Регуляризация вероятностной тематической модели для выделения ядер тем

Потапенко Анна Александровна

Студентка

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: anya_potapenko@mail.ru.ru

Вероятностная тематическая модель коллекции текстовых документов позволяет описать каждый документ d дискретным распределением $p(t|d)$ на множестве латентных тем, а каждую тему t — дискретным распределением $p(w|t)$ на множестве слов. Такое представление оказывается полезным в задачах информационного поиска, классификации, категоризации, аннотирования документов.

Задача построения тематической модели является некорректно поставленной, так как имеет много решений. Поэтому наряду с максимизацией правдоподобия необходимо вводить дополнительные регуляризирующие критерии. Важным требованием является интерпретируемость тем. Каждая тема должна состоять из характерных терминов своей предметной области, которую специалисты в состоянии обособить и дать ей адекватное название. Интерпретируемость является трудно формализуемым понятием. Известные способы количественного оценивания интерпретируемости требуют привлечения экспертов-асессоров. Известно также, что критерий когерентности, вычисляемый полностью автоматически, неплохо коррелирует с экспертными оценками метода интрузий. Однако они строятся лишь по 5–10 верхним словам в каждой теме, при этом не оценивается ни размер темы, ни её интерпретируемость в целом.

В данной работе предлагается другая формализация, основанная на предположении, что интерпретируемая тема имеет *ядро* из характерных слов, отличающих данную тему от остальных. Будем считать, что слово w относится к ядру W_t темы t, если в распределении $p(t|w)$ не более двух тем, включая t, имеют суммарную вероятность 0.5. На основе понятия ядра вводятся две новые меры интерпретируемости — средняя *чистота* и *контрастность* тем. *Чистота темы* — это суммарная вероятность слов ядра $\sum_{t\in W_t} p(w|t)$. *Контрастность темы* характеризует различность ядер — это средняя вероятность данной темы для слов ядра, $|W_t|^{-1}\sum_{t\in W_t} p(t|w)$.

Для улучшения интерпретируемости модели предлагается использовать *аддитивную регуляризацию* [1], комбинируя несколько регуляризаторов, повышающих чистоту и контрастность тем.

1. Для *разреживания* распределений $p(w|t)$ и $p(t|d)$ вводится регуляризатор, максимизирующий сумму дивергенций Кульбака–Лейблера между искомыми распределениями и равномерным распределением. В результате темы очищаются от слишком редких слов, а документы приписываются к малому числу основных тем.

2. Для повышения различности тем используется регуляризатор *декоррелирования*, минимизирующий сумму попарных ковариаций между распределениями $p(w|t)$ [3]. Он приводит к вытеснению из тем общеупотребительных слов, частотных по всей коллекции.

3. Действие разреживающего и декоррелирующего регуляризаторов на основные *предметные* темы компенсируется введением дополнительных *фоновых* тем, к которым применяется регуляризатор *сглаживания*. Он притягивает распределения $p(w|t)$ фоновых тем к общему распределению слов в коллекции.

На рис. 1 регуляризованная модель сравнивается с базовой моделью PLSA. Эксперименты проводились на коллекции 1700 статей научной конференции NIPS (Neural Information Processing Systems) на английском языке. Число тем равно 100 (90 предметных и 10 фоновых), число итераций 40. Качество тематической модели измеряется множеством показателей.

Контрольная перплексия моделей различается несущественно. Разреженность измеряется долей нулевых вероятностей и достигает в регуляризованной модели 96% для $p(w|t)$ и 87% для $p(t|d)$, в то время как в PLSA разреженности нет. Существенно улучшается качество ядер тем: чистота увеличивается с 13% до 65%, контрастность с 41% до 59%, размер ядра сокращается со 100 до 69 слов. Согласно [4], с асессорскими оценками интерпретируемости хорошо коррелирует *когерентность*, оценивающая частоту совместной встречаемости слов темы в коллекции текстов. Все когерентности улучшаются: по 10 и по 100 наиболее вероятным словам в темах и по ядрам тем.

Основной эффект предложенной комбинации регуляризаторов заключается в очищении предметных тем от слов общей лексики, которые концентрируются в фоновых темах. В результате улучшается специфичность и интерпретируемость предметных тем.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (проекты 14-07-31240, 14-07-31176)

Иллюстрации

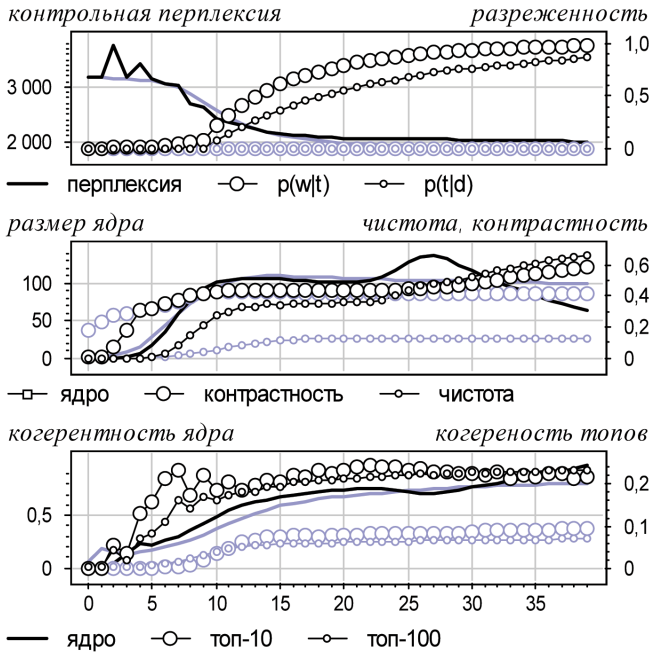

Рис. 1: PLSA (серым) и регуляризованная модель (черным).

Литература

1. Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. Т. 455, № 3.

2. Hofmann T. Probabilistic latent semantic indexing. // 22nd ACM SIGIR conference, 1999, pp. 50–57.

3. Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th International Symposium on ISCSLP, 2010, pp. 224–228.

4. Newman D., Lau J. H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // HLT '10 Human Language Technologies, 2010, pp. 100–108.