Балансируемые тематические модели (Balanced Topic Models)

К. В. Воронцов

28 июня 2018 г.

1 Проблема балансирования тем

Вероятностные тематические модели, основанные на матричном разложении, стремятся к сбалансированному распределению коллекции по темам. Чтобы максимизировать правдоподобие, модель должна полностью задействовать все свои параметры для описания данных. Модели не выгодно сокращать число тем, так как это означало бы уменьшение числа параметров. Также ей не выгодно сокращать доли отдельных тем в коллекции, так как это привело бы к неполному использованию параметров, а в пределе и к сокращению числа тем. Из этого можно заключить, что модели выгодно использовать все темы примерно в равных долях. *Мощностью темы* t будем называть число n_t слов данной темы в коллекции. Она связана с оценкой вероятности темы: $\hat{p}(t) = \frac{n_t}{n}$. Эксперименты показывают, что отношение максимальной и минимальной мощности тем, как правило, не превышает 3–4.

В то же время, пропорции тем в реальной текстовой коллекции определяются не принципом максимума правдоподобия, а историей формирования коллекции, и могут оказаться сколь угодно несбалансированными. Если в коллекции 980 документов по биологии, 10 документов по математике и 10 документов по социологии, то в тематической модели с тремя темами все три темы будут, скорее всего, по биологии, а 20 небиологических документов распределятся между ними как попало. Для максимизации правдоподобия оказывается выгоднее описать тонкие тематические различия основной массы документов, чем более сильные различия небольшой части коллекции. Если же строить 100 тем, то, скорее всего, 98 из них окажутся по биологии и будут очень похожи друг на друга, а ещё две темы, по математике и по социологии, будут сильно отличаться от них и друг от друга.

В тематических моделях часто наблюдаются эффекты слияния и расщепления тем. Некоторые темы дублируются, тогда как в других образуются семантически разнородные смеси. Это объясняется тем, что семантически однородные темы могут отличаться по мощности в разы или даже на порядки. Но, поскольку модель выравнивает их по мощности, наиболее мощные темы окажутся разделёнными на много мелких, отличающихся незначительными нюансами, тогда как наименее мощные будут вынуждены объединиться, рис. 1. Нужен новый критерий, отличный от максимума правдоподобия, чтобы балансировать темы не по мощности, а по радиусу семантической близости, позволяя объединять семантически близкие темы и разделять семантически разнородные темы независимо от их мощности.

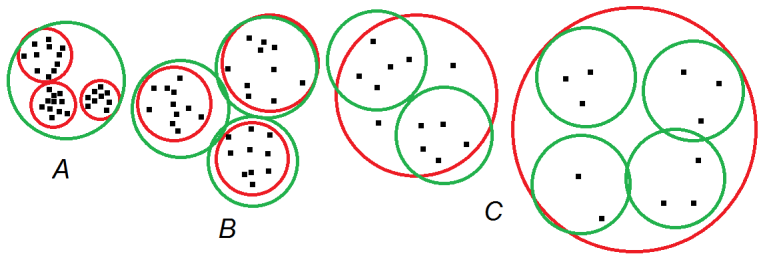

Рис. 1. Тематические модели стремятся выравнивать темы по их мощности (красные кластеры). Это приводит к появлению тем-дубликатов (A) и семантически разнородных тем (C). Лишь часть тем оказываются семантически однородными и нераздробленными (B). Выравнивание тем по радиусу семантической однородности (зелёные кластеры) должно решать обе проблемы.

2 Итеративное балансирование тем

Рассмотрим задачу максимизации регуляризированного log-правдоподобия в ARTM [1, 5]. Точка локального экстремума удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t \mid d, w)$:

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{2.1}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{2.2}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}. \tag{2.3}$$

Обозначим через n_{tdw} число раз, когда слово w в документе d относилось к теме t. Несбалансированность тем можно устранить, если домножить n_{tdw} на величину, обратно пропорциональную n_t.

Допустим, что n_{tdw} увеличилось в k_t раз по всей коллекции: $n'_{dwt} = k_t n_{dwt}$. Тогда вероятности p_{tdw} модифицируются следующим образом:

$$p'_{tdw} = \frac{n'_{dwt}}{\sum_s n'_{dws}} = \frac{k_t n_{dwt}}{\sum_s k_s n_{dws}} = k_t p_{tdw}\frac{\sum_s n_{dws}}{\sum_s k_s n_{dws}} = \underset{t\in T}{\operatorname{norm}}\big(k_t p_{tdw}\big).$$

Положим $k_t = \frac{1}{n_t}$, причём мощности тем n_t будем оценивать через немодифицированные вероятности p_{tdw}, а параметры модели — через модифицированные p'_{tdw}:

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big);$$

$$p'_{tdw} = \underset{t\in T}{\operatorname{norm}}\left(\frac{\varphi_{wt}\theta_{td}}{n_t}\right); \qquad n_t = \sum_{d\in D}\sum_{w\in d} n_{dw}p_{tdw}. \tag{2.4}$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D} n_{dw}p'_{tdw};$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w\in d} n_{dw}p'_{tdw}.$$

Можно обобщить (2.4) и ввести показатель степени $\beta \in [0, 1]$:

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}} \left(\frac{\varphi_{wt} \theta_{td}}{(n_t)^\beta} \right).$$

Тогда при $\beta \to 0$ имеем обычную модель, при $\beta \to 1$ включается максимальная балансировка тем.

3 Оценивание семантической однородности тем

Рассмотрим кластерную структуру тематической модели, рис. 1. Каждая тема t представляет собой кластер на единичном симплексе размерности $|W|$. Центром этого кластера является распределение слов в теме $p(w \,|\, t)$, точками кластера — распределения слов темы в документах $p(w \,|\, d, t)$. Согласно гипотезе условной независимости, эти распределения должны быть равны, следовательно, каждый кластер должен стягиваться в точку. Однако на практике гипотеза условной независимости может не выполняться или выполняться приблизительно. Кроме того, сами эти распределения не известны, мы можем пользоваться лишь их частотными оценками:

$$\hat{p}(w \,|\, t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(w \,|\, d, t) = \frac{n_{tdw}}{n_{dt}}.$$

Введём *нулевую гипотезу* о том, что выборка слов темы t в документе d с эмпирическим распределением $\hat{p}(w \,|\, d, t)$ порождается из распределения $\hat{p}(w \,|\, t)$. Для проверки этой гипотезы можно воспользоваться критерием согласия, основанным на статистике хи-квадрат Пирсона:

$$X^2 = n_{td} \sum_{w \in d} \frac{\big(\hat{p}(w \,|\, d, t) - \hat{p}(w \,|\, t)\big)^2}{\hat{p}(w \,|\, t)}.$$

Для корректного применения критерия Пирсона требуется, чтобы ожидаемая частота каждого слова в документе $n_{td}\, \hat{p}(w \,|\, t)$ была не менее 5. Это нестрогая рекомендация, но в случае тематического моделирования она, скорее всего, не выполняется для большинства слов, поскольку распределения $\hat{p}(w \,|\, t)$ могут быть сильно разрежены и объём словаря $|W|$ может превышать объём выборки n_{td}. В таких случаях распределение статистики X^2 не описывается асимптотикой χ^2 и может зависеть от объёма выборки [2], что затрудняет применение критерия. Обычно в таких случаях критерий модифицируют, разбивая множество слов W на непересекающиеся подмножества $u \subset W$, и определяя оценки вероятностей на этих подмножествах:

$$\hat{p}(u \,|\, t) = \frac{n_{ut}}{n_t} = \sum_{w \in u} \frac{n_{wt}}{n_t}, \qquad \hat{p}(u \,|\, d, t) = \frac{n_{tdu}}{n_{dt}} = \sum_{w \in u} \frac{n_{tdw}}{n_{dt}}, \qquad W = \bigsqcup_{u \in U} u.$$

Недостаток этого подхода в том, что результат теста может зависеть от способа разбиения множества слов на подмножества.

Имеется ещё одна причина отказаться от классического критерия Пирсона: отклонение распределения $\hat{p}(u \,|\, d, t)$ от $\hat{p}(u \,|\, t)$ можно измерить огромным числом способов. Хотелось бы иметь возможность выбрать из них лучший. Подходящее обобщение

критерия Пирсона строится с помощью параметрического семейства статистик — *дивергенции Кресси-Рида* между двумя распределениями [3]:

$$\mathrm{CR}_\lambda\big(\hat{p}(u\,|\,d,t) : \hat{p}(u\,|\,t)\big) = \frac{2n_{td}}{\lambda(\lambda+1)} \sum_{u\in U} \hat{p}(u\,|\,d,t)\left(\left(\frac{\hat{p}(u\,|\,d,t)}{\hat{p}(u\,|\,t)}\right)^{\lambda} - 1\right) = \\ = \frac{2}{\lambda(\lambda+1)} \sum_{u\in U} n_{tdu}\left(\left(\frac{n_{tdu}n_t}{n_{td}n_{ut}}\right)^{\lambda} - 1\right). \tag{3.1}$$

При некоторых значениях параметра λ дивергенция Кресси–Рида переходит в известные функции различности дискретных распределений, включая статистику хи-квадрат Пирсона ($\lambda = 1$), модифицированную статистику Пирсона ($\lambda = -1$), дивергенцию Кульбака–Лейблера ($\lambda \to 0$), расстояние Хеллингера ($\lambda = -\frac{1}{2}$), взвешенное евклидово расстояние ($\lambda = -2$). Удобство этого параметрического семейства статистик заключается в том, что свободой выбора параметра λ можно распорядиться путём оптимизации заданного внешнего критерия качества.

При условии истинности нулевой гипотезы дивергенция Кресси–Рида имеет то же асимптотическое распределение χ^2, и при тех же условиях применимости, что и статистика хи-квадрат Пирсона [3].

Осталось определиться со способом группирования слов.

Возможен подход, когда слова не группируются, а вместо асимптотического распределения статистики χ^2 используется её эмпирическое распределение, которое строится путём сэмплирования слов из распределения $p(w\,|\,t)$. Недостаток этого подхода в том, что он слишком чувствителен к незначительным нарушениям гипотезы.

Гипотеза условной независимости является избыточно сильным предположением, и многие явления естественного языка могут приводить к её нарушению. Например, явление повторяемости слов (word burstiness): если слово встретилось в тексте один раз, то оно с большой вероятностью встретится ещё [4, 7]. Тема может содержать много синонимичных терминов, но каждый автор, в силу индивидуальных особенностей стиля, может использовать только часть из них. Следствием этих явлений может быть то, что в документе встретится намного меньше слов темы, но некоторые из них будут встречаться чаще, чем можно было бы ожидать в условиях строгого выполнения нулевой гипотезы.

Таким образом, имеет смысл использовать ослабленные статистические тесты, проверяющие нулевую гипотезу против более узкого множества альтернатив. Одна из идей ослабления теста состоит в том, чтобы группировать только те слова, которые встретились в документе. Другая идея — уменьшать число подмножеств $|U|$. Можно фильтровать словарь, игнорируя при группировании нетематические слова или редкие слова, вероятность которых меньше равномерного распределения, $p(w\,|\,t) < \frac{1}{|W|}$, как предлагается в [6].

Заметим, что во всех перечисленных случаях проверка нулевой гипотезы основана на статистике Кресси–Рида (3.1). Отличия могут заключаться в способах группирования слов и методах вычисления квантилей распределения статистики.

Обозначим через S_{dt} значение статистики (3.1) для темы t и документа d.

Радиусом семантической однородности R^α_{dt} темы t для документа d с уровнем значимости α назовём $(1-\alpha)$-квантиль распределения статистики S_{dt} при условии истинности нулевой гипотезы. Радиус семантической однородности зависит от уровня

значимости, и может также зависеть от объёма выборки n_{td}. Он показывает, на какое расстояние точка кластера может отдалиться от его центра, не нарушая нулевую гипотезу при выбранном уровне значимости.

Степенью семантической неоднородности темы t назовём долю документов d, содержащих тему t, для которых значение статистики больше радиуса семантической однородности: $S_{dt} > R^{\alpha}_{dt}$. Степень семантической неоднородности показывает, какая доля точек кластера лежит за пределами радиуса однородности, тем самым нарушая нулевую гипотезу. Тему t назовём *семантически однородной*, если степень её семантической неоднородности не превышает α.

Степенью семантической загрязнённости темы t назовём долю документов d, содержащих тему t, для которых дивергенция Кресси–Рида меньше радиуса семантической однородности не только для темы t, но и для некоторой другой темы t':

$$S_{dt} = \mathrm{CR}_{\lambda}\big(\hat{p}(u \,|\, d,t) : \hat{p}(u \,|\, t)\big) < R^{\alpha}_{dt};$$
$$S_{dtt'} = \mathrm{CR}_{\lambda}\big(\hat{p}(u \,|\, d,t) : \hat{p}(u \,|\, t')\big) < R^{\alpha}_{dt'}.$$

Степень семантической загрязнённости показывает, какая доля точек кластера относится также и к другим кластерам.

4 План исследования

1. Показать проблему несбалансированности тем на примере ПостНауки. Выбрать монотематичные документы по заданному числу тем, составить из них серию из нескольких коллекций различной степени несбалансированности. Решая задачу о назначениях (венгерский алгоритм), построить зависимость числа правильно определяемых тем от степени несбалансированности коллекции.

2. Добавить регуляризатор декоррелирования, построить те же зависимости и сравнить с базовым решением. Верна ли гипотеза, что декоррелирование решает проблему несбалансированности, но лишь частично?

3. Реализовать итеративное балансирование в ЕМ-алгоритме, построить те же зависимости и сравнить с базовым решением. Верна ли гипотеза, что модификация решает проблему несбалансированности окончательно?

4. Реализовать статистические тесты для вычисления радиусов семантической однородности, степени семантической неоднородности, числа семантически неоднородных тем и семантической загрязнённости тем,

5. Построить их зависимости от числа тем. Проверить, что с увеличением числа тем степень семантической неоднородности уменьшается (что свидетельствует об увеличении точности тематического описания коллекции), в то же время степень семантической загрязнённости увеличивается (что свидетельствует о чрезмерном измельчении тем и появлении тем-дубликатов).

6. Проверить, как регуляризатор декоррелирования влияет на степени семантической неоднородности и загрязнённости тем.

7. Проверить, как итеративное балансирование тем влияет на степени семантической неоднородности и загрязнённости тем.

8. Возможно, предложить альтернативные постановки задачи тематического моделирования, основанные на принципе балансирования тем по радиусу, а не по мощности.

9. Можно ли обосновать EM-алгоритм с итеративным перевзвешиванием (Iteratively Reweighted EM) через модификацию функционала log-правдоподобия, а не формулы E-шага, что выглядит как эвристика.

Список литературы

[1] *Воронцов К. В.* Обзор вероятностных тематических моделей // Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Под ред. Е. И. Большакова, К. В. Воронцов, М. И. Ионов, Э. С. Клышинский, Н. В. Лукашевич. — М.: Изд-во НИУ ВШЭ, 2017. — 272 с.

[2] *Целых В. Р., Воронцов К. В.* Критерии согласия для разреженных дискретных распределений и их применение в тематическом моделировании // *Машинное обучение и анализ данных.* — 2012. — Т. 1, № 4. — С. 437–447.

[3] *Cressie N., Read T. R. C.* Multinomial goodness-of-fit tests // *Journal of the Royal Statistical Society, Series B.* — 1984. — Vol. 46, no. 3. — Pp. 440–464.

[4] *Doyle G., Elkan C.* Accounting for burstiness in topic models // Proceedings of the 26th Annual International Conference on Machine Learning. — ICML'09. — New York, NY, USA: ACM, 2009. — Pp. 281–288.

[5] *Kochedykov D., Apishev M., Golitsyn L., Vorontsov K.* Fast and modular regularized topic modelling // Proceeding Of The 21St Conference Of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW). Helsinki, Finland, November 6-10, 2017. — IEEE, 2017. — P. 182–193.

[6] *Koltcov S., Koltsova O., Nikolenko S.* Latent Dirichlet allocation: Stability and applications to studies of user-generated content // Proceedings of the 2014 ACM Conference on Web Science. — WebSci'14. — New York, NY, USA: ACM, 2014. — Pp. 161–165.

[7] *Lei S., Zhang J., Weng S., Zhang C.* Topic model with constrainted word burstiness intensities // The 2011 International Joint Conference on Neural Networks. — 2011. — Pp. 68–74.