Статистические критерии адекватности вероятностных тематических моделей коллекции текстовых документов

6 июня 2013 г.

1 Введение

Тематическое моделирование (topic modeling) — одно из активно развивающихся приложений машинного обучения к анализу текстов [9]. *Тематическая модель* коллекции текстовых документов определяет, к каким темам относится каждый документ, и какие термины образуют каждую тему. *Вероятностная тематическая модель* описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем. Это позволяет решать задачи классификации, кластеризации и категоризации текстов, а также создавать тематические поисковые системы, позволяющие по тексту произвольной длины находить документы схожей тематики. Кроме того, тематические модели широко используются в области компьютерного зрения: для классификации изображений [1], определения подписей к ним [2] и построения иерархий [3, 4].

Исходными данными для тематической модели является множество (коллекция) текстовых документов D и множество (словарь) терминов W. Каждый документ $d \in D$ представляется последовательностью терминов $(w_1, \ldots, w_{n_d})$ из W, где n_d — длина документа. Через n_{dw} обозначается число вхождений термина w в документ d.

Латентно-семантический анализ [5] (латентно-семантическое индексирование в информационном поиске)—метод обработки информации, анализирующий взаимосвязь между документами и встречающимися в них терминами путем представления документов и терминов в пространстве так называемых “тем”. Для этого используется сингулярное разложение матрицы слов-на-документы A, обычно содержащей в качестве элементов веса, учитывающие частоты использования каждого термина в каждом документе и участие термина во всех документах (tf-idf). Иначе говоря, матрица A представляется в виде произведения трех матриц:

$$A = U\Lambda V^T,$$

где U и V—ортогональные матрицы, а Λ—диагональная матрица, на диагонали которой—собственные значения AA^T, называемые сингулярными значениями матрицы A. Оставляя k наибольших диагональных элементов матрицы Λ и взяв соответствующие им столбцы матриц U и V, получается матрица

$$A_k = U_k\Lambda_k V_k^T$$

ранга k, которая является лучшей аппроксимацией исходной матрицы A среди матриц заданного ранга k (минимизирует норму Фробениуса разности матриц [16]). Таким образом, каждый термин и документ представляется в общем пространстве размерности k, что позволяет определять близость между документами, терминами или документами и терминами как косинус угла между соответствующими векторами. Латентно-семантический анализ используется для классификации документов, кластеризации, поиска информации, позволяет решать проблемы, связанные с синонимией терминов [6].

Вероятностный латентно-семантический анализ, появившийся в 1999 году [10], является дальнейшим развитием *латентно-семантического анализа*, в отличие от которого имеет строгое статистическое обоснование и определяет модель порождения коллекции документов. Кроме того, данный метод позволяет разделять различные значения слов, тем самым решая проблемы, связанные с полисимией (многозначностью). Вероятностный латентный семантический анализ основывается на представлении вероятности появления пары "документ-термин" следующим образом:

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

где t—скрытая переменная(тема). Неизвестные вероятности $p(w|t)$ и $p(t|d)$ для всех $w \in W, t \in T, d \in D$ определяются из решения задачи максимизации логарифма правдоподобия выборки:

$$\sum_{d\in D}\sum_{w\in W} n_{dw} \log \sum_{t\in T} p(w|t)p(t|d) \longrightarrow \max_{\{p(w|t)\},\{p(t|d)\}}$$

при ограничениях:

$$\begin{cases} p(w|t) \geqslant 0 & \forall w \in W, t \in T \\ p(t|d) \geqslant 0 & \forall t \in T, d \in D \\ \sum\limits_{w\in W} p(w|t) = 1 & \forall t \in T \\ \sum\limits_{t\in T} p(t|d) = 1 & \forall d \in D \end{cases}$$

Для решения данной задачи применяется итерационный процесс, каждая итерация которого состоит из двух шагов (EM-алгоритм [11]).

Вероятностная тематическая модель с априорными распределениями Дирихле была предложена Дэвидом Блэем и др. в [12] и названа *латентным размещением Дирихле* (Latent Dirichlet Allocation, LDA). Одним из ключевых предположений модели является то, что распределения тем в документах подчиняются распределению Дирихле. При этом вводится двухуровневая модель порождения каждого документа коллекции $d \in D$, описанная в Алгоритме 1.1. Каждая тема представляет собой распределение вероятностей над словами $p(w|t)$, а каждый документ—распределение вероятностей над темами $p(t|d)$.

Одним из методов решения задачи тематического моделирования LDA является сэмплирование Гиббса, предложенное в [13](Gibbs Sampling, GS). В [14] описан обзор и анализ основных моделей обучения LDA, а в [15]— строгий вывод формул LDA-GS.

Все вероятностные модели основаны на следующих предположениях [10, 12].

Алгоритм 1.1. Порождение документа d в коллекции.

Вход: $p(w \mid t)$ для всех $w \in W, t \in T$, α—параметр распределения Дирихле;
Выход: последовательность слов документа d;

1: выбрать длину документа n_d;
2: выбрать распределение тем в документе $p(t \mid d) \sim Dir(\alpha)$;
3: **для всех** $i = 1 \dots n_d$
4: выбрать тему $t \sim p(t|d)$;
5: выбрать слово $w \sim p(w|t)$;

Во-первых, предполагается, что для выявления тематики достаточно знать, какие термины встречаются в каких документах, но не важен ни порядок терминов в документах (*гипотеза «мешка слов»*), ни порядок документов в коллекции (*гипотеза «мешка документов»*). Другими словами, предполагается, что тематику документа можно узнать даже после случайной перестановки терминов, хотя для человека такой текст теряет смысл.

Во-вторых, предполагается, что существует конечное множество тем T и дискретное распределение $p(d, w, t)$ на $D \times W \times T$, порождающее последовательность независимых наблюдений — троек (d_i, w_i, t_i), $i = 1, \dots, n$. Переменная t является латентной (скрытой), и наблюдаемая коллекция документов представляет собой последовательность пар (d_i, w_i), $i = 1, \dots, n$, оставшихся после отбрасывания всех тем.

В-третьих, предполагается, что условное распределение вероятностей терминов $p(w \mid d, t)$ в любом документе d зависит только от темы t, но не от самого документа. Это предположение называется *гипотезой условной независимости*:

$$p(w \mid d, t) = p(w \mid t). \tag{1}$$

Его можно представить эквивалентном виде:

$$p(d \mid w, t) = p(d \mid t). \tag{2}$$

Согласно формуле полной вероятности и гипотезе условной независимости,

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d). \tag{3}$$

Построить тематическую модель коллекции — означает по известной левой части $p(w \mid d) = n_{dw}/n_d$ найти 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 $|T|$.

Большинство тематических моделей [10, 12, 13, 14] оценивают вероятности тем $p(t \mid d, w)$ для каждого слова w в каждом документе d. Зная эти вероятности, возможно оценить число троек:

$n_{dwt} = n_{dw} p(t \mid d, w)$ — в которых термин w документа d связан с темой t,
$n_{dt} = \sum\limits_{w \in W} n_{dwt}$ — в которых термин документа d связан с темой t,
$n_{wt} = \sum\limits_{d \in D} n_{dwt}$ — в которых термин w связан с темой t,
$n_t = \sum\limits_{d \in D} \sum\limits_{w \in W} n_{dwt}$ — связанных с темой t,

и затем по ним найти частотные оценки искомых условных вероятностей:

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

Выполнение гипотезы условной независимости (2) является важным требованием к вероятностной тематической модели. В [17] предлагается критерий, оценивающий степень несоответствия темы t гипотезе условной независимости. Он основан на дивергенции Кульбака-Лейблера и может быть вычислен в ЕМ-алгоритме:

$$\mathrm{KL}_t = \mathrm{KL}(\hat{p}(d, w \mid t) \| \hat{p}(d \mid t)\hat{p}(w \mid t)) = \sum_{d,w} \frac{n_{dwt}}{n_t} \ln \frac{n_{dwt} n_t}{n_{dt} n_{wt}} \tag{5}$$

Проверка гипотезы условной независимости (2) для каждой пары слово–тема (w, t) является важной задачей, имеющей большую область применения в построении и оценивании вероятностных тематических моделей. Оба распределения $\hat{p}(d \mid t)$ и $\hat{p}(d \mid w, t)$ оцениваются согласно (4) в процессе построения тематической модели.

Критерий хи-квадрат Пирсона—один из статистических тестов, применяемый для проверки согласия экспериментальных данных с теоретическим распределением. Он имеет свои границы применимости и, в частности, плохо подходит для разреженных распределений [8, 7], когда число возможных значений наблюдаемой переменной значительно превосходит число наблюдений, либо когда многие значения имеют крайне низкие, хотя и ненулевые, вероятности. В этих случаях распределение статистики хи-квадрат уже не описывается классической асимптотикой, и может зависеть от длины выборки и вида теоретического распределения.

Разреженные распределения естественным образом возникают в прикладных задачах статистического анализа текстов. В работе предлагается эффективный способ оценивания функции распределения и квантилей статистики хи-квадрат, основанный на сэмплировании Монте-Карло.

!!! ДАЛЬНЕЙШЕЕ ОПИСАНИЕ

2 Критерий согласия хи-квадрат

Пусть имеется выборка n независимых наблюдений $\{x_1, \ldots, x_n\}$ 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 гипотезу о том, что случайная величина имеет заданное распределение $p(x)$, $x \in \Omega$. Для этого вычисляется статистика хи-квадрат:

$$X^2 = n \sum_{x \in \Omega} \frac{(\hat{p}(x) - p(x))^2}{p(x)}. \tag{6}$$

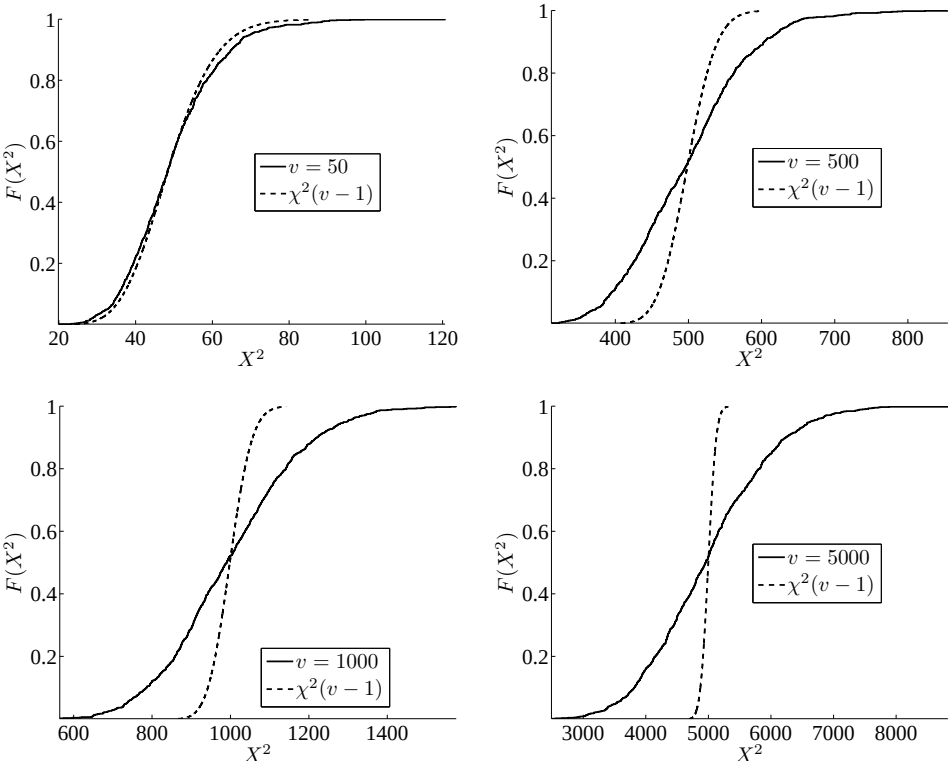

Рис. 1: Функции распределения статистики X^2 при $s = 1$, $v = 50,\ 500,\ 1000,\ 5000$, $n = 100$, $N = 1000$ и соответствующие функции распределения $\chi^2(v-1)$.

Распределение статистики X^2 стремится к распределению хи-квадрат с $k = |\Omega| - 1$ степенями свободы: $X^2 \sim \chi^2(k)$. Нулевая гипотеза отвергается на уровне значимости α, если значение статистики превышает $(1-\alpha)$-квантиль этого распределения: $X^2 > \chi^2_{1-\alpha}(k)$.

Считается, что асимптотика хи-квадрат применима, если объём выборки $n \geqslant 50$ и ожидаемое число наблюдений $np(x) \geqslant 5$ для каждого $x \in \Omega$. В случаях *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 очень больших выборках [8]. Стандартная рекомендация — объединять значения $x \in \Omega$ в группы — для сильно разреженных распределений оказывается неприемлемой, так как результат существенно зависит от способа группирования, который выбирается произвольно.

В качестве иллюстрации рассмотрим распределение, называемое *законом Ципфа*:

$$p(x) = Ax^{-s}, \quad x \in \Omega = \{1, \ldots, v\}, \tag{7}$$

где $A = \left(\sum_{x=1}^{v} x^{-s}\right)^{-1}$ — нормировочный множитель, s — параметр. Этот закон неплохо описывает частоты слов в текстах на естественных языках, если за x принимать номера слов, упорядоченных по убыванию частоты. Параметр s зависит от языка и от корпуса текстов, по которому делается оценка, но в большинстве экспериментов значение s близко к 1 и находится в пределах от 0.9 до 1.2 [19, 20].

Чем больше значение параметра s и размер словаря v, тем более разрежено распределение $p(x)$. Проведём простой вычислительный эксперимент. Возьмём типичные значения параметра $s = 1$ и размера словаря $v \in \{50,\ 500,\ 1000,\ 5000\}$. Сгене-

ририруем $N = 1000$ выборок (искусственных текстов) длины $n = 100$ из распределения (7), и для каждой выборки вычислим значение статистики X^2.

На рис. 1 сплошными линиями показаны эмпирические распределения статистики X^2, пунктирными линиями — распределения $\chi^2(v-1)$. Чем больше размер словаря, тем сильнее разрежено распределение $p(x)$, и тем сильнее отличаются $(1-\alpha)$-квантили этих распределений (при типичном значение $\alpha = 0.05$).

Таким образом, распределение хи-квадрат не может быть использовано в практических задачах анализа текстов, когда требуется проверить, является ли заданный текст $\hat{p}(x)$ случайной выборкой из корпуса текстов $p(x)$.

3 Тест на основе сэмплирования

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

Построение теста. Генерируется N независимых выборок объёма n из заданного дискретного распределения $p(x)$. Для каждой выборки вычисляется эмпирическое распределение $\hat{p}_j(x)$, $j = 1, \ldots, N$ и значение статистики X_j^2 по формуле (6). По полученным значениям $X_1^2, \ldots, X_N^2$ строится эмпирическая функция распределения статистики

$$\hat{F}_n(X^2) = \frac{1}{N} \sum_{j=1}^{N} \left[X_j^2 < X^2 \right]$$

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 [7].

Применение теста. Пусть задана выборка объёма n, по которой построено эмпирическое распределение $\hat{p}(x)$ и вычислено значение статистики X^2 согласно (6). Если $X^2 > \hat{F}_{n,1-\alpha}$, то нулевая гипотеза о том, что данная выборка порождена распределением $p(x)$, отклоняется.

Рекуррентное построение теста. Как будет показано ниже, в случае разреженных распределений значение квантили $\hat{F}_{n,1-\alpha}$ может зависеть от объёма выборки n. Строить тест заново для каждой выборки довольно накладно. Поэтому предлагается рекуррентный метод, позволяющий при заданном распределении $p(x)$ вычислить квантили для всех значений n один раз, и затем быстро осуществлять проверку нулевой гипотезы для выборок различного объёма n.

В рекуррентном методе N выборок $\{x_{j1}, \ldots, x_{jn}\}$ наращиваются одновременно, где $j = 1, \ldots, N$ — номер выборки, $n = 1, \ldots, n_{\max}$ — объём выборки. Для каждого j строится эмпирическая гистограмма $H_j(x) = n\hat{p}_j(x)$. При добавлении каждого нового наблюдения $\xi = x_{j,n+1}$, сэмплированного из распределения $p(x)$, обновляется гистограмма и пересчитывается значение статистики $X_{j,n+1}^2$ по значению $X_{j,n}^2$. Сэмплированные выборки не сохраняются. В процессе работы алгоритм формирует двумерный массив значений статистики $X_{j,n}^2$ и одномерный массив эмпирических гистограмм $H_j(x)$. В случае $|\Omega| \gg n_{\max}$ для хранения эмпирических гистограмм лучше использовать специальные структуры данных — разреженные векторы, не выделяющие память под нулевые значения $H_j(x)$. В таком случае расход памяти для данного

Алгоритм 3.1. Построение теста путём рекуррентного вычисления значений статистики X^2 по N одновременно растущим выборкам объёма n.

Вход: $p(x)$, N, $n_{\max}$, α;
Выход: $\hat{F}_{n,1-\alpha}$ для всех $n = 1, \ldots, n_{\max}$;

для всех $j := 1, \ldots, N$
  сэмплировать первый элемент j-й выборки $\xi \sim p(x)$;
  инициализировать эмпирическую гистограмму для j-й выборки: $H_j(x) := [x = \xi]$ для всех $x \in \Omega$;
  инициализировать значение статистики X^2 для j-й выборки: $X^2_{j1} := 1/p(\xi) - 1$;
для всех $n := 1, \ldots, n_{\max} - 1$
  для всех $j := 1, \ldots, N$
    сэмплировать $(n+1)$-й элемент j-й выборки $\xi \sim p(x)$;
    обновить эмпирическую гистограмму для j-й выборки: $H_j(\xi) := H_j(\xi) + 1$;
    обновить значение статистики X^2 для j-й выборки:
    $$X^2_{j,n+1} := \frac{nX^2_{j,n} + 1}{n+1} + \frac{2H_j(\xi) - 1}{(n+1)p(\xi)} - 2;$$
для всех $n := 1, \ldots, n_{\max}$
  упорядочить $X^2_{1,n}, \ldots, X^2_{N,n}$ по возрастанию;
  $\hat{F}_{n,1-\alpha} := X^2_{N(1-\alpha),n}$;

алгоритма составляет $O(n_{\max}N)$; вычислительная сложность $O(n_{\max}N \log N)$. Детали реализации показаны в Алгоритме 3.1.

4 Регрессионный тест

Рассмотрим частную постановку задачи: проверяется нулевая гипотеза о том, что выборка с эмпирическим распределением $\hat{p}(x)$ порождена распределением Ципфа (7) с параметром s. Будем строить распределение статистики X^2 с помощью сэмплирования и исследовать зависимость квантиля $\hat{F}_{n,1-\alpha}$ от параметров n, s и v.

На рис. 2 показана зависимость 0.95-квантиля от объёма выборки n и её интерполяция функцией $\tilde{F}_{1-\alpha}(n) = A + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4}$ с параметрами A, B, C, D, E.

На рис. 3 показана зависимость 0.95-квантиля от показателя s в законе Ципфа и её интерполяция функцией $\tilde{F}_{1-\alpha}(s) = F + GH^s$ с параметрами F, G, H.

На рис. 4 показана зависимости 0.95-квантиля от параметра $v = |\Omega|$ и её линейная интерполяция $\tilde{F}_{1-\alpha}(v) = I + Jv$ с параметрами I, J.

Построение регрессионного теста. Чтобы найти общий вид зависимости $\tilde{F}_{1-\alpha}(s, v, n)$, применим эмпирический подход. Сформируем обучающую выборку из 1000 троек (s, v, n), равномерно выбранных из параллелепипеда $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$. Для каждой тройки вычислим значение $\hat{F}_{n,0.95}$.

Для поиска нелинейной регрессионной зависимости используем алгоритм символьной регрессии MVR-composer [?, ?]. Преимущество этого алгоритма в том, что он автоматически подбирает формулу регрессии среди всевозможных супер-

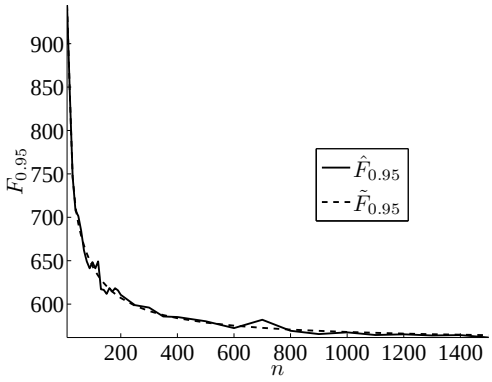

Рис. 2: Зависимость 0.95-квантиля X^2 от объёма выборки n при $s = 1$, $v = 500$, $N = 1000$ и ее интерполяция.

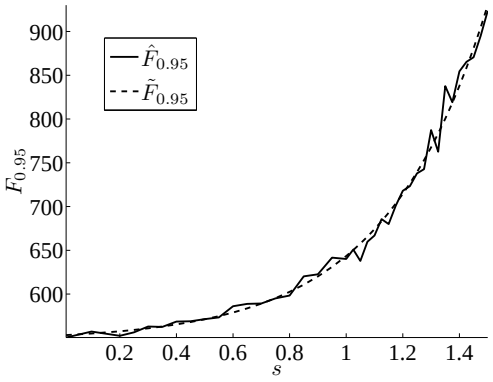

Рис. 3: Зависимость 0.95-квантиля X^2 от параметра s при $n = 100$, $v = 500$, $N = 1000$ и ее интерполяция.

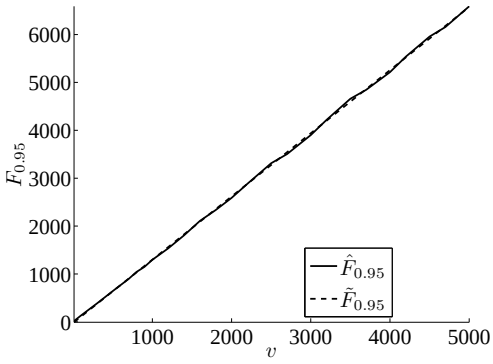

Рис. 4: Зависимость 0.95-квантиля X^2 от $v = |\Omega|$ при $s = 1$, $n = 100$, $N = 1000$ и ее интерполяция.

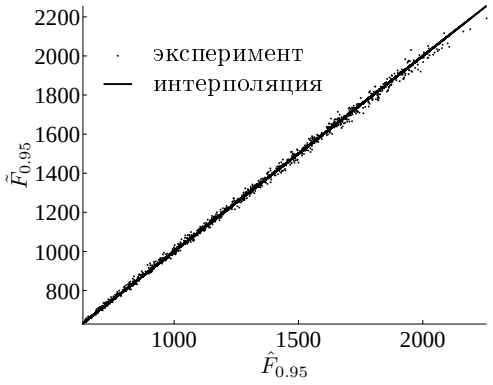

Рис. 5: Зависимость 0.95-квантилей, аппроксимированных моделью $\tilde{F}^4_{1-\alpha}$, от их эмпирических значений при различных s, n, v.

позиций заданного множества элементарных функций. В нашем случае MVR-composer находит следующую модель регрессии: $\tilde{F}^1_{1-\alpha}(s,v,n) = (A + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4})(F + GH^s)(I + Jv)$ и определяет оптимальные значения 10 параметров $A, B, C, D, E, F, G, H, I, J$. Рассмотрим также некоторые упрощения этой модели:

$$\tilde{F}^2_{1-\alpha}(s,v,n) = A(1 + Bn^{-1} + Cn^{-2} + Dn^{-3} + En^{-4})(1 + GH^s)(1 + Jv);$$
$$\tilde{F}^3_{1-\alpha}(s,v,n) = A(1 + Bn^{-c})(1 + GH^s)(1 + Jv);$$
$$\tilde{F}^4_{1-\alpha}(s,v,n) = Av(1 + Bn^{-c})(1 + GH^s);$$
$$\tilde{F}^5_{1-\alpha}(s,v,n) = Av(1 + GH^s);$$
$$\tilde{F}^6_{1-\alpha}(s,v,n) = Av(1 + Bn^{-c}).$$

Параметры этих моделей настроим с помощью функции nlinfit программы Matlab. Начальные приближения всех параметров положим равными 1, кроме параметра A, который инициализируем средним значением $\hat{F}_{n,1-\alpha}/v$ по всей выборке. Получим следующие значения среднеквадратичной ошибки (СКО) на обучающей и контрольной выборках из 1000 случайных троек (s,v,n) каждая:

модель	$\tilde{F}^1_{1-\alpha}$	$\tilde{F}^2_{1-\alpha}$	$\tilde{F}^3_{1-\alpha}$	$\tilde{F}^4_{1-\alpha}$	$\tilde{F}^5_{1-\alpha}$	$\tilde{F}^6_{1-\alpha}$
число параметров	10	8	6	5	3	3
СКО (обучение)	16.3	16.8	16.8	16.7	52.2	43.7
СКО (контроль)	15.8	16.1	16.0	16.0	50.9	43.8

Сравнение СКО на обучающей и контрольной выборках показывает, что переобучения нет ни в одной из моделей. Модель $\tilde{F}^4_{1-\alpha}$ представляется оптимальной по точности и числу параметров. Дальнейшее упрощение модели приводит к резкому увеличению СКО. Оптимальные значения параметров для неё: $A = 0.913$, $B = 3.98$, $c = 0.636$, $G = 0.00458$, $H = 36.8$.

На рис. 4 показан график зависимости 0.95-квантилей, аппроксимированных моделью $\tilde{F}^4_{0.95}$, от их эмпирических значений при различных s, n, v. Сплошной линией изображена «идеальная» прямая $\tilde{F} = \hat{F}$.

Таким образом, в отличие от классического критерия хи-квадрат, квантиль распределения статистики X^2 существенно зависит от объёма выборки n и от вида распределения $p(x)$, в частности, от показателя степени s в законе Ципфа, отвечающего за разреженность распределения. Построенная регрессионная модель довольно точно описывает зависимость 0.95-квантили от параметров s, n, v. Эту зависимость можно построить один раз вместо того, чтобы строить тест для каждого распределения $p(x)$. Предварительно необходимо убедиться, что распределение $p(x)$ описывается законом Ципфа и найти значение параметра s. Данное обстоятельство сужает область применимости регрессионного теста.

Анализ качества регрессионного теста. Оценим вероятности ошибок первого и второго рода предложенного регрессионного теста в эксперименте.

Ошибкой первого рода называется отклонение нулевой гипотезы при условии её истинности. Вероятность ошибки первого рода равна уровню значимости $\alpha = 0.05$. Для эксперимента сгенерируем контрольную выборку из 500 различных троек (s, v, n), равномерно распределённых на параллелепипеде $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$. Для каждой тройки сгенерируем 1000 выборок объёма n из распределения Ципфа $p(x)$ с параметрами v и s и вычислим значение статистики X^2. Оценим вероятность ошибки первого рода как долю выборок, для которых нулевая гипотеза отклонялась: $X^2 > \tilde{F}^4_{0.95}(s, v, n)$. Оценка вероятности ошибки первого рода составляет 0.0496 ± 0.0141 с доверительной вероятностью 0.95.

Ошибкой второго рода называется принятие гипотезы $H_0\colon p(x)$ при условии истинности ее альтернативы $H_1\colon p'(x)$. Вероятность ошибки второго рода существенно зависит от альтернативы — чем более похожи распределения $p(x)$ и $p'(x)$, тем больше вероятность ошибки. Исследуем способность теста различать распределения, отличающиеся на небольшом числе элементов x из Ω. Выделим из множества $\Omega = \{1, \ldots, v\}$ подмножество элементов с наибольшими вероятностями: $\Omega_0 = \{x\colon p(x) > \mu p(1)\}$ при заданном $\mu \in (0, 1)$. Построим распределение $p'(x)$ из $p(x)$ следующим образом: выберем K различных случайных элементов множества Ω_0 и их вероятности поменяем местами с вероятностями K различных случайных элементов множества $\Omega \backslash \Omega_0$.

Из полученного распределения $p'(x)$ сгенерируем выборки, для каждой построим эмпирическое распределение $\hat{p}(x)$ и вычислим статистику X^2. Если $X^2 \leqslant \tilde{F}^4_{0.95}(s, v, n)$, то для данной выборки гипотеза H_0 ошибочно принимается. Долю выборок, при которых это происходит, примем в качестве оценки вероятности ошибки второго рода.

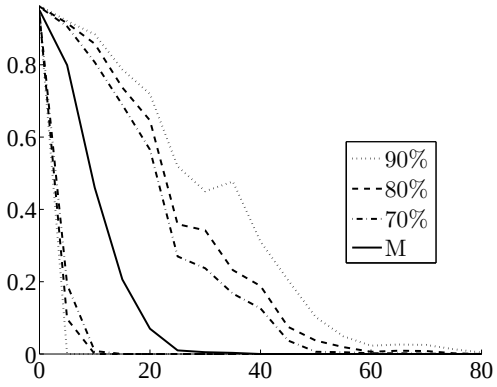

Рис. 6: Зависимость вероятности ошибки второго рода от K при $\mu = 0.01$.

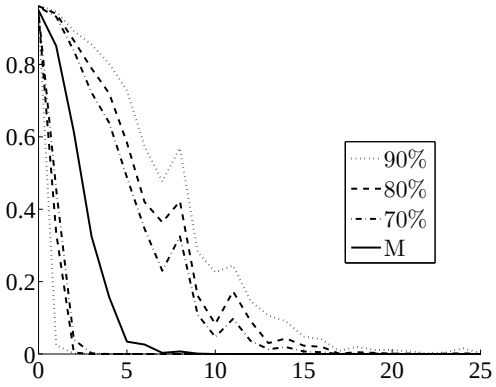

Рис. 7: Зависимость вероятности ошибки второго рода от K при $\mu = 0.05$.

Для каждого K сгенерируем 200 различных троек (s, v, n) из равномерного распределения на параллелепипеде $s \in [0.9, 1.1]$, $v \in [500, 1500]$, $n \in [50, 150]$ и вычислим 200 оценок вероятности ошибки второго рода. На рис. 6 и рис. 7 показаны зависимости медианы M и доверительных границ 90%, 80%, 70% вероятности ошибки второго рода от числа перестановок K при $\mu = 0.01$ и $\mu = 0.05$. По мере увеличения K распределения $p(x)$ и $p'(x)$ все сильнее отличаются, и вероятность ошибки второго рода уменьшается. По мере увеличения μ различия становятся менее контрастными, и вероятность ошибки второго рода убывает медленнее. При $\mu = 0.01$ она становится меньше 0.1 при $K = 20$, при $\mu = 0.05$ она достигает этого значения при $K = 5$.

Отсюда, в частности, можно сделать вывод, что различные тексты, отличающиеся лишь 5 высокочастотными терминами, в среднем довольно надёжно различаются по их случайным фрагментам.

5 Вероятностные тематические модели

Чтобы оценить качество тематической модели, необходимо проверить, выполняется ли гипотеза условной независимости (1) — важнейшее базовое предположение модели (3) — для каждой пары документ–тема (d, t). Тема t описывается распределением $\hat{p}(w \mid t)$. Выборка слов документа d, относящихся к теме t, согласно модели, образует эмпирическое распределение $\hat{p}(w \mid d, t)$. Оба распределения оцениваются согласно (4) в процессе построения тематической модели. Чтобы проверить, действительно ли данная выборка могла быть получена из распределения $\hat{p}(w \mid t)$, воспользуемся критерием согласия, основанным на статистике хи-квадрат (6):

$$X^2_{dt} = n_{dt} \sum_{w:\, n_{wt}>0} \frac{\big(\hat{p}(w \mid d, t) - \hat{p}(w \mid t)\big)^2}{\hat{p}(w \mid t)}. \tag{8}$$

Число различных слов в теме может быть намного больше, чем число слов в документе. Следовательно, мы имеем дело с разреженными распределениями, к которым неприменим асимптотический критерий хи-квадрат. Поэтому будем строить статистические тесты методом сэмплирования, для каждой темы $t \in T$ отдельно.

Экспериментально установлено, что для больших корпусов текстов на естественных языках закон Ципфа или более сложные параметрические законы (например

Ципфа–Мандельброта) выполняются с неплохой точностью [19, 20]. Для ускорения проверки гипотезы условной независимости предлагается двухэтапный тест. Сначала проверяется согласие каждой темы t с выбранным параметрическим законом. Если согласие есть, то строится один регрессионный тест для всех таких тем. Для каждой из остальных тем строится отдельный тест на основе сэмплирования.

6 Сэмплирование без возвращений

Проверки согласия документных эмпирических распределений $\hat{p}(w \mid d, t)$, $d \in D$ с распределением $\hat{p}(w \mid t)$, вообще говоря, не являются независимыми, поскольку имеется тождество, связывающие эти распределения друг с другом:

$$\hat{p}(w \mid t) = \sum_{d \in D} \hat{p}(w \mid d, t)\hat{p}(d \mid t). \tag{9}$$

Документы являются выборками без возвращений из распределения $\hat{p}(w \mid t)$, тогда как обычно критерии согласия предполагают выборку с возвращениями. Наличие дополнительного ограничения (9) может и не влиять на результаты тестов или влиять несущественно, особенно на коллекциях большого размера. Однако это лишь предположение, которое необходимо проверить. Для этого построим более точный тест на основе сэмплирования *без возвращений*, учитывающий, что последовательность слов, образующих тему t, разрезается на документы в пропорциях $\hat{p}(d \mid t)$.

Построение теста сэмплированием без возвращений. Возьмём последовательность терминов длины n_t, образующую распределение $\hat{p}(w \mid t)$. Сгенерируем N случайных перестановок этой последовательности. Разрежем каждую из полученных последовательностей W_j, $j = 1, \ldots, N$ на «документы» — подпоследовательности терминов W_{jd} длины n_{dt} каждая, $d \in D$. По каждому «документу» W_{jd} построим эмпирическое распределение $\hat{p}_j(w \mid d, t)$ и вычислим значение статистики хи-квадрат X^2_{jd}. Для каждого $d \in D$ по множеству значений статистики $X^2_{1d}, \ldots, X^2_{Nd}$ построим эмпирическую функцию распределения $\hat{F}_d(X^2)$ и вычислим её $(1-\alpha)$-квантиль $\hat{F}_{d,1-\alpha}$. Число N рекомендуется брать не менее 1000 при типичном значении $\alpha = 0.05$.

Отметим, что в тесте без возвращений квантиль строится для каждого документа d, тогда как тест с возвращениями строился для каждого значения длины документа n. Построение теста без возвращений более ресурсоёмко и требует $O(n_t N \log N)$ операций вместо $O(n_{\max} N \log N)$, где $n_{\max} = \max\limits_{d \in D} n_{td}$.

Применение теста сэмплированием без возвращений. Проверка гипотезы условной независимости для пары документ–тема (d, t) заключается в вычислении статистики X^2_{dt} по формуле (8) и проверке неравенства $X^2_{dt} > \hat{F}_{d,1-\alpha}$. Если оно выполнено, то гипотеза условной независимости отвергается для данной пары (d, t).

7 Вычислительные эксперименты

Эксперименты проводились на коллекции из $|D| = 2000$ авторефератов диссертаций на русском языке. Мощность словаря после предварительной обработки данных (лемматизации и удаления стоп-слов) составляет $|W| = 20211$ слов, длина документов от 1000 до 4000 слов. Строились две тематические модели — PLSA [10] и LDA-GS [12, 13] с помощью алгоритма, описанного в [**?**]. Число тем $|T| = 100$.

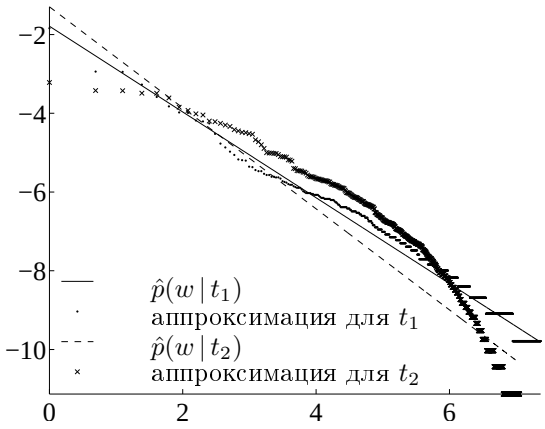

Рис. 8: Аппроксимация эмпирических распределений слов законом Ципфа (для двух тем, в логарифмических осях).

Выполняется ли закон Ципфа для тем? На рис. 8 показаны графики эмпирических распределений и закона Ципфа для двух из 100 тем t_1 и t_2 в модели LDA, в логарифмических осях. По горизонтальной оси откладывается логарифм номера слова, слова упорядочены по частоте. По вертикальной оси откладывается логарифм вероятности слова. Оптимальные значения параметра закона Ципфа: $s = 1.04$ для t_1, $s = 1.28$ для t_2. Хотя «на глаз» соответствие неплохое, особенно для t_1, нулевая гипотеза отклоняется для обоих тем. Более того, большинство тем согласуются с законом Ципфа лишь при крайне низких уровнях значимости, меньших 0.05. Это объясняется тем, что при выборках длины n_t порядка 10^3–10^5 критерии согласия чувствительны даже к незначительным различиям распределений, и одного параметра в законе Ципфа не достаточно для описания эмпирических распределений.

Сравнение тестов без возвращений и с возвращениями.

Для модели PLSA рассматривается одна тема из $|D_t| = 1992$ документов суммарной длины $n_t = 87026$ слов. В тестах без возвращений и с возвращениями нулевая гипотеза принимается для 1674 и 1688 документов соответственно. Решения отличаются на 22 документах из 1992. Оба теста дают примерно одинаковый результат: гипотеза условной независимости отклоняется для 15% документов.

Для модели LDA-GS рассматривается тема из $|D_t| = 1114$ документов суммарной длины $n_t = 63805$ слов. Нулевая гипотеза принимается для 1032 и 1035 документов соответственно. Решения отличаются на 7 документах из 1114. Оба теста снова дают примерно одинаковый результат: нулевая гипотеза отклоняется для 7% документов.

Таким образом, результаты тестов без возвращений и с возвращениями почти одинаковы, однако тест с возвращениями менее ресурсоёмкий.

Определение оптимального числа тем.

Тест может быть использован для определения оптимального числа тем в коллекции. Если на последней итерации алгоритма LDA-GS гипотеза условной независимости принимается для всех тем, то достигнуто оптимальное число тем. Например, на модельной коллекции $D = 500$, $W = 200$, $n_d = 120$, $T = 10$ зависимость средней доли слов, не прошедших гипотезу на уровне значимости 0.05, от задаваемого числа тем изображена на рис. 7. Видно, что при числе тем, равном или большем оптимального $T = 10$, гипотеза условной независимости не отвергается.

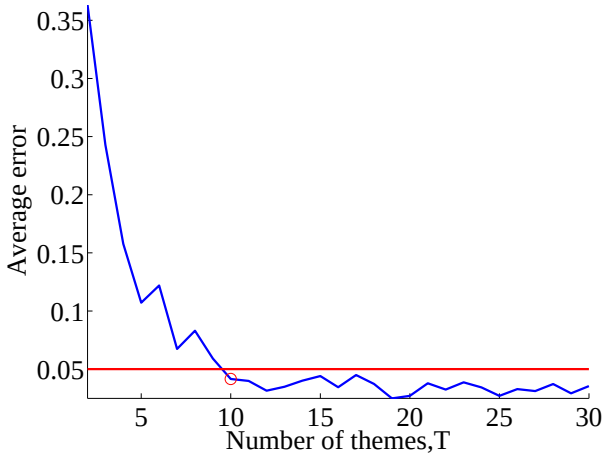

Рис. 9: Зависимость средней доли слов, не прошедших гипотезу, от числа тем.

8 Точный тест Фишера

Точный тест Фишера [18] используется для проверки отсутствия взаимосвязи между двумя переменными в таблице сопряженности размерности 2×2. При этом уровень значимости вычисляется, как если бы значения на границах таблицы были известны. Например, рассматриваются данные, разделенные на две категории двумя способами, с таблицей сопряженности X:

	Столбец 1	Столбец 2	Суммы по строкам
Строка 1	a	b	$a+b$
Строка 2	c	d	$c+d$
Суммы по столбцам	$a+c$	$b+d$	$a+b+c+d=n$

Тогда вероятность получить данную таблицу сопряженности при условии истинности нулевой гипотезы об отсутствии взаимосвязи между категориями задается гипергеометрическим распределением:

$$P(X) = \frac{C_{a+b}^{a} C_{c+d}^{c}}{C_{n}^{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}, \tag{10}$$

где C_n^k—биномиальный коэффициент, $(a+b)$ и $(c+d)$—маргинальные суммы по строкам, $(a+c)$ и $(b+d)$—по столбцам, n—общее число наблюдений. Уровень значимости для двустороннего варианта теста Фишера:

$$\text{Pvalue}_2 = \sum_{Y:P(Y)\leqslant P(X)} P(Y),$$

где Y—всевозможные таблицы сопряженности с той же суммой элементов в каждой строке и каждом столбце, что и в таблице X. Для проверки гипотезы против односторонней альтернативы о том, что наблюдения из строки 1 чаще попадают в столбец 1, чем в столбец 2, уровень значимости вычисляется по формуле:

$$\text{Pvalue}_1 = \sum_{Z} P(Z), \tag{11}$$

где Z—таблицы сопряженности с теми же маргинальными суммами, что и X, но обладающие не меньшими элементами первой строки первого столбца, чем X. Другими

Алгоритм 9.1. Односторонний тест Фишера

1: **ПРОЦЕДУРА** `Fisher`(X)
2: r=rand(0,1)
3: Pvalue $= \sum_{Z \neq X} P(Z) + rP(X)$
4: **return** Pvalue

словами, таблицы Z имеют вид:

	Столбец 1	Столбец 2	Суммы по строкам
Строка 1	a'	b'	$a' + b'$
Строка 2	c'	d'	$c' + d'$
Суммы по столбцам	$a' + c'$	$b' + d'$	$a' + b' + c' + d' = n'$

Причем выполняется:

$$\begin{cases} a' \geqslant a \\ a' + b' = a + b \\ c' + d' = c + d \\ a' + c' = a + c \\ b' + d' = b + d \\ n' = n \end{cases} \tag{12}$$

В случае больших выборок используется тест хи-квадрат, но он неприменим, когда математические ожидания значений в любой из ячеек таблицы с заданными границами ниже 10. Дело в том, что приближение выборочного распределения испытуемой статистической величины распределением хи-квадрат оказывается неадекватным при неравноценном распределении данных среди ячеек таблицы, а также при малых размерах выборки. В таких случаях применяется точный тест Фишера, который не зависит от особенностей выборки.

9 Применение точного теста Фишера для проверки гипотезы условной независимости

Проверка гипотезы для фиксированной тройки (d,w,t). Для каждого слова $w \in W$ и документа $d \in D$ в фиксированной теме $t \in T$ предлагается выполнить проверку гипотезы об их независимости с помощью точного теста Фишера. Таблица сопряженности для рассматриваемой тройки (d,w,t) имеет вид:

	w	$W \backslash w$	Суммы по строкам
d	n_{dwt}	$n_{dt} - n_{dwt}$	n_{dt}
$D \backslash d$	$n_{wt} - n_{dwt}$	$n_t - n_{dt} - n_{wt} + n_{dwt}$	$n_t - n_{dt}$
Суммы по столбцам	n_{wt}	$n_t - n_{wt}$	n_t

При анализе текстов математическое ожидание числа наблюдений в первом ячейке таблицы обычно оказывается меньше 10, поэтому асимптотика хи-квадрат неприменима. Рассматривается односторонняя альтернатива, заключающаяся в том, что слово w из темы t слишком часто встречается в документе d. При этом уровень значимости вычисляется по формуле (11) и может принимать лишь дискретный набор

Алгоритм 9.2. Проверка гипотезы условной независимости для темы t

Вход: t, n_t, α, n_{dwt}, n_{dt}, $n_{wt}\, \forall d \in D, w \in W$;
Выход: принять или отклонить гипотезу на уровне значимости α;

1: **для всех** документов $d \in D$
2: **для всех** слов $w \in W$
3: **если** $n_{dt} > 0$ **и** $n_{wt} > 0$ **то**
4: проверить независимость d и w в теме t:
составить таблицу сопряженности X_{dwt}:

	w	$W \backslash w$
d	n_{dwt}	$n_{dt} - n_{dwt}$
$D \backslash d$	$n_{wt} - n_{dwt}$	$n_t - n_{dt} - n_{wt} + n_{dwt}$

5: провести точный тест Фишера и получить рандомизированное значение уровня значимости:
$\text{Pvalue}_{dwt} = \text{Fisher}(X_{dwt})$
6: проверить гипотезу о равномерности распределения Pvalue_{dwt} на уровне значимости α с помощью критерия Пирсона:

значений. Для того, чтобы избежать дискретности, предлагается вычислять уровень значимости следующим образом:

$$Pvalue_{rand} = \sum_{Z \neq X} P(Z) + rP(X), \tag{13}$$

где суммирование вероятностей производится для таблиц сопряженности Z, не совпадающих с X и обладающих свойством (12), r—случайная величина из равномерного распределения на отрезке [0,1] (Процедура 9.1).

МОЖНО ПРОИЛЛЮСТРИРОВАТЬ НА КАРТИНКЕ

Проверка гипотезы условной независимости для темы t. Для того, чтобы определить, описывает ли тематическая модель данную тему t, предлагается алгоритм 9.2. Проверка гипотезы условной независимости слов в теме от документов сводится к проверке гипотез о независимости каждого слова от каждого документа в теме. Другими словами, для каждой пары (d,w) с $n_{dt} > 0$ и $n_{wt} > 0$ (принадлежащих теме) проводится точный тест Фишера, описанный выше. В условиях истинности нулевой гипотезы, распределение вычисленных уровней значимости Pvalue_{dwt} должно быть близко к равномерному, что проверяется с помощью критерия согласия Пирсона.

Список литературы

[1] L. Fei-Fei and P. Perona *A Bayesian hierarchical model for learning natural scene categories*, IEEE Computer Vision and Pattern Recognition, pages 524-531, 2005.

[2] D.Blei and M.Jordan *Modelling annotated data* In Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 127-134. ACM Press, 2003.

[3] E.Bart, M.Welling, and P. Perona *Unsupervised organization of image collections: Taxonomies and beyond*, Transactions on Pattern Recognition and Machine Intelligence, 2010.

[4] J.Li, C.Wang, Y.Lim, D.Blei, and L.Fei-Fei *Building and using a semantivisual image hierarchy*, In Computer Vision and Pattern Recognition, 2010.

[5] S.Deerwester; S. T. Dumais; T. K. Landauer; G. W. Furnas and R. A. Harshman *Indexing by latent semantic analysis*, Journal of the Society for Information Science, 41(6), 391-407, 1990.

[6] Papadimitriou, Christos; Raghavan, Prabhakar; Tamaki, Hisao; Vempala, Santosh *Latent Semantic Indexing: A probabilistic analysis*, Proceedings of ACM PODS, 1998.

[7] von Davier M. *Bootstrapping goodness-of-fit statistics for sparse categorical data-results of a monte carlo study*, Methods of Psychological Research Online,1997.

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

[9] Daud, Ali and Li, Juanzi and Zhou, Lizhu and Muhammad, Faqir *Knowledge discovery through directed probabilistic topic models: a survey*, Frontiers of Computer Science in China,Pp. 280-301, 2010.

[10] Hofmann, Thomas *Probabilistic latent semantic indexing*,Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Pp.50-57,1999.

[11] Dempster A.P., Laird N. M., Rubin D. B. *Maximum likelihood from incomplete data via the EM algorithm*, J. of the Royal Statistical Society, Series B., no.34, Pp.1-38, 1977.

[12] Blei, David M. and Ng, Andrew Y. and Jordan, Michael I. *Latent Dirichlet allocation*, Journal of Machine Learning Research, Pp.993-1022, 2003.

[13] Steyvers, Mark and Griffiths, Tom, *Finding scientific topics* ,Proceedings of the National Academy of Sciences, Pp.5228-5235, 2004.

[14] A. Asuncion and M. Welling and P. Smyth and Y. W. Teh, *On Smoothing and Inference for Topic Models*, Proceedings of the International Conference on Uncertainty in Artificial Intelligence, 2009.

[15] Yi Wang, *Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details*,2008.

[16] G.Golub and C.Reinsch, *Handbook for matrix computation II, Linear Algebra*, Springer-Verlag, New York, 1971.

[17] Mimno D., Blei D., *Bayesian checking for topic models*, 11th Conference on Empirical Methods in Natural Language Processing.— Association for Computantional Linguistics, 2011.—Pp. 227-237.

[18] Fisher, R. A. *On the interpretation of χ^2 from contingency tables, and the calculation of P*, Journal of the Royal Statistical Society,1922 85(1):87-94.

[19] Бриллюэн Л. *Наука и теория информации*, М.: «Государственное издательство физико-математической литературы», 1960.—391с.

[20] Gelbukh A., Sidorov G. *Zipf and heaps laws' coefficients depend on language* //Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City, Lecture Notes in Computer Science.— No. 2004.— Springer-Verlag, 2001.—P. 332–335.