Использование контекстной документной кластеризации для улучшения качества построения тематических моделей

А. В. Гринчук

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

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

Задача тематического моделирования

Дано:

- D — множество документов коллекции,
- W — множество слов коллекции (словарь коллекции),
- n_{dw} — число вхождений слова $w \in W$ в документ $d \in D$,
- $\hat{p}(w|d) = \frac{n_{wd}}{n_d}$ - частота встречаемости слова в документе,
- $F = \|\hat{p}(w|d)\|$ — представление коллекции в виде матрицы.

Задача:

Определить множество тем T и представить матрицу F в виде

$$\underset{W\times D}{F} \approx \underset{W\times T}{\Phi} \cdot \underset{T\times D}{\Theta}.$$

$\Phi = \|p(w|t)\|$ — матрица распределения слов по темам,
$\Theta = \|p(t|d)\|$ — матрица распределения тем по документам.

Возникающие проблемы

1. Неединственность матричного разложения:

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

Способ решения проблемы: регуляризация

2. Сходимость решения к локальным экстремумам.

Способы решения проблемы:

- мультистарт
- выбивание из локальных экстремумов (jogging of weights)
- выбор начального приближения

Обзор предыдущих работ

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

Цель работы

- Предложить интерпретируемый подход к выбору тем в тематических моделях.
- Разработать метод, позволяющий находить хорошее начальное приближение матриц Φ и Θ.
- Добиться того, чтобы полученная из начального приближения тематическая модель была лучше моделей, инициализируемых стандартными методами.

Базовые определения

Определение 1

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

Определение 2

Документной частотой называется число документов коллекции, в которых данное слово встретилось

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

Определение 3

Энтропией слова w называется энтропия его контекста:

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

Понятие узкого контекста

Предположение: каждая тема содержит уникальные слова (термины), которые её хорошо определяют. Вместе они встречаются гораздо чаще, чем по отдельности и редко встречаются со словами из других тем.

Такие слова назовём **узкими контекстами**.

Идея:

Отобрать все узкие контексты и кластеризовать их. После использовать полученные распределения в качестве столбцов матрицы Φ при её инициализации.

Выделение узких контекстов

Пусть $W(D_w) = \{u \in W : n_{du} > 0,\ d \in D_w\}$ — множество всех слов, которые встречаются в документах коллекции вместе со словом w. Максимальное значение энтропии достигается, когда контекст слова является равномерным распределением и равно

$$H_{max}(w) = \log |W(D_w)|.$$

Чем меньше энтропия слова, тем меньше слов составляет его контекст. Такие слова, как правило, и образуют темы, а, значит, могут считаться узкими контекстами.

Выделение узких контекстов

По закону Хипса, словарь коллекции имеет размер Kn^{β}, где n — это размер всей коллекции в словах, а K и $\beta < 1$ — некоторые постоянные. Пусть k — средний размер в словах документов из W_d. Тогда

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

Как видно, энтропия зависит от документной частоты. Поэтому в качестве узких контекстов будем выбирать слова с минимальной энтропией для каждого значения документной частоты.

Алгоритм контекстной кластеризации слов

Выделение узких контекстов

Вход: коллекция текстовых документов

Выход: множество узких контекстов

Метод:

- сегментация слов по частоте N_w [Добрынин, 2004]
- квантильная регрессия

Инициализация тематической модели

Вход: множество узких контекстов

Выход: матрицы Φ и Θ

Метод:

- кластеризация методом k-средних
- расстояние Хеллингера

Вычислительный эксперимент: отбор слов

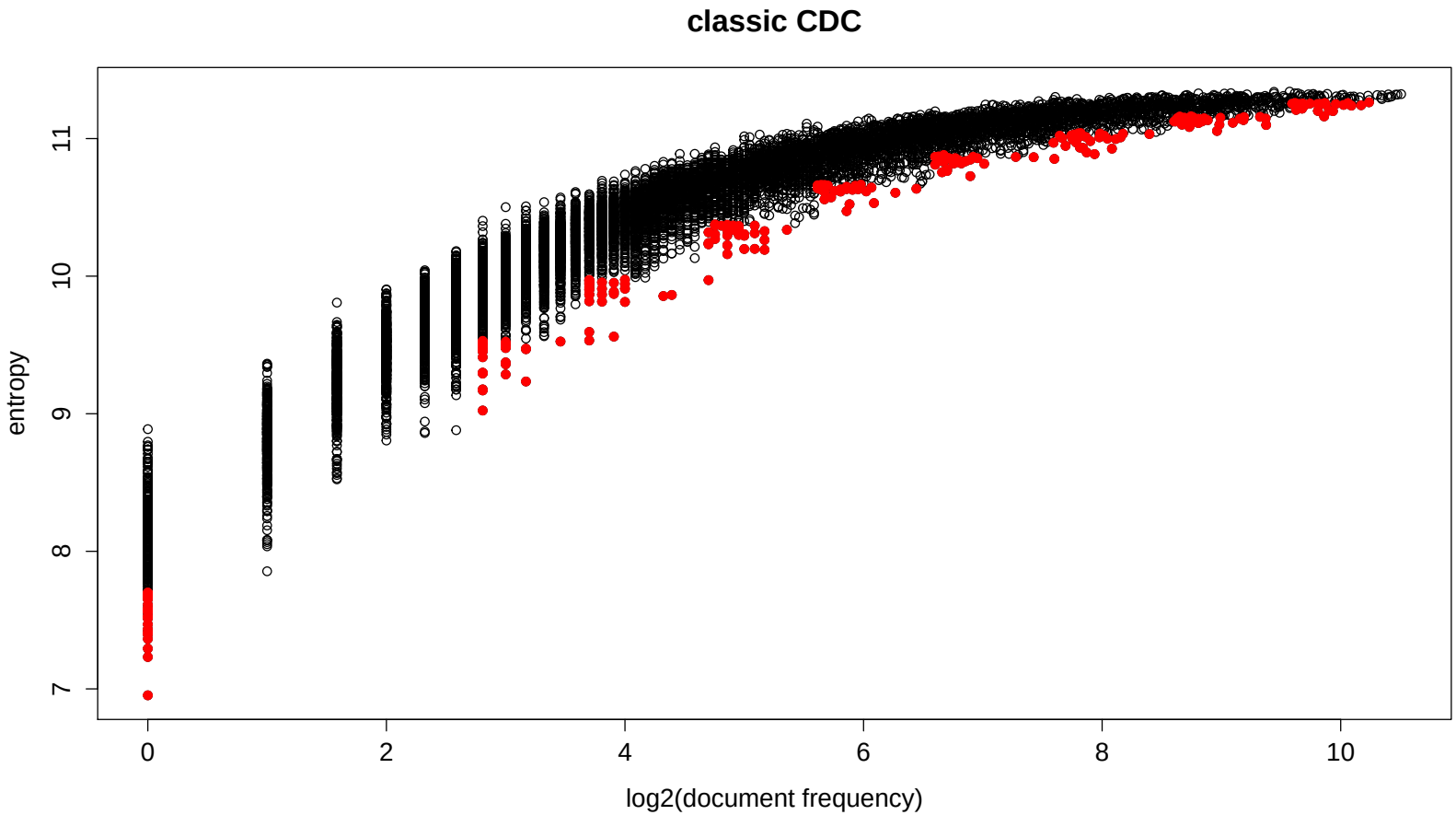

Рис.: Отбор узких контекстов классической CDC для коллекции NIPS

Вычислительный эксперимент: отбор слов

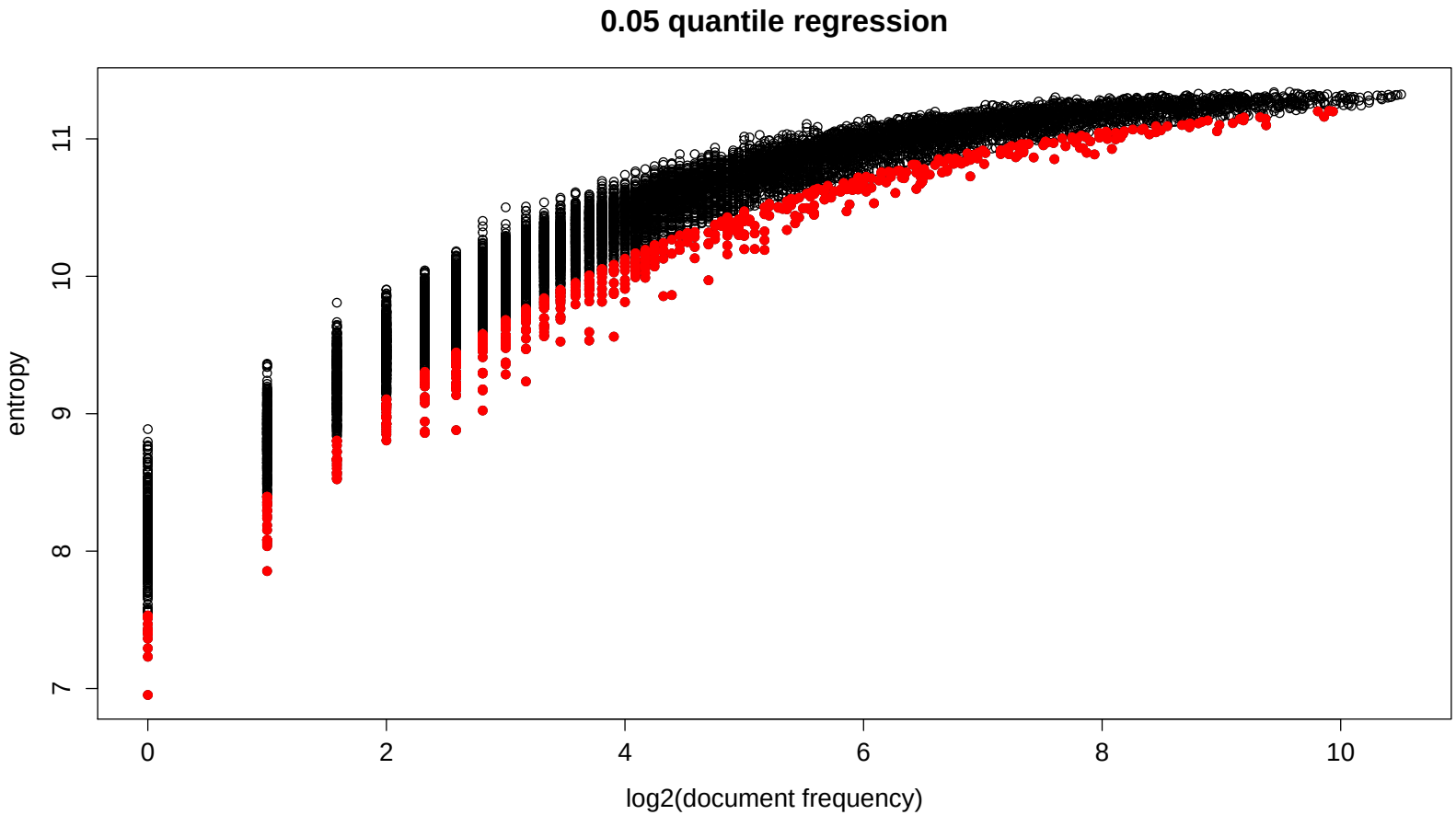

Рис.: Отбор узких контекстов квантильной регрессией для NIPS

Описание эксперимента по инициализации

- В эксперименте использовалась русскоязычная коллекция статей конференции ММРО-ИОИ.
- Встречающимеся вместе считались слова, которые встретились хотя бы раз в одном и том же документе.
- Для отбора слов применялась квантильная регрессия с квантилью $\tau = 0.05$.
- Тематическая модель строилась в библиотеке BigARTM (www.bigartm.org) без использования регуляризаторов.

Вычислительный эксперимент: перплексия

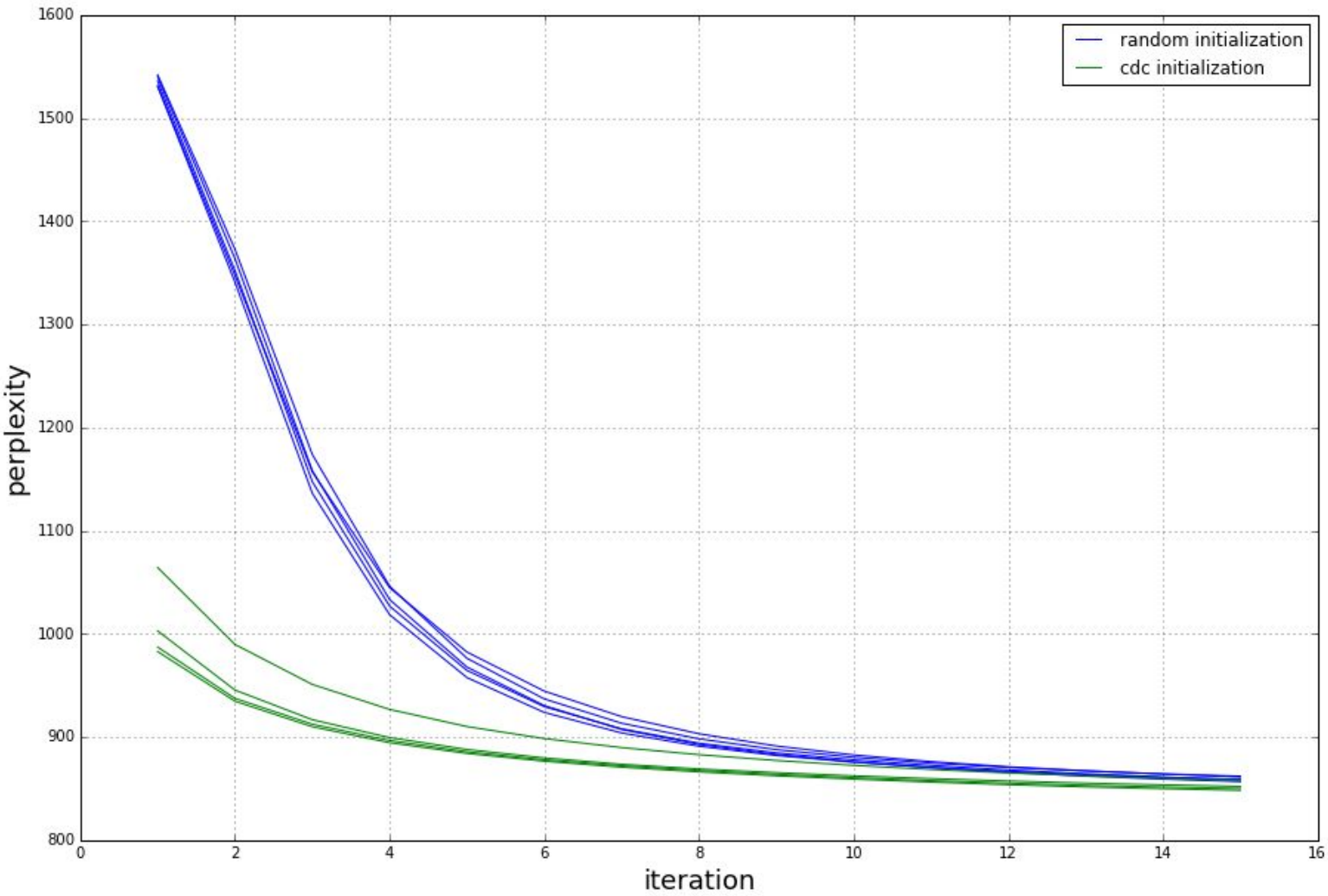

Рис.: Сравнение прплексии для различных инициализаций

Вычислительный эксперимент: интерпретируемость

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* — тема в модели без высокочастотных слов

Вычислительный эксперимент: интерпретируемость

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

Заключение

- 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.
- Показано, что удаления слов с большими значениями документной частоты улучшают интерпретируемость модели.