Построение тематической классификации коллекции документов с неизвестным числом тем

С. Ю. Лобастов

24 июня 2013 г.

Научный руководитель: К. В. Воронцов

Постановка задачи

Дано:
$D = \|n_{dw}\|$ — коллекция текстовых документов
W — словарь
Предположение:

1. Существует распределение $p(d, w, t)$ на $\mathcal{D} \times W \times \mathcal{T}$
2. $\Phi = \|p(w|t)\| \in Mat_{|W| \times |\mathcal{T}|}, \quad \Theta = \|p(t|d)\| \in Mat_{|\mathcal{T}| \times |\mathcal{D}|}$

$$D = \Phi \times \Theta$$

Найти:

1. Число тем $|\mathcal{T}|$
2. Параметры Φ и Θ

Распределение Дирихле

$\overrightarrow{\phi_t} = (\phi_{1t}, \ldots, \phi_{|W|t})$ — параметры мультиномиального распределения слов в теме t.

Предположение: $\exists\alpha : \forall t \in \mathcal{T} \quad \overrightarrow{\phi_t} \sim Dir(\alpha)$, где $Dir(\alpha)$ — распределение Дирихле:

Свойства:

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

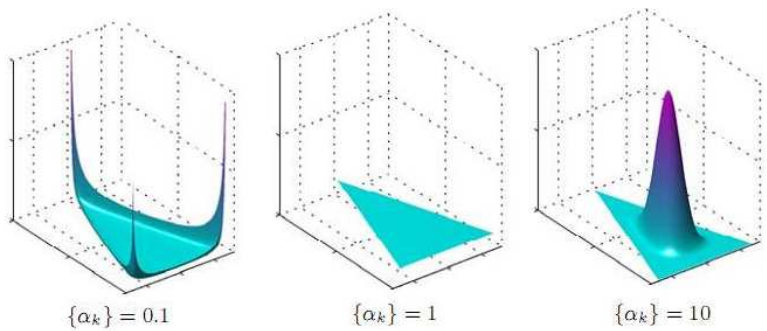


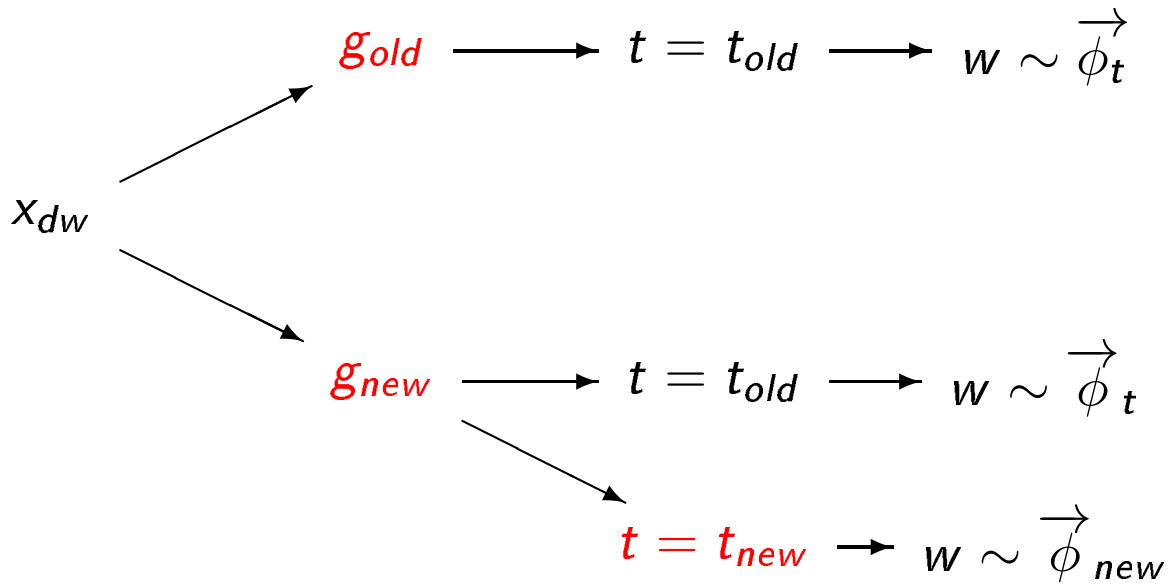
Рис.: Плотность распределения Дирихле

Порождающие модели

1. LDA. Каждое слово x_{dw} принадлежит ровно одной теме t_{dw}. Число тем $|\mathcal{T}|$ задано.

$$x_{dw} \longrightarrow t_{dw} = t_{old} \rightarrow w \sim Dir(\vec{\phi}_t)$$

2. HDP. Слова разбиты на группы, темы определяются для них.



Семплирование

1. LDA. Необходимо семплировать темы для слов.
   - $p(t|w,d) \sim \widehat{p}(t|d)\widehat{p}(w|d)$
2. HDP. Необходимо семплировать группы для слов и темы для групп.
   - $p(g|w,d) \sim \widehat{p}(t_{dg}|d)\widehat{p}(w|t_{dg})$
   - $p(t|g,d) \sim \widehat{p}(g|t)\widehat{p}(t|d)$

Анализ HDP

docs count: 50 word count: 200 words in doc: 50 $\alpha = 2.0$; $\beta = 10$; $\gamma = 0.01$; Истинное число тем — 38.

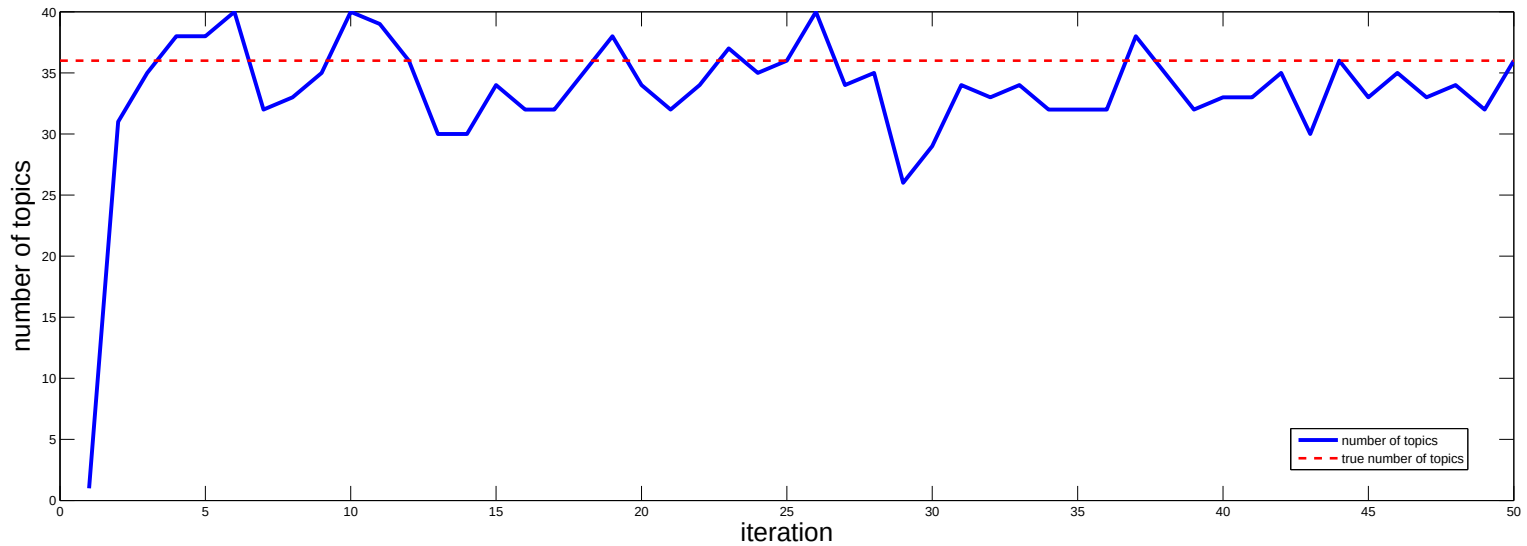


Рис.: Зависимость числа тем от номера итерации

Анализ HDP. Устойчивость.

docs count: 50 word count: 200 words in doc: 50 $\alpha = 2.0$; $\beta = 10$; $\gamma = 0.01$;
Истинное число тем — 38; число экспенриментов — 30.

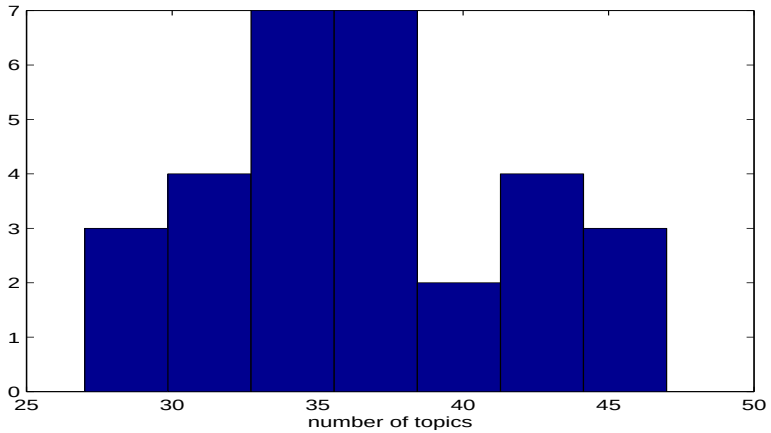


Рис.: Число тем

Дисперсия результатов HDP.

Устойчивость по параметрам

docs count: 50 word count: 200 words in doc: 50 $\alpha = 2.0$; $\beta = 10$; $\gamma = 0.01$;

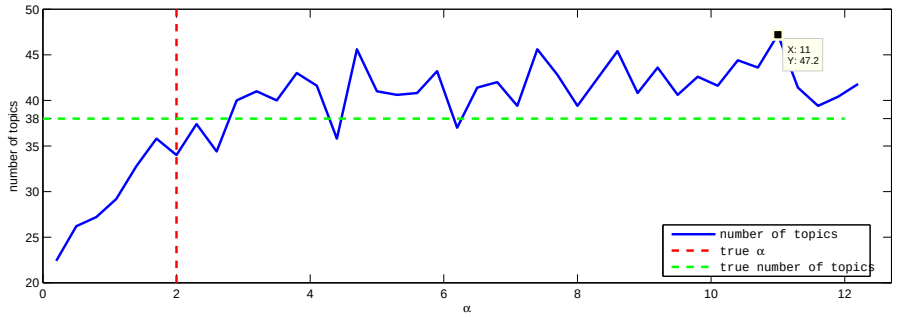


Рис.: Зависимость числа тем, определенных алгоритмом HDP, от параметра α

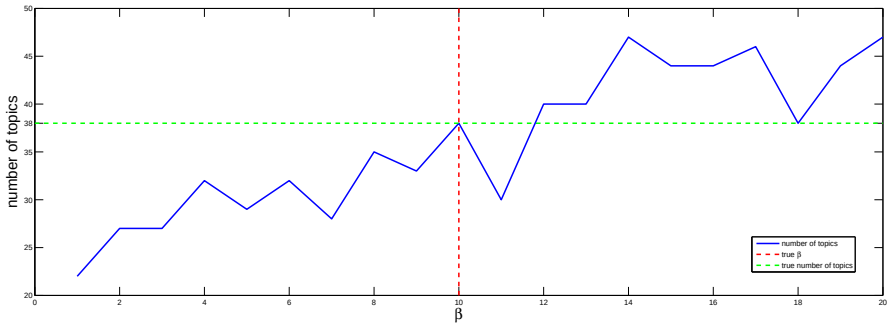


Рис.: Зависимость числа тем, определенных алгоритмом HDP, от параметра β

Запуск на разреженной коллекции

Параметры генерации: $\alpha = 0.01$; $\beta = 0.01$; Истинное число тем — 30.

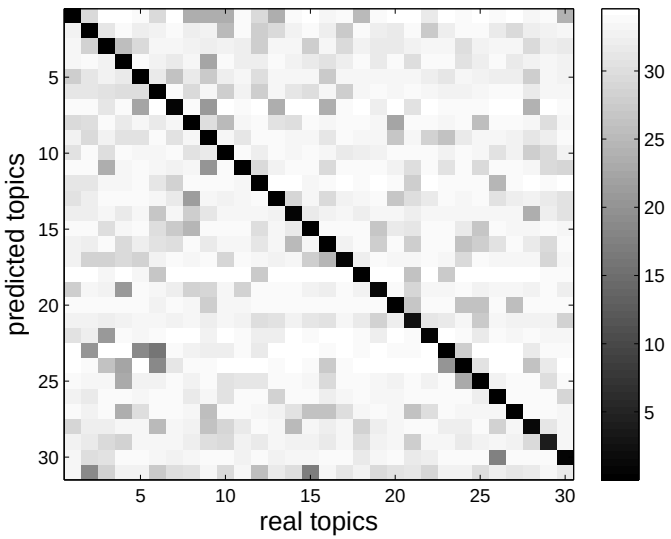


Рис.: Расстояния между истинными и найденными темами

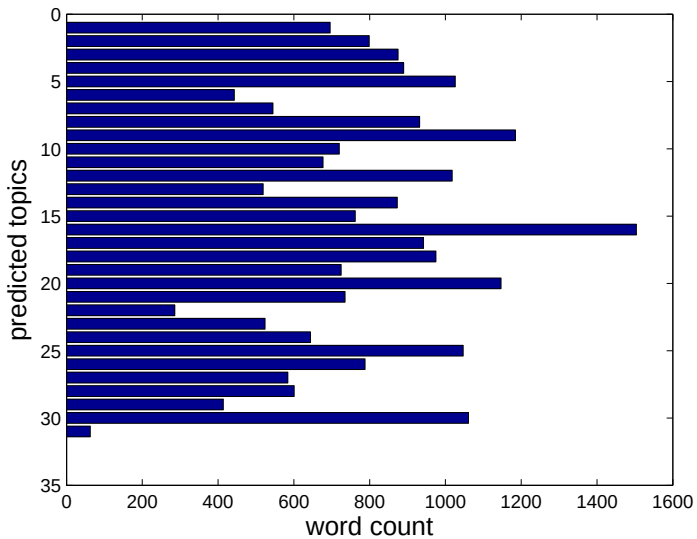


Рис.: Число слов в найденных темах

Запуск на неразреженной коллекции

Параметры генерации: $\alpha = 0.1$; $\beta = 0.1$; Истинное число тем — 10.

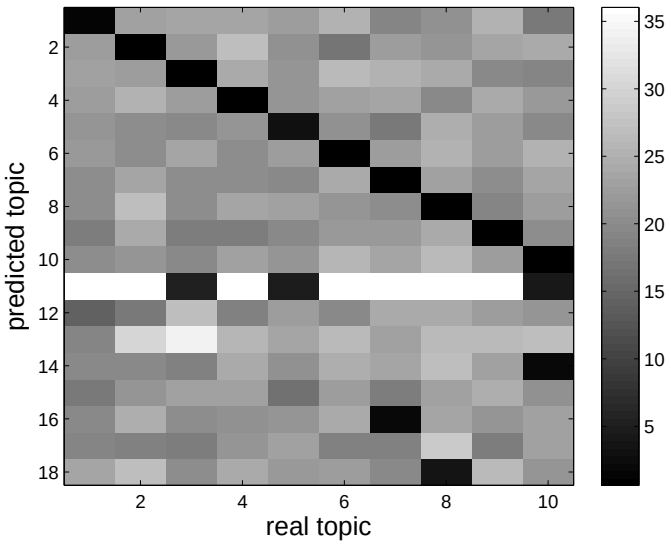


Рис.: Расстояния между истинными и найденными темами

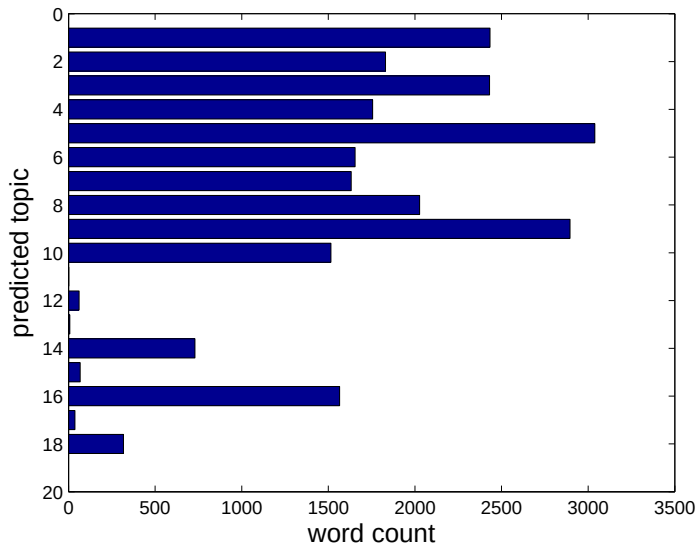


Рис.: Число слов в найденных темах

Предположение

Параметр распределения Дирихле α существенно влияет на результат: чем он больше, тем сильнее могут различаться профили темы в разных документах.
Идея: будем увеличивать этот параметр на каждой итерации.

Результат

Параметры генерации: $\alpha = 0.1$; $\beta = 0.1$; Истинное число тем — 10.

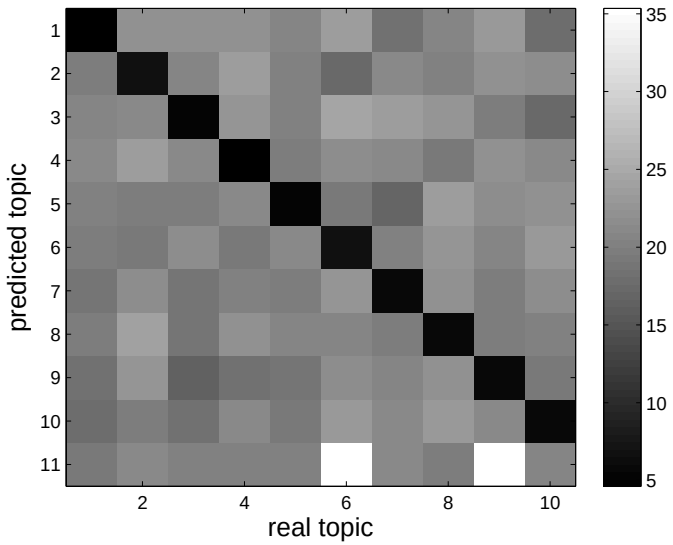


Рис.: Расстояния между истинными и найденными темами

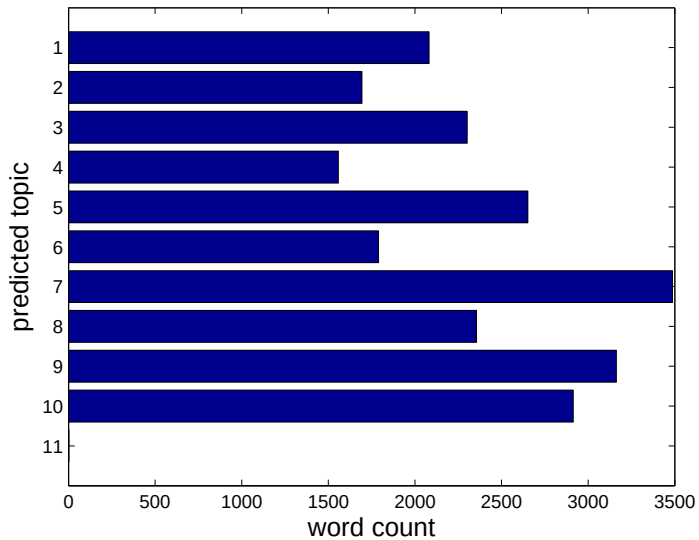


Рис.: Число слов в найденных темах

Сходимость

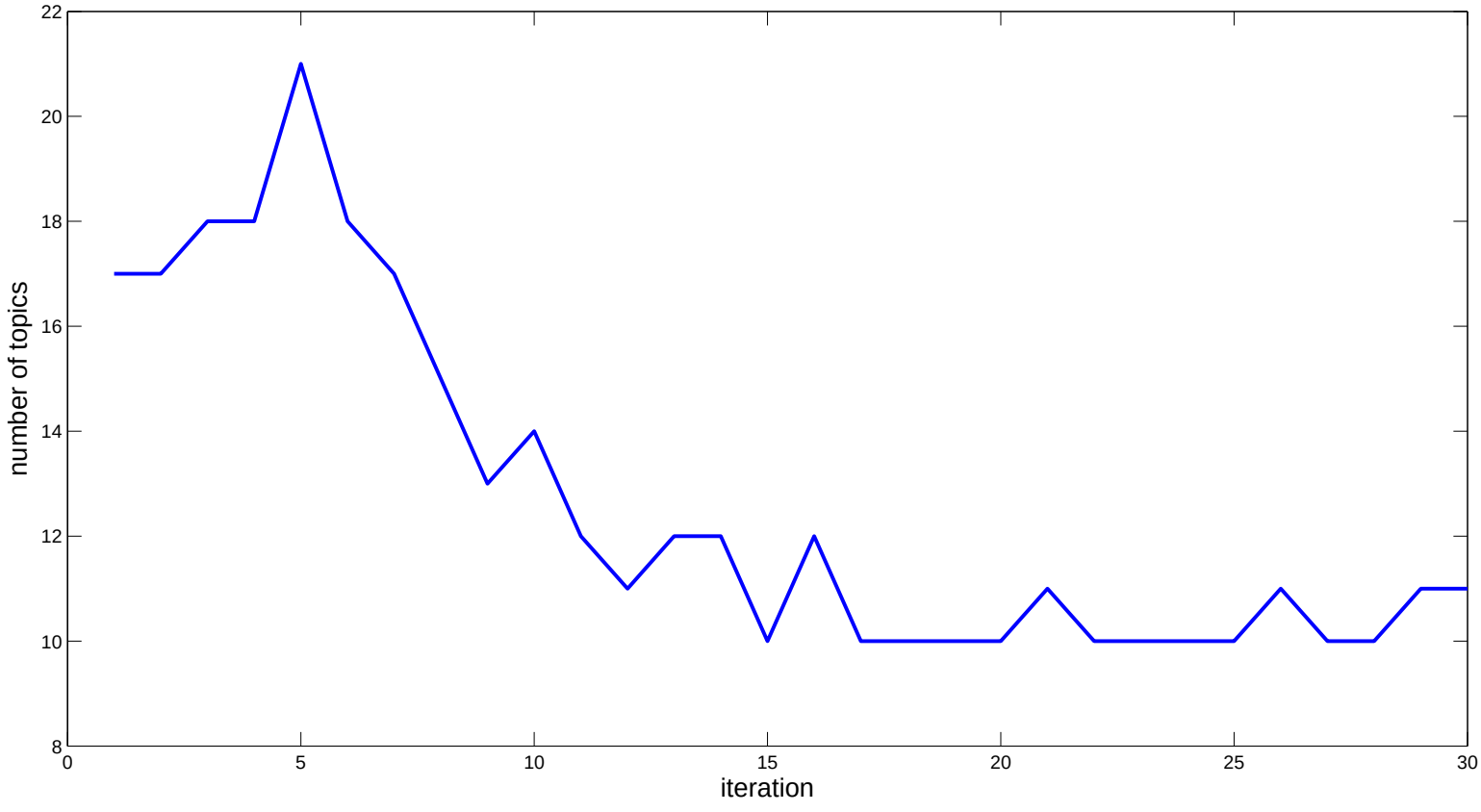


Рис.: Зависимость числа тем от номера итерации

Выводы

- Алгоритм HDP способен автоматически определять число тем, содержащихся в коллекции документов.
- Однако это вероятностный алгоритм, и дисперсия его результатов велика.
- Предложен метод решения этой проблемы путем подбора параметров.