в данной выборке значение 1 реализовалось n_1 раз, 2 — n_2 раз, ..., k — n_k раз. Пусть во второй выборке значение 1 реализовалось c_1 раз, 2 — c_2 раз, ..., k — c_k раз. Тогда вероятность того, что вторая выборка получена из того же мультиномиального распределения, равна

$$\frac{\prod_{i=1}^{k}(\alpha+n_i)_{c_i}}{(k\alpha+\sum_{i=1}^{k}n_i)_{(\sum_{i=1}^{k}c_i)}}$$

Доказательство: Воспользуемся тождеством:

$$P(A,B|C)=P(A|B,C)P(B|C)$$

Получим:

$$\begin{aligned}
&P(c_1,c_2,\ldots,c_k|n_1,\ldots,n_k)=\\
&\qquad=P(c_1,\ldots,c_{k-1}|c_k,n_1,\ldots,n_k)P(c_k|n_1,\ldots,n_k)=\\
&\quad=P(c_1,\ldots,c_{k-1}|n_1,\ldots,n_{k-1},n_k+c_k)P(c_k|n_1,\ldots,n_k)=\\
&=P(c_1|n_1,n_2+c_2,\ldots,n_k+c_k)\cdot P(c_2|n_1,n_2,n_3+c_3,\ldots,n_k+c_k)\cdot\ldots\\
&\qquad\qquad\cdot P(c_k|n_1,\ldots,n_k)
\end{aligned}$$

Согласно лемме 1,

$$\begin{aligned}
&P(c_1|n_1,n_2+c_2,\ldots,n_k+c_k)\cdot\cdots\cdot P(c_k|n_1,\ldots,n_k)=\\
&\quad=\frac{(\alpha+n_1)_{c_1}}{(k\alpha+\sum_{i=1}^{k}n_i+\sum_{i=2}^{k}c_k)_{c_1}}\cdot\cdots\cdot\frac{(\alpha+n_k)_{c_k}}{(k\alpha+\sum_{i=1}^{k}n_i)_{c_k}}=\\
&\qquad\qquad=\frac{\prod_{i=1}^{k}(\alpha+n_i)_{c_i}}{(k\alpha+\sum_{i=1}^{k}n_i)_{(\sum_{i=1}^{k}c_i)}}
\end{aligned}$$

Лемма доказана.

Если не принимать во внимание, что параметры мультиномиального распределения были получены из распределения Дирихле, то оценка параметров мультиномиального распределения была бы $\frac{n_t}{\sum_{i=1}^{k}n_i}$, а оценка искомой вероятности была бы

$$\prod_{i=1}^{k}\Big(\frac{n_i}{\sum_{j=1}^{k} n_j}\Big)^{c_i} \approx \frac{\prod_{i=1}^{k}(\alpha+n_i)_{c_i}}{(k\alpha+\sum_{i=1}^{k} n_i)_{(\sum_{i=1}^{k} c_i)}} \quad \text{при} \quad c \ll n_t, \alpha \ll n_t.$$

Таким образом, мы снова получили регуляризированные оценки максимального правдоподобия.

Опишем теперь сам процесс семплирования.

Семплирование группы слов для i-го слова в документе d. Для семплирования группы слов необходимо найти вероятность того, что принадлежности этого слова к различным группам:

$$P(g_{di} = g | G \backslash g_{di}, T)$$

Воспользовавшись теоремой Байеса и отбросив все множители, не зависящие от g, получим:

$$P(g_{di} = g | G \backslash g_{di}, T) \propto P(g_{di} = g | G \backslash g_{di}, T) P(x_{di} | G, T, X \backslash x_{di})$$

Для того, чтобы найти неизвестные вероятности, обратимся к порождающей модели. Пересчитаем значения n и m с учетом исключения слова x_{di} из коллекции. Пусть до этого слово $x_{di} = w$ принадлежало к g-й группе слов и относилось к t-й теме. Тогда

$n_{dgt} := n_{dgt} - 1$;
$n_{dg\cdot} := n_{dg\cdot} - 1$;
$n_{d\cdot t} := n_{d\cdot t} - 1$;
$c_{wt} := c_{wt} - 1$;
$c_{\cdot t} := c_{\cdot t} - 1$;
Если после этого $n_{dgt} = 0$ (группа слов стала пустой), то
$m_{dt} := m_{dt} - 1$;
$m_{d\cdot} := m_{d\cdot} - 1$;

$m_{\cdot t} := m_{\cdot t} - 1$;

Вероятность $P(g_{di} = g|G\backslash g_{di}, T)$ — это вероятность того, что слово x_{di}, поступив в документ, будет отнесено к группе слов g. Согласно порождающей модели, эта вероятность равна

$$\frac{n_{dg\cdot}}{n_{d\cdot\cdot} + \beta}.$$

Вероятность $P(x_{di}|G, T, X\backslash x_{di})$ — это вероятность того, что слово $x_{di} = w$ было выбрано из того же распределения, что и все остальные слова, относящиеся к теме $t = t_{dg}$. Согласно лемме 1, эта вероятность равна

$$\frac{c_{wt} + \alpha}{c_{\cdot t} + \alpha|W|}.$$

Таким образом, при семплировании группы слов для слова x_{di}:

—С вероятностью, пропорциональной

$$p(g) = n_{dg\cdot}\frac{c_{wt_{dg}} + \alpha}{c_{\cdot t_{dg}} + \alpha|W|}, \quad \text{при} \quad g = 1, \ldots, m_{d\cdot} \tag{4}$$

слово будет отнесено к группе слов g;

—С вероятностью, пропорциональной

$$p(t_{m_{d\cdot}+1}, t) = \beta\frac{c_{wt} + \alpha}{c_{\cdot t} + \alpha|W|} \cdot \frac{m_{\cdot t}}{m_{\cdot\cdot} + \gamma} \tag{5}$$

для слова будет создана новая группа с темой t , $t = 1, \ldots, N$

—С вероятностью, пропорциональной

$$p(t_{m_{j\cdot}+1}, k_{N+1}) = \beta\frac{1}{|W|} \cdot \frac{\gamma}{m_{\cdot\cdot} + \gamma} \tag{6}$$

для слова будет создана новая группа с новой темой

После семплирования вновь пересчитаем параметры n, m и c с учетом добавленного слова.

Семплирование темы для j-й группы слов в d-м документе. Пусть в группе, для которой семплируется тема, содержится p_1 слов w_1, p_2 слов w_2, ..., $p_{|W|}$ слов $w_{|W|}$. Обозначим за $\{x_{dj}\}$ множество слов, относящихся к d-му документу и j-й группе. Для этого множества $\{x_{dj}\}$ мы и будем сейчас выбирать тему. Для семплирования темы для группы слов необходимо найти вероятность

$$P(t_{dj}|G, T\backslash t_{dj}, X)$$

Воспользовавшись формулой Байеса и отбросив множители, не зависящие от t_{dj}, получим:

$$P(t_{dj} = t|G, T\backslash t_{dj}, X) \propto P(\{x_{dj}\}|G, T, X\backslash\{x_{dj}\})P(t_{dj} = t|G, T\backslash t_{dj}, X\backslash\{x_{dj}\})$$

Для того, чтобы найти неизвестные вероятности, вновь обратимся к порождающей модели. Пересчитаем значения n и m с учетом исключения слов из $\{x_{dj}\}$ из коллекции. Пусть до этого эти слова принадлежали к j-й группе слов и относилось к t-й теме. Кроме того, пусть $|\{x_{dj}\}| = n_{djt}$. Тогда

$n_{djt} := 0$;
$n_{dj\cdot} := 0$;
$n_{d\cdot t} := n_{d\cdot t} - n_{djt}$;
$c_{wt} := c_{wt} - p_w \quad \forall w \in W$;
$c_{\cdot t} := c_{\cdot t} - n_{djt}$
$m_{dt} := m_{dt} - 1$;
$m_{d\cdot} := m_{d\cdot} - 1$;
$m_{\cdot t} := m_{\cdot t} - 1$;

Вероятность $P(t_{dj} = t|G, T\backslash t_{dj}, X\backslash\{x_{dj}\})$ — это вероятность того, что при появлении в d-м документе новой группы слов для нее была выбрана тема t. Согласно порождающей модели, эта вероятность равна

$$P(t_{dj} = t|G, T\backslash t_{dj}, X\backslash\{x_{dj}\}) = \frac{m_{\cdot t}}{m_{\cdot\cdot} + \gamma}.$$

Вероятность $P(\{x_{dj}\}|G, T, X\backslash\{x_{dg}\}$ — это вероятность того, что слова $\{x_{dj}\}$ были выбраны из того же распределения, что и все остальные слова темы t. Согласно лемме 2, эта вероятность равна

$$P(\{x_{dj}\}|G, T, X\backslash\{x_{dj}\}) = \frac{\prod_{w\in W}(c_{wt}+\alpha)_{(p_w)}}{(c_{\cdot t}+\alpha|W|)_{n_{djt}}}.$$

Таким образом, при семплировании темы для группы слов g в документе d:

—С вероятностью, пропорциональной

$$p(t) = m_{\cdot t}\frac{\prod_{w\in W}(c_{wt}+\alpha)_{(p_w)}}{(c_{\cdot t}+\alpha|W|)_{n_{dgt}}}, \quad g_{dg} = t, \quad t = 1, \ldots, N \tag{7}$$

—С вероятностью, пропорциональной

$$p(t_{new}) = \beta\frac{\prod_{w\in W}(\alpha)_{(p_w)}}{(\alpha|W|)_{n_{djt}}} \tag{8}$$

для группы слов будет создана новая тема t_{N+1}

После семплирования вновь пересчитаем параметры n, k и c с учетом добавленной группы слов.

Полностью алгоритм семплирования представлен на 2.

Data: matrix D word-document
Result: matrixes T and G
Select first approximation of T and G;
Compute X, c, m, n according to definition;
while *T and G not stabilize* **do**
    // Computing G
    foreach *document d* **do**
        foreach *word x_{di} in document d* **do**
            $w := x_{di}; g := g_{di}; t = t_{dg}$;
            $n_{dgt} - -; n_{dg\cdot} - -; n_{d\cdot t} - -$;
            $c_{wt} - -; c_{\cdot t} - -$;
            if $n_{dgt} == 0$ **then** $m_{dt} - -; m_{d\cdot} - -; m_{\cdot t} - -$;
            compute priors $p(g), p(g_{new}, t), p(g_{new}, t_{new})$ according to (4), (5), (6) ;
            sample g and t from computed priors;
            $n_{dgt} + +; n_{dg\cdot} + +; n_{d\cdot t} + +$;
            $c_{wt} + +; c_{\cdot t} + +$;
            $g_{di} := g$;
            if $g == g_{new}$ **then**
                $m_{dt} + +; m_{d\cdot} + +; m_{\cdot t} + +$;
                $t_{dg} := t$;
            end
        end
    end
    // Computing T
    foreach *document d* **do**
        foreach *wordgroup j in document d* **do**
            $t = t_{dj}; n = n_{djt}$;
            $n_{djt} := 0; n_{dj\cdot} := 0; n_{d\cdot t} := n_{d\cdot t} - n$;
            $m_{dt} - -; m_{d\cdot} - -; m_{\cdot t} - -$;
            foreach *w in W* **do**
                $p_w :=$ count of words w in $\{x_{dj}\}$ $c_{wt} - = p_w; c_{\cdot t} - = n$
            end
            compute priors $p(t), p(t_{new})$ according to (7),(8);
            sample k from computed priors;
            $n_{djt} + = n; n_{dj\cdot} + = n; n_{d\cdot t} + = n$;
            $m_{dt} + +; m_{d\cdot} + +; m_{\cdot t} + + \; g_{dj} := t$;
            foreach *w in W* **do**
                $c_{wt} + = p_w; c_{\cdot t} + = n$;
            end
        end
    end
end

Algorithm 2: Sampling values of T and G

8 Вычислительный эксперимент.

1. Сгенерируем при помощи порождающей модели коллекцию документов. Параметры порождающей модели:

число документов: 50;

число слов в словаре: 200;

число слов в документе: 50;

α: 2.0;

β: 10;

γ: 0.01;

Число тем в коллекции получилось равным 36.

Применим к сгенерированной коллекции алгоритм HDP с теми же параметрами. На рис. 4 и 3 представлены зависимости перплексии и числа тем в зависимости от номера итерации.

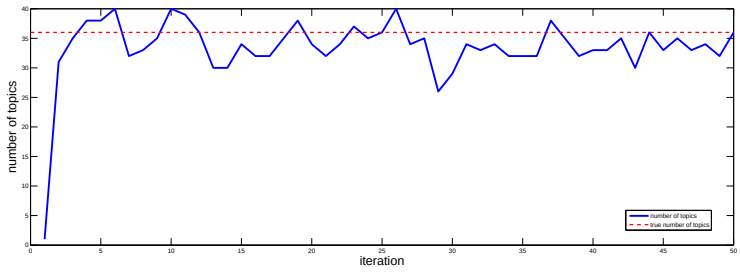

Рис. 3: Зависимость числа тем от номера итерации

Как мы видим, 25ти итераций достаточно для того, чтобы перплексия достигла минимального значения. После этого ее значение стабилизируется, но число тем постоянно меняется.

Проведем следующий эксперимент. На одной и той же сгенерированной коллекции запустим 30 раз алгоритм HDP с одними и теми же параметрами. После этого построим гистограмму 5 числа тем, определенных алгоритмом.

Истинное число тем в коллекции — 38.

Как мы видим, результаты работы алгоритма HDP нестабильны.

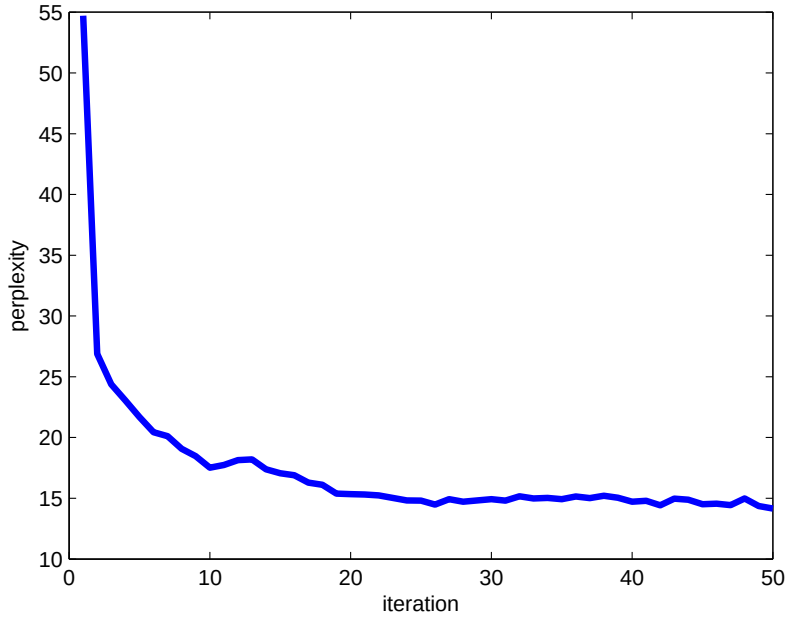

Рис. 4: Зависимость перплексии от номера итерации

Кроме того, важной характеристикой тематической модели, помимо перплексии, является разреженность, то есть доля нулей в матрицах тема-слово и документ-тема. Изменение разреженности в зависимости от итерации представлено на рис. 6 и 7. Разреженность матрицы документ-тема достигает 88%, а матрицы тема-слово — 97.5%.

Теперь обратимся к следующему аспекту задачи. Порождающая модель зависит от трех параметров — α, β, γ. Соответственно, алгоритм HDP тоже зависит от этих параметров. Имея искусственно сгенерированную коллекцию, мы знали эти параметры и запускали алгоритм HDP, используя их. В реальности же эти параметры нам неизвестны, поэтому представляесть целесообразным проверить, как поведет себя алгоритм HDP при различных значениях параметров α, β, γ. Эксперименты проводились по 5 раз, их результаты усреднялись. На рис. 8, 9, 10 представлена зависимость перплексии и числа тем при вариациях переменных α, β, γ соответственно.

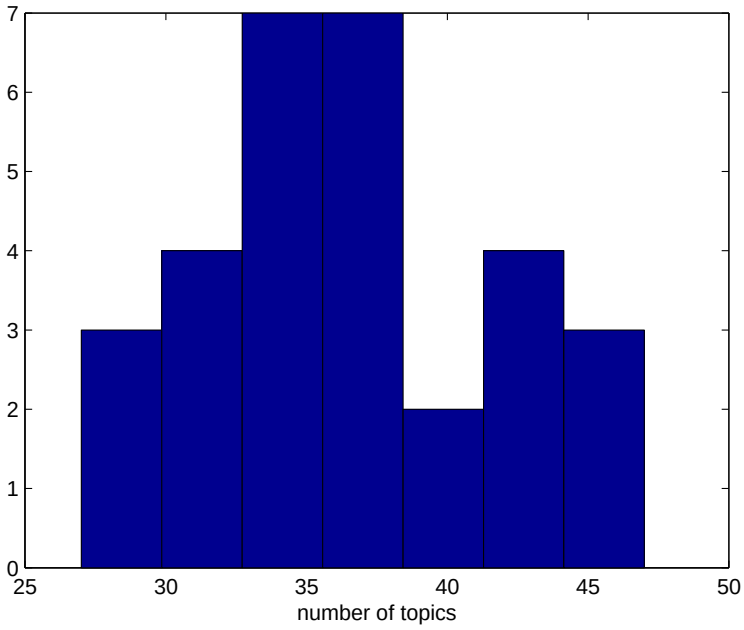

Рис. 5: Число тем

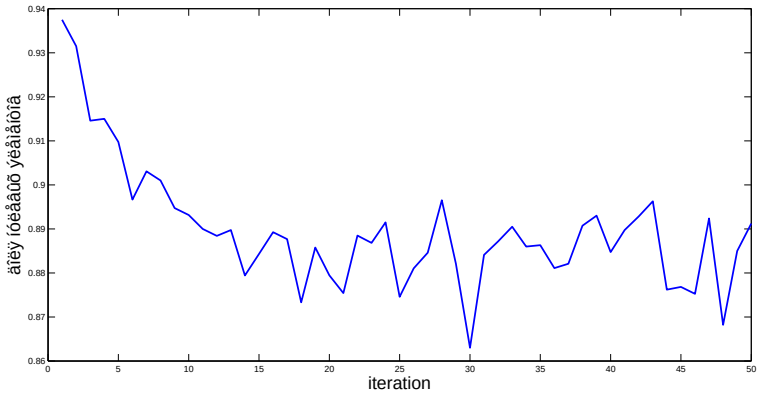

Рис. 6: Доля нулей в матрице документ-тема

Таким образом мы видим, что модель HDP неустойчива и существенно зависит от входных параметров даже при восстановлении

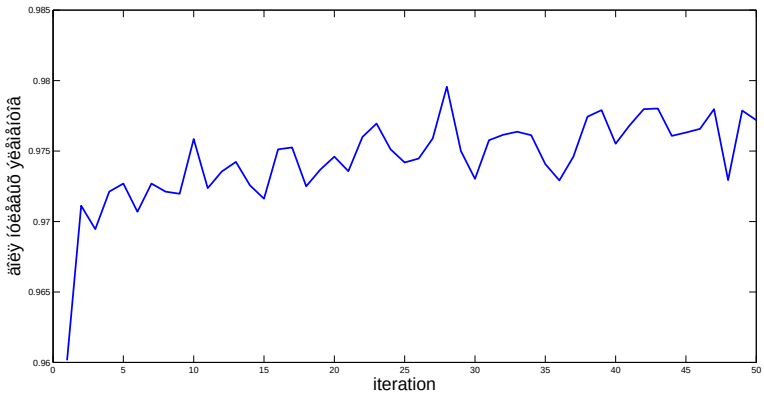

Рис. 7: Доля нулей в матрице тема-слово

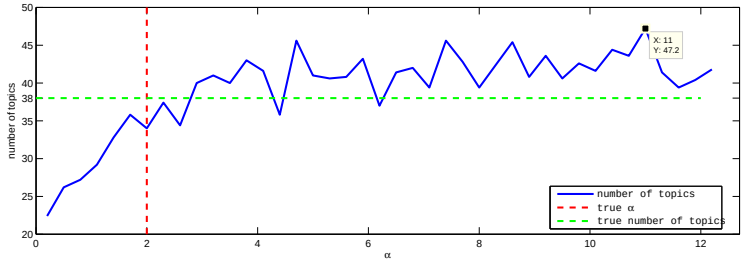

Рис. 8: Зависимость числа тем, определенных алгоритмом HDP, от параметра α

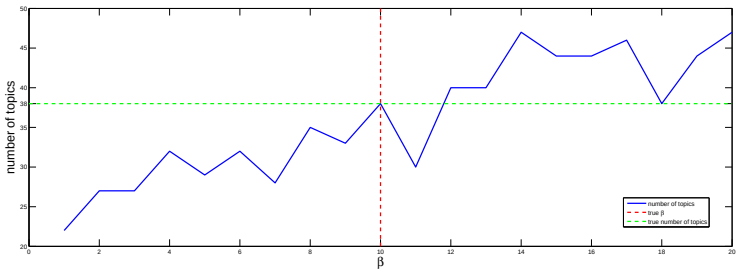

Рис. 9: Зависимость числа тем, определенных алгоритмом HDP, от параметра β

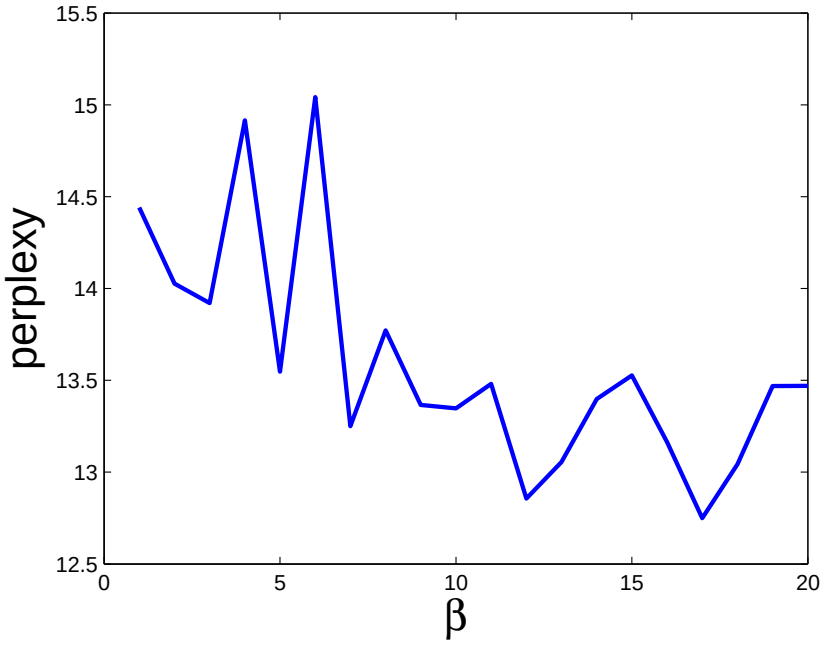

Рис. 10: Зависимость перплексии от параметра β

классификации, полученной из соответствующей ей порождающей модели. Отсюда следует необходимость введения какой-либо эвристики, которая позволила бы стабилизировать модель.

Попробуем теперь восстановить при помощи HDP коллекцию документов, полученную из порождающей модели LDA.

Сначала посмотрим, как HDP восстанавливает простую модель, в которой распределения слов в темах и тем в документах существенно разреженные ($\alpha = 0.01, \beta = 0.01$). Проведем следующий эксперимент.

На рис. 11 и 12 представлен результат восстановления алгоритмом HDP разреженной модели. Параметры генерации: $\alpha = 0.01$; $\beta = 0.01$; Истинное число тем — 30.

Мы видим, что настоящие и определенные при помощи HDP распределения слов в темах совпадают, а тема, для которой не существу-

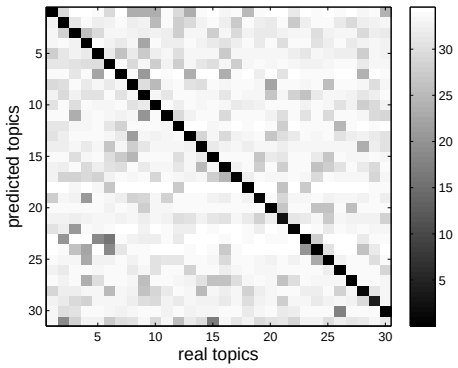

Рис. 11: Расстояния между истинными и найденными темами

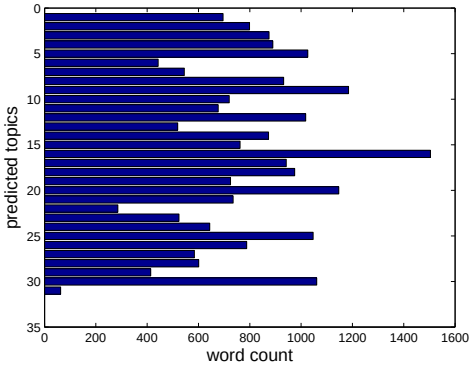

Рис. 12: Число слов в найденных темах

ет соответствия содержит существенно меньше слов, чем остальные.

Совсем иначе происходит восстановление не настолько разреженной коллекции. На рис. 13 и 14 представлен результат восстановления алгоритмом HDP неразреженной модели. Параметры генерации: $\alpha = 0.1$; $\beta = 0.1$; Истинное число тем — 10.

Мы видим, что в этом случае алгоритм HDP пределяет лишние темы, число слов в которых уже нельзя считать незначительным. Заметим, что параметром модели, от которого зависит, насколько различными могут быть распределения слов внутри групп, принадлежащих одной и той же теме, является параметр регуляризации α. Чем он больше, тем больше может быть расстояние Кульбака-Лейбнера между эмпирическими функциями распределения слов в группах, принадлежащих одной теме.

Введем эвристику. На каждой итерации (в ходе обхода коллекции при семплировании) будем увеличивать параметр α (в данном случае на каждой итерации $\alpha = \alpha \cdot 1.2$).Результаты работы измененного алгоритма, запущенного на той же неразреженной коллекции документов, представлены на 15 и 16.

Как мы видим, результатом работы HDP с примененной эвристи-

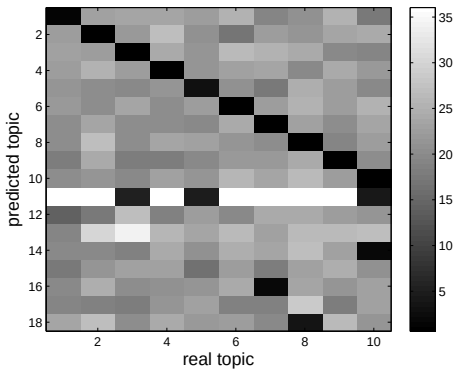

Рис. 13: Расстояния между истинными и найденными темами

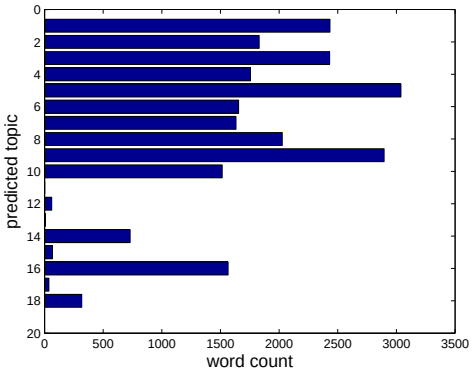

Рис. 14: Число слов в найденных темах

кой является верное определение тем (лишняя определенная тема содержит значительно меньше слов, чем все остальные темы). Кроме того, число тем устойчиво и стабилизируется на 7й итерации, как показано на 17.

9 Выводы

Алгоритм, основанный на использовании иерархического процесса Дирихле(HDP) в качестве порождающей модели позволяет производить тематическую классификацию коллекций документов и способен автоматически определять число тем, содержащихся в коллекции документов. Он хорошо подходит для классификации коллекций, в которых распределения слов в темах и тем в документах являются сильно разреженными. Однако, этот алгоритм является вероятностным, и, как показали эксперименты, его результаты обладают большой дисперсией. Было показано также, что результат работы этого метода существенно зависит от входных параметров модели.

Кроме того, при при классификации коллекций, в которых распределения слов в темах и тем в документах не являются разрежен-

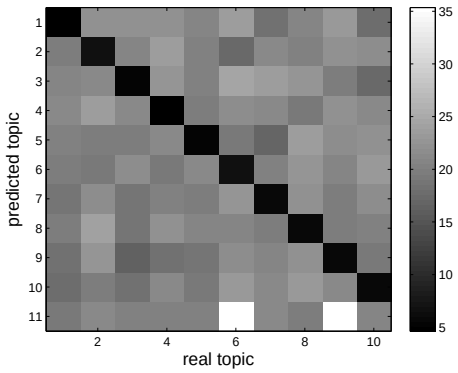

Рис. 15: Расстояния между истинными и найденными темами

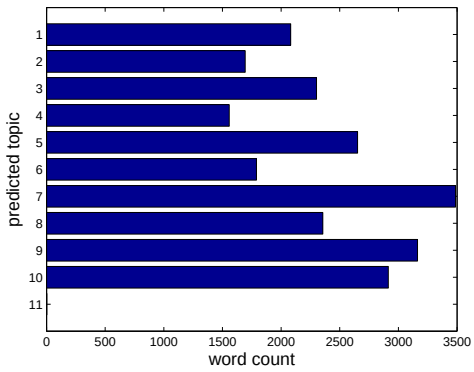

Рис. 16: Число слов в найденных темах

ными, качество построенной HDP модели существенно падает. В связи с этим была предложена модификация этого алгоритма, которая позволяет существенно улучшить качество получаемой алгоритмом модели путем корректировки его параметров.

Список литературы

[1] Daniel D. Lee H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, pages 556–562, 2001.

[2] Deerwester S. et al. Improving information retrieval with latent semantic indexing. In *Proceedings of the 51st Annual Meeting of the American Society for Information Science 25*, pages 36–40, 1988.

[3] A. Price, R Zukas. Application of latent semantic indexing to processing of noisy text intelligence and security informatics. *Lecture Notes in Computer Science*, 3495:602–603, 2005.

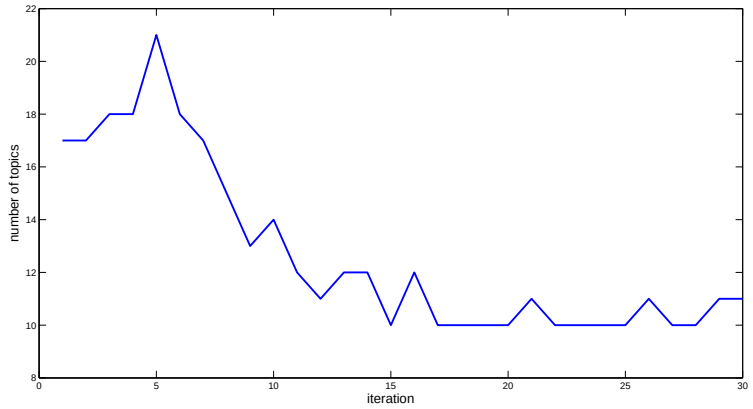

Рис. 17: Зависимость числа тем от номера итерации

[4] J. Zhao, L. Callan. Term necessity prediction. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM 2010)*, 2010.

[5] David M. Blei John D. Lafferty. Topic models. *Text Mining: Classification, Clustering, and Applications Chapman & Hall/CRC Data Mining and Knowledge Discovery Series,*, 2009.

[6] D. Blei A. Ng M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[7] D. Blei J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1:1:17–35, 2007.

[8] M. Hoffman D. Blei F. Bach. Online learning for latent dirichlet allocation. *Neural Information Processing Systems*, 2010.

[9] D. Blei J. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[10] C. Wang D. Blei D. Heckerman. Continuous time dynamic topic models. *Uncertainty in Artificial Intelligence [UAI]*, 2010.

[11] Daniel D. Lee H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 (6755):788–791, 1999.

[12] Vorontsov K. V. Potapenko A. A. Regularization, robustness and sparsity of probabilistic topic models.

[13] Chemudugunta C. Smyth P. Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model. *Advances in Neural Information Processing Systems*, 19:241–248, 2006.

[14] Asuncion A. Welling M. Smyth P. Teh Y. W. On smoothing and inference for topic models. In *Intl conf. on Uncertainty in Artificial Intelligence.*, 2009.

[15] Wallach H. Mimno D. McCallum A. Rethinking lda: Why priors matter. *Advances in Neural Information Processing Systems*, 22:1973–1981, 2009.

[16] Y. Teh M. Jordan M. Beal D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101[476]:1566–1581, 2006.

[17] $http://en.wikipedia.org/wiki/dirichlet_distribution$.

[18] Mark Johnson. Chinese restaurant processes. CG168 notes.

[19] Phil Blunsom Sharon Goldwater Trevor Cohn Mark Johnson. A note on the implementation of hierarchical dirichlet processes.