Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Гринчук Алексей Валерьевич

Использование контекстной документной кластеризации для улучшения качества построения тематических моделей

010900 — Прикладные математика и физика

Бакалаврская диссертация

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

Москва
2015 г.

Содержание

1 Введение

Актуальность темы. С интенсивным развитием информационных технологий возникает задача обработки больших объёмов текстовой информации. Когда человек читает какой-либо текст, он легко может составить самое общее представление о нём — определить тему и идею, выделить ключевую информацию и общие слова. Однако, когда информации оказывается слишком много (например, труды научной конференции), на её обработку могут уйти недели, а то и месяцы. Хотелось бы иметь возможность автоматически получать краткую, но точную выжимку из всей имеющийся информации, не тратя впустую время на отсев ненужного текста. Инструментом, позволяющим решить данную задачу, является тематическое моделирование.

Хорошая тематическая модель может быть использована в таких областях, как информационный поиск, обработка больших текстовых коллекций или анализ контента социальных сетей. Однако, на данном этапе развития науки тематического моделирования остро стоит задача качественного сравнения различных моделей и выбор лучшей из них, чему и посвящена данная работа.

Задача тематического моделирования. Пусть имеется коллекция текстовых документов. В задаче тематического моделирования предполагается, что каждый документ состоит не просто из набора слов, а из некоторых тем, которые раскрываются этими словами. Таким образом, каждый документ представляет собой некоторое число тем, а каждая тема, в свою очередь, может быть представлена некоторым числом слов. Если считать, что порядок слов в документах не важен для определения их тем, то всю коллекцию можно представить в виде матрицы частот встречаемости слов в документах, а задача тематического моделирования может быть сведена к нахождению её разложения в произведение двух матриц.

Обзор литературы. В процессе поиска решения поставленной задачи было предложено множество методов, но лучше всех показали себя различные вероятностные методы, такие как вероятностный латентный семантический анализ [1] или латентое размещение Дирихле [8]. Эти методы основаны на предположении, что вся коллекция получена из некоторого вероятностного распределения на множестве слов и документов. Для нахождения решения используется итерационный EM-алгоритм.

В общем случае задача поиска матричного разложения имеет бесконечное чис-

ло решений. Одним из способов борьбы с имеющейся неопределённостью является сужение класса искомых матриц посредством наложения дополнительных ограничений. Такие ограничения называются регуляризаторами и при их правильном подборе можно прийти к лучшему решению. Описанный подход лежит в основе метода аддитивной регуляризации тематических моделей [11].

Вследствие невыпуклости задачи тематического моделирования возникает проблема сходимости решения к локальным экстремумам. Существует несколько способов борьбы с данной проблемой: мультистарт, встряхивание коэффициентов, удачная инициализация. В данной работе рассматривается поиск начального приближения из которого можно прийти к лучшему локальному максимуму и получить более качественную тематическую модель.

Ключевой идеей предложенной инициализации является использование метода контекстной документной кластеризации [2]. Предполагается, что каждая тема хорошо описывается некоторым набором специальных слов (терминов). Вместе они встречаются гораздо чаще, чем по отдельности и редко встречаются со словами из других тем. Начальное приближение задаётся распределением всех слов коллекции на множестве этих слов.

Ещё одной проблемой вероятностных тематических моделей является низкая интерпретируемость. Во всех существующих методах документы считаются "мешками слов" и не учитывают внутренней структуры текста, что с точки зрения лингвистики является очень грубым упрощением. Особенностью предлагаемой инициализации является высокая интерпретируемость и частичный отказ от формата "мешка слов".

Существует много различных метрик качества для сравнения тематических моделей [6]. В данной работе используются перплексия (показатель правдоподобия) и когерентность (показатель интерпретируемости).

Цель работы. Предложить метод поиска начального приближения тематической модели, который приводит к построению лучшей и более интерпретируемой модели по сравнению с методами, выбирающими начальное приближение произвольным образом.

Методы исследований. Алгоритм поиска лучшей инициализации использует вероятностную тематическую модель и метод контекстной документной кластеризации с квантильной регрессией.

Положения, выносимые на защиту:

- Усовершенствован критерий отбора слов в методе контекстной документной кластеризации с помощью квантильной регрессии.

- Показано, что кластеризация локальных контекстов является хорошей инициализацией для тематических моделей, которая ведёт к лучшим локальным максимумам правдоподобия.

- Разработан устойчивый алгоритм поиска хорошего начального приближения тематических моделей.

2 Постановка задачи

2.1 Вероятностная модель коллекции текстовых документов

Пусть заданы два конечных множества: D — коллекция текстовых документов d и W — множество всех её слов w (словарь коллекции). Определим вероятностную тематическую модель, исходя из следующих предположений.

1. Существует некоторое множество тем T такое, что каждое слово в каждом документе связано с некоторой темой из этого множества. Коллекция документов является случайной выборкой, полученной из дискретного распределения $p(d, w, t)$ на $D \times W \times T$.

2. Порядок слов в документах не важен. Пусть n_{dw} — число вхождений слова w в документ d, а $\hat{p}(w|d) = \frac{n_{wd}}{n_d}$ - частота встречаемости слова в документе. Тогда вся коллекция может быть представлена в виде стохастической матрицы $F = \|\hat{p}(w|d)\|$.

3. Вероятность порождения слова w в документе d темой t зависит только от темы, но не от самого документа:

$$p(w|d, t) = p(w|t)$$

2.2 Задача тематического моделирования

Исходя из гипотезы условной независимости порождения слова в документе можно записать:

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

Пусть $\Phi = \|\varphi_{wt}\| = \|p\,(w|t)\,\|$ — матрица распределения слов по темам, а $\Theta = \|\theta_{td}\| = \|p\,(t|d)\,\|$ — матрица распределения тем по документам. Тогда для матрицы коллекции F верно разложение:

$$F = \Phi \times \Theta.$$

Задача тематического моделирования заключается в определении множества тем T и восстановлении скрытых распределений $p(w|t)$ и $p(t|d)$ для всех $w \in W$ и $d \in D$. Для поиска приближённого решения максимизируется логарифм правдоподобия:

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

Максимизация производится EM-алгоритмом, в котором итерационно повторяются два шага.

На E-шаге по формуле Байеса оцениваются условные распределения латентных тем $p(t|d, w)$ для всех слов в документах:

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

На M-шаге по этим условным вероятностям вычисляются частотные оценки искомых условных вероятностей:

$$\varphi_{wt} \propto \hat{n_{wt}} = \sum_{d \in D} n_{dw} p(t|d, w),$$

$$\theta_{td} \propto \hat{n_{td}} = \sum_{w \in d} n_{dw} p(t|d, w).$$

2.3 Трудности поиска решения

При решении задачи тематического моделирования возникают две важные проблемы.

Во-первых, задача матричного разложения имеет бесконечное число решений. Действительно, пусть S - произвольная невырожденная матрица размера $|T|$. Тогда

$$F = \Phi\Theta = \Phi S^{-1} S\Theta = (\Phi S^{-1})(S\Theta) = \tilde{\Phi}\tilde{\Theta}.$$

Во-вторых, решение может оказаться неустойчивым, а EM-алгоритм — сходиться к локальным максимумам, что приводит к худшему решению задачи.

Первая проблема решается регуляризацией, а вторая — подбором качественной и хорошо интерпретируемой инициализации матриц Φ и Θ.

2.4 Функционалы качества

Для оценивания качества построенных тематических моделей и сравнения их между собой используются следующие функционалы.

Перплексия - мера несоответствия модели $p(w|d)$ терминам $w \in W$, наблюдаемым в документах $d \in D$, определяемая через логарифм правдоподобия:

$$\mathbf{P}(D;p) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d)\right).$$

Чем меньше эта величина, тем лучше модель p предсказывает появление терминов w в документах d коллекции D.

Когерентность. Тема называется *когерентной*, если термины, наиболее частые в данной теме, неслучайно часто совместно встречаются рядом в документах коллекции. Когерентность может оцениваться по сторонней коллекции, либо по той же коллекции, по которой строится модель. Наиболее адекватной мерой когерентности является *логарифм условной вероятности* (log conditional probability, LCP), оценивающий вероятность менее частого слова при условии более частого:

$$\mathbf{LCP}(t) = \sum_{i=1}^{k-1}\sum_{j=i}^{k}\log\frac{N(w_i,w_j)}{N(w_i)}.$$

2.5 Выделение узких контекстов

Определение 2.1. ***Контекстом слова*** *w называется дискрентное распределение $p(u|w)$ на W. Это вероятностное распределение слов, которые встречаются вместе со словом w в документах коллекции.*

Основной идеей контекстной документной кластеризации является выделение тех документов, которые относятся к конкретной узкой тематике. Такие документы, как правило, содержат уникальные слова (термины), которые очень хорошо определяют тематику. Вместе они встречаются гораздо чаще, чем по отдельности и редко встречаются со словами из других тем. На первом этапе задача CDC — отобрать такие слова.

Для каждого слова из словаря коллекции определим документную частоту и энтропию.

Определение 2.2. ***Документной частотой*** *называется число документов коллекции, в которых данное слово встретилось хотя бы один раз*

$$N_w = |D_w|, \quad D_w = \{d \in D : n_{dw} > 0\}.$$

Определение 2.3. ***Энтропией*** *слова w называется энтропия контекста этого слова $p(u|w)$:*

$$H(w) = -\sum_{u \in W} p(u|w) \log p(u|w).$$

Максимальное значение энтропии достигается при равномерном распределении и равно $H_{max}(w) = \log |W(D_w)|$, где $W(D_w) = \{u \in W : n_{du} > 0,\ d \in D_w\}$ — это словарь тех документов, в которых встречается слово w. Малые значения энтропии свидетельствуют о том, что контекст слова описывается относительно небольшим количеством слов, а, значит, само слово может оказаться термином в какой-то тематике.

Эмпирический закон Хипса гласит, что размер словаря коллекции равен Kn^β, где n — это размер всей коллекции в словах, а K и $\beta < 1$ - некоторые постоянные. Пусть k — средний размер в словах документа из D_w для некоторого слова w, тогда $|W(D_w)| = K(k \cdot N_w)^\beta$ и

$$H(w) \leqslant H_{max}(w) = \log |W(D_w)| = \log K(k \cdot N_w)^\beta =$$

$$= \log K + \beta \log k + \beta \log(N_w) = C + \beta \log N_w.$$

Для того, чтобы учесть зависимость между $H(w)$ и N_w, словарь коллекции разбивается на подмножества так, чтобы документные частоты всех слов из данного подмножества лежали в одном интервале:

$$W = \bigcup_i W_i, \quad W_i = \left\{w : w \in W,\ N_w^{(i)} \leqslant N_w < N_w^{(i+1)}\right\}, \quad i = 1, \ldots, r.$$

Как видно, для каждого интервала можно выбрать пороговое значение энтропии $H^{(i)}$ и отобрать все слова, при некотором $i = 1, \ldots, r$ удовлетворяющие условию:

$$\begin{cases} N_w^{(i)} \leqslant N_w < N_w^{(i+1)}, \\ H(w) \leqslant H^{(i)}. \end{cases} \tag{2.1}$$

Пороговые значения энтропии выбираются эмпирически на основании её распределения по всем документным частотам слов коллекции.

Определение 2.4. *Слово $w \in W$ называется* ***узким контекстом****, если для некоторого $i = 1, \ldots, r$ оно удовлетворяет условию (2.1). Множество всех узких контекстов коллекции обозначим за* $\mathbf{N}$.

Однако, можно предложить и другой подход к определению узких контекстов в предположении что их ровно K. Пусть, как и ранее, $W = \bigcup_i W_i$. Для каждого $i = 1, ..., r$ определим множества $\mathbf{N_i} \subset W_i$ следующим образом:

$$|\mathbf{N_i}| = \frac{K \cdot W_i}{|W|}, \quad \forall w_1 \in \mathbf{N_i},\ \forall w_2 \in W_i \backslash \mathbf{N_i} : H(w_1) < H(w_2).$$

Множество узких контекстов определим как $\mathbf{N} = \bigcup_i \mathbf{N_i}$. Этот подход хорош тем, что не требует эмпирического задания пороговых энтропий $H^{(i)}$, однако для его использования нужно хотя бы представлять, сколько тем содержит коллекция.

2.6 Кластеризация документов

Выбранные узкие контексты выступают в роли центров кластеров. В процессе кластеризации каждый документ коллекции относят к ближайшему такому центру в смысле расстояния Йенсена-Шеннона.

Определение 2.5. ***Расстоянием Йенсена-Шеннона*** *между двумя вероятностными распределениями $p_1(u)$ и $p_2(u)$ называется число*

$$JS_{\{\pi_1,\pi_2\}}[p_1, p_2] = H[\overline{p}] - \pi_1 H[p_1] - \pi_2 H[p_2],$$

где $\pi_1 \geqslant 0$, $\pi_2 \geqslant 0$, $\pi_1 + \pi_2 = 1$, $\overline{p} = \pi_1 p_1 + \pi_2 p_2$.

Расстояние Йенсена-Шеннона обладает следующими свойствами:

1. Является неотрицательной ограниченной функцией от p_1 и p_2.
2. $JS_{\{\pi_1,\pi_2\}}[p_1, p_2] = 0 \ \Leftrightarrow \ p_1 \equiv p_2$.
3. Является вогнутой функцией от π_1 и π_2 с единственным максимумом в точке $\{0.5, 0.5\}$.

Пусть $d \in D$ - документ коллекции. Тогда вероятностное распределение слов в нём задается плотностью:

$$p(u|d) = \frac{p(u,d)}{p(d)}.$$

Схожесть документа d и узкого контекста w будем вычислять как расстояние Йенсена-Шеннона между двумя вероятностными распределениями с плотностями $p(u|w)$ и $p(u|d)$. Мы будем минимизировать наибольшее расстояние, поэтому:

$$w = \arg\min_{w' \in \mathbf{N}} JS_{\{0.5,0.5\}}[p(u|w'), p(u|d)].$$

Это значит, что документ d будет отнесён к кластеру, центром которого является узкий контекст w.

2.7 Алгоритм контекстной документной кластеризации

Алгоритм контекстной кластеризации состоит из двух частей. В первой части определяются слова, являющиеся узкими контекстами. Во второй, используя полученные узкие контексты как центры кластеров, происходит жёсткая кластеризация документов коллекции. Документ относят к кластеру, центр которого находится на наименьшем расстоянии от него в смысле метрики Йенсена-Шеннона.

Алгоритм 2.1 Нахождение узких контекстов

Вход: Коллекция текстовых документов D, $\left\{N^{(i)}\right\}_{i=1}^{r}$, $\left\{H^{(i)}\right\}_{i=1}^{r}$

Выход: Множество узких контекстов $\mathbf{N} \subset W$

1: $\mathbf{N} = \emptyset$
2: **для всех** $w \in W$ **:**
3: Найти распределение $p(u|w)$
4: Подсчитать $H(w)$ и N_w
5: **если** $H(w)$ и N_w удовлетворяют (2.1) **то:**
6: $\mathbf{N} = \mathbf{N} \cup w$

Алгоритм 2.2 Контекстная документная кластеризация

Вход: Коллекция текстовых документов D, множество узких контекстов $\mathbf{N} \subset W$

Выход: Множество кластеров документов $\{C_i\}_{i=1}^{|\mathbf{N}|} \subset D$, $\bigcup_{i=1}^{|\mathbf{N}|} C_i = D$

1: Пронумеруем все элементы множества $\mathbf{N}$: $\mathbf{N} = \{w_1, w_2, ..., w_{|\mathbf{N}|}\}$

2: **для всех** $i = 1, ..., |\mathbf{N}|$ **:**

3: $\quad C_i = \emptyset$

4: **для всех** $d \in D$ **:**

5: $\quad w_i = \arg\min_{w \in \mathbf{N}} JS_{\{0.5,0.5\}}[p(u|w), p(u|d)]$

6: $\quad C_i = C_i \cup d$

3 Решение

Решение задачи можно разбить на три этапа:

1. Сегментация слов по документной частоте.
2. Выделение узких контекстов.
3. Инициализация тематической модели.

3.1 Сегментация слов по документной частоте

Для двух слов u и w коллекции подсчитаем условную вероятность $p(u|w)$:

$$p(u|w) = \frac{p(u,w)}{p(w)} = \frac{\sum_{d \in D} p(d)p(u|d)p(w|d)}{p(w)} = \frac{\sum_{d \in D} \sum_{w \in W} p(w,d)p(u|d)p(w|d)}{\sum_{d \in D} p(w,d)}.$$

Если положить вероятности $p(d)$ равными для всех документов коллекции, то можно получить более простую эмпирическую оценку:

$$p(u|w) = \frac{\sum_{d \in D_w} n_{du}}{\sum_{d \in D_w} \sum_{v \in d} n_{dv}}.$$

Подсчитав для каждого слова все такие вероятности, можно найти его энтропию:

$$H(w) = -\sum_{u \in W} p(u|w) \log p(u|w).$$

Если теперь изобразить все слова коллекции на графике в координатах $(N_w, H(w))$, то, исходя из верхней оценки энтропии слова, мы получим график некоторой сублинейной функции.

3.2 Выделение узких контекстов

Как видно из определения энтропии, чем она меньше, тем дальше распределение от равномерного, что соответствует понятию узкого контекста. Однако, нельзя выбрать один порог для всего словаря и взять слова, энтропия которых меньше порога. Слова с большей документной частотой имеют априори большую энтропию, т.к. их контексты шире.

В оригинальном методе контекстной документной кластеризации для решения этой проблемы диапазон всех документных частот разбивается на дискретные интервалы, что приводит к фрагментарному отбору слов. Кроме того, нужно задать само разбиение и пороги, что делается эмпирически для каждой конкретной коллекции.

Для того, чтобы избавиться от вышеперечисленных недостатков, в моей работе предлагается использовать квантильную регрессию. С её помощью мы для каждого значения документной частоты N мы выберем τN слов с наименьшей энтропией, где τ — квантиль регрессии.

3.3 Инициализация тематической модели

В методе контекстной документной кластеризации выделенные узкие контексты играют роль центров кластеров для дальнейшей кластеризации документнов. Однако, их отбирается гораздо больше, чем всего тем в коллекции, а потому предлагается кластеризовать их.

В данной работе кластеризуются контексты отобранных слов методом k-средних, в качестве метрики близости используется расстояние Хеллингера. Полученные в итоге центры кластеров являются хорошими начальными приближениями для словесных профилей тем, а потому задают столбцы матрицы Φ.

Инициализацию матрицы Θ вручную можно не делать, т.к. её легко можно получить из матриц Φ и F:

$$p(t|d) = \sum_{w \in W} p(t|w)p(w|d) = \sum_{w \in W} \frac{p(w|t)p(w|d)p(t)}{p(w)}.$$

3.4 Алгоритм

Объединив все три этапа, получим алгоритм поиска начального приближения тематической модели.

Алгоритм 3.1 Инициализация тематической модели

Вход: Коллекция текстовых документов D, число тем $|T|$, квантиль τ

Выход: Матрицы Φ и Θ для тематической модели

$\mathbf{N} = \emptyset$

Сегментировать слова по документной частоте N_w:

для всех $w \in W$ **:**

  Найти распределение $p(u|w)$

  Подсчитать $H(w)$ и N_w

Квантильной регрессией отобрать узкие контексты:

$\mathrm{Q}(\log N_w) = \mathrm{Quantreg}(\log N_w, H(w))$

для всех $w \in W$ **:**

  если $H(w) < Q(\log N_w)$ **то:**

    $\mathbf{N} = \mathbf{N} \cup w$

Кластеризовать контексты отобранных слов:

$\vec{t}_1, \cdots, \vec{t}_{|T|}$ — центры кластеров

Найти матрицы Φ и Θ:

$\Phi = \|\vec{t}_1 \cdots \vec{t}_{|T|}\|$

для всех $d \in D, t \in T$ **:**

  Подсчитать $p(t|d)$

$\Theta = \|p(t|d)\|$

4 Вычислительные эсперименты

4.1 Описание экспериментов

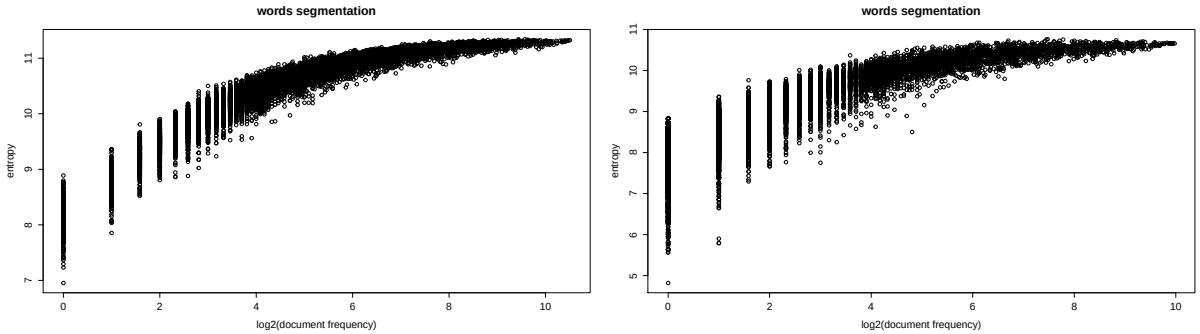
Эксперименты проводились на двух коллекциях текстовых документов, представляющих собой статьи научных конференций — англоязычной NIPS (1500 документов, 12419 уникальных слов) и русскоязычной ММРО-ИОИ (1061 документ, 17242 уникальных слова). Встречающимися вместе считались слова, которые встретились хотя бы раз в одном и том же документе.

Для отбора узких контекстов использовалась квантильная регрессия с квантилью $\tau = 0.05$. Отобранные слова кластеризовались методом k-средних, в качестве метрики близости использовалось расстояние Хеллингера для дискретных вероятностных распределений.

Тематическая модель строилась библиотекой тематического моделирования BigARTM (www.bigartm.org) без использования регуляризаторов.

4.2 Сегментация и отбор слов

Из теоретической верхней оценки $H(w) \leqslant C + \beta \log N_w$ следует, что энтропия слова является сублинейной функцией от логарифма его документной частоты. Это хорошо видно на графиках зависимости $H(\log N_w)$ для наших коллекций.



(a) NIPS (b) ММРО-ИОИ

Рис. 1: Вид зависимости энтропии от документной частоты

Для каждого такого графика узкими контекстами мы считаем слова, находящиеся в нижней части полученной "косы". Для их выделения авторами метода контекстной документной кластеризации было предложено разбить все документные частоты на интервалы и для каждого из них подобрать свою пороговую энтропию. Кроме того, что этот метод требует ручного подбора параметров, отбираются не все слова, которые нам нужны, что можно увидеть на графике.

При таком подходе теряется значительное число слов с небольшими значениями документной частоты, которые, в основном, и образуют темы. Также отбирается слишком много слов с большими значениями N_w, которые встречаются почти во всех документах коллекции и являются общеупотребительными.

Всех вышеперечисленных недостатков лишён метод отбора слов при помощи квантильной регрессии.

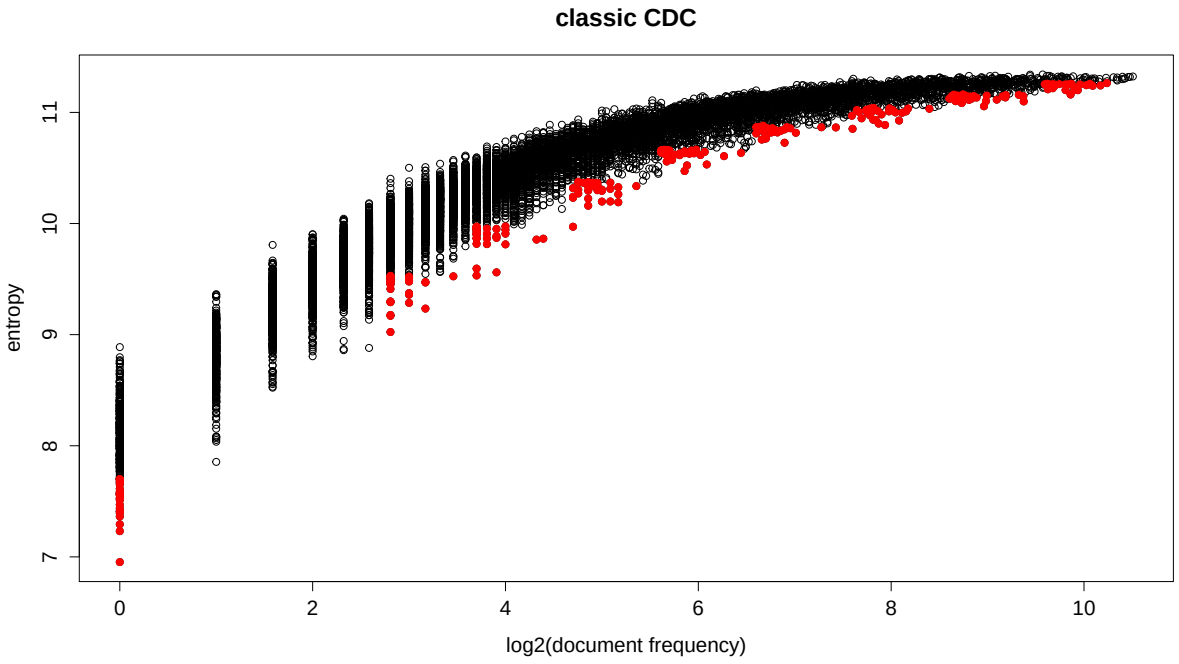


Рис. 2: Узкие контексты для разбиения N_w на интервалы на коллекции NIPS

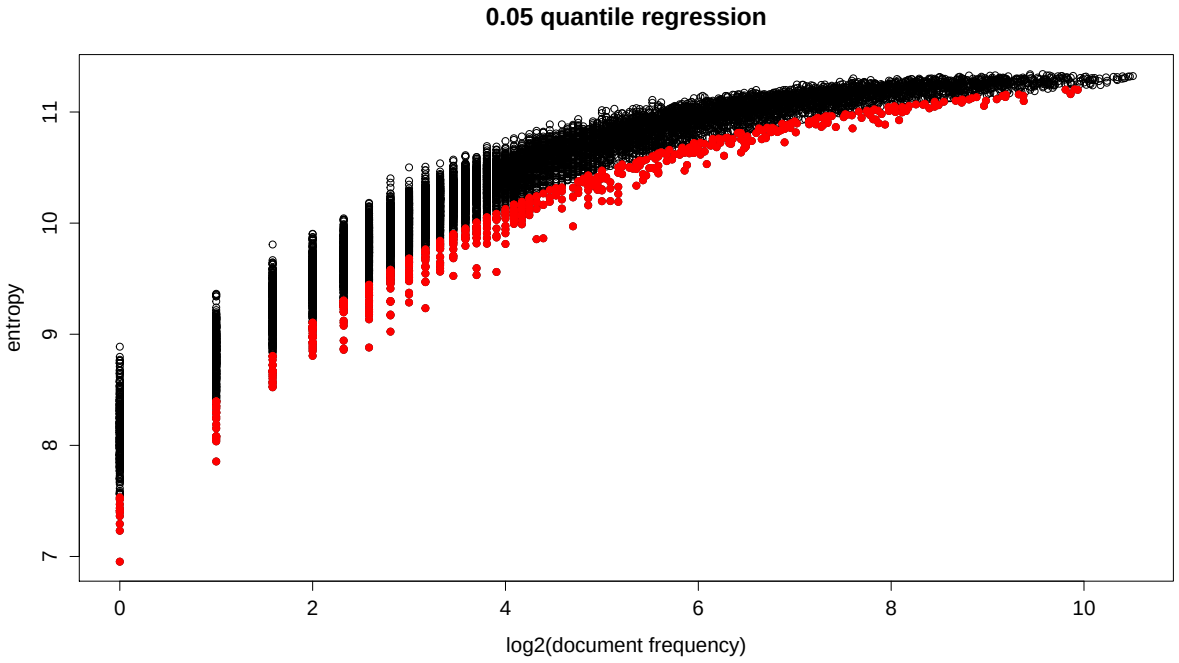


Рис. 3: Узкие контексты для квантильной регрессии на коллекции NIPS

4.3 Инициализация тематической модели

Отобранные слова кластеризовались, а центры полученных кластеров использовались для инициализации матрицы Φ. Далее полученная тематическая модель

сравнивалась с моделью, инициализируемой случайно.

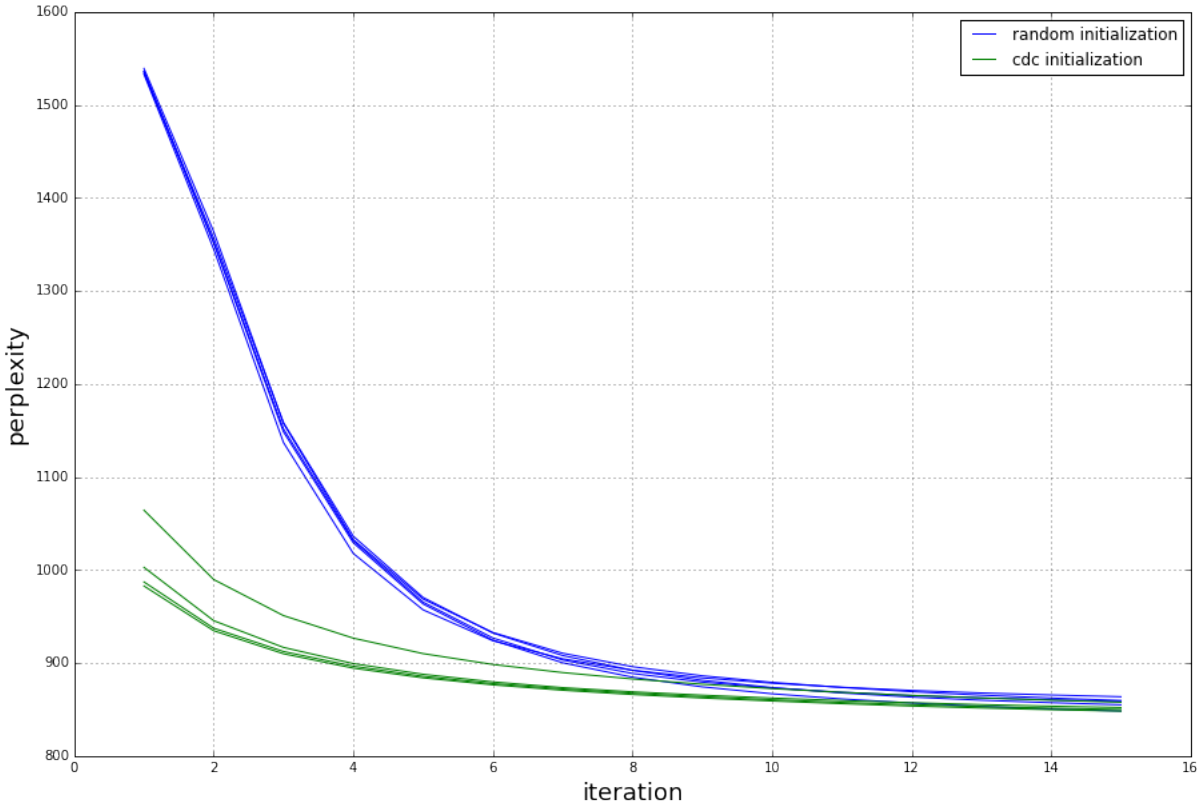


Рис. 4: Зависимость перплексии модели от числа итераций EM-алгоритма для коллекции ММРО-ИОИ

На графике показано изменение перплексии тематической модели в зависимости от числа итераций EM-алгоритма. Пучку синих линий отвечают различные случайные инициализации, а пучку зелёных — предлагаемая инициализация для различного числа итераций алгоритма кластеризации.

По графику видно, что инициализированная нами модель в целом лучше модели со случайной инициализацией. Несмотря на то, что перплексии обеих моделей сходятся к одному значению, инициализированная модель ведёт себя значительно лучше и может быть использована для экономии числа итераций в EM-алгоритме. Значение перплексии этой модели на нулевой итерации (1225) значительно ниже того же значения для случайной инициализации (16900), что говорит о хорошей её обусловленности. Ещё можно добавить, что достаточно одного-двух шагов алгоритма кластеризации для получения хорошего устойчивого решения.

4.4 Интерпретируемость

Если посмотреть на наиболее вероятные слова в темах построенных тематических моделей, то можно сделать вывод о плохой интепретируемости тем. В одних сложно определить о чём идёт речь, в других же встречается много общеупотребительных и неинформативных слов. Как показывают наблюдения, такие слова имеют большие значения документной частоты. В следующем эксперименте из коллекции ММРО-ИОИ были выброшены все слова с документной частотой $N_w > 500$ (порядка 100 слов).

Тема 9*	Тема 10	Тема 29
изображение(0.136)	точка(0.026)	изображение(0.081)
преобразование(0.020)	распознавание(0.010)	преобразование(0.021)
форма(0.015)	трёхмерный(0.010)	точка(0.017)
яркость(0.013)	объект(0.010)	объект(0.014)
пиксель(0.008)	плоскость(0.010)	метод(0.013)
координата(0.008)	фильтр(0.009)	контур(0.011)
размер(0.007)	координата(0.009)	быть(0.011)
плоскость(0.007)	изображение(0.009)	описание(0.010)
фрагмент(0.007)	поверхность(0.008)	область(0.010)
обработка(0.006)	задача(0.008)	являться(0.008)

Тема 9 является одной из тем модели, полученной на коллекции без высокочастотных слов. Её можно легко интерпретировать — речь ведётся об обработке изображений. Две другие темы выбраны из модели на всей коллекции и в числе наиболее встречаемых слов явно заметны как общеупотребительные (быть, являться), так и неинформативные (объект, задача).

Тема 19*
ладонь(0.016)
человек(0.015)
идентификация(0.014)
палец(0.014)
эксперт(0.008)
движение(0.008)
изображение(0.008)
база(0.007)
экспертный(0.007)
лицо(0.007)

Тема 12	ошибка(0.009)
изображение(0.043)	быть(0.009)
алгоритм(0.013)	ладонь(0.009)
который(0.011)	представление(0.009)
объект(0.011)	палец(0.009)
классификатор(0.011)	

В этом примере также хорошо просматривается тема идентификации человека по изображению его ладони или лица (Тема 19*). Тема представляет собой целое научное направление в области обработки изображений. Им занимается Местецкий Л. М., работы которого представлены в коллекции и, скорее всего, образовали данную тему. В модели на всей коллекции похожую тему найти не удалось. Наиболее близкая к ней Тема 12 сама по себе плохо интерпретируется и содержит лишние слова.

5 Заключение

В данной работе было исследовано влияние инициализации на качество построения тематических моделей. Был предложен метод нахождения хорошего начального приближения, объединяющего в себе кластеризацию и тематическое моделирование. Также была выполнена программная реализация предложенного метода и проведены вычислительные эксперименты, подтверждающие его применимость. В работе был тщательно изучен и усовершенствован метод контекстной документной кластеризации.

Проведённые исследования и полученные результаты являются хорошим стимулом для пересмотра существующего подхода к тематическому моделированию. В частности, к переходу от достаточно грубого формата представления текстовой коллекции – "мешка слов" к рассмотрению их локальных окружений.

Список литературы

[1] *Hoffman, T.* Probabilistic latent semantic analysis. Uncertainity in Artificial Intelligence, UAI'99, Stockholm, 1999.

[2] *Dobrynin, V., Patterson, D., Rooney, N.* Contextual document clustering. In Proceedings of the 26th European Conference on Information Retrieval Research, LNCS 2997, pp. 167-180. Berlin/Heidelberg: Springer, 2004.

[3] *Dobrynin, V., Patterson, D., Galushka, M., Rooney, N.* SOPHIA: an interactive cluster-based retrieval system for the OHSUMED collection. Information Technology in Biomedicine, IEEE Transactions on, Vol. 9, pp. 256-265, 2005.

[4] *Patterson, D. W., Rooney, N., Dobrynin, V., Galushka, M.* Sophia: A novel approach for Textual Case-based Reasoning. Leslie Pack Kaelbling and Alessandro Saffiotti, ed., IJCAI , Professional Book Center, pp. 15-20, 2005.

[5] *Rooney, N., Patterson, D., Galushka, M., Dobrynin, V., Smirnova, E.* An investigation into the stability of contextual document clustering. JASIST 59(2): pp. 256-266, 2008.

[6] *Chuang, J., Gupta, S., Manning, C. D., Heer, J.* Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment. ICML(3), JMLR.org, pp. 612-620, 2013.

[7] *Blei, D. M.* Probabilistic topic models. Communications of the ACM 55 , no. 4, pp. 77-84, 2012.

[8] *Blei, D. M., Ng, A. Y., Jordan, M. I., Lafferty, J.* Latent Dirichlet allocation. Journal of Machine Learning Research 3 , 2003.

[9] *Potapenko, A., Vorontsov, K.* Robust PLSA Performs Better Than LDA. 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. — Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, pp. 784-787, 2013.

[10] *Vorontsov, K.* Additive Regularization for Topic Models of Text Collections. Doklady Mathematics. 2014, Pleiades Publishing, Ltd., Vol. 89, No. 3, pp. 301-304, 2014.

[11] *Potapenko, A., Vorontsov, K.* Additive Regularization of Topic Models. Machine Learning Journal, Special Issue „Data Analysis and Intelligent Optimization“, Springer, 2014.