РЕГУЛЯРИЗАЦИЯ ВЕРОЯТНОСТНЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ПОВЫШЕНИЯ ИНТЕРПРЕТИРУЕМОСТИ И ОПРЕДЕЛЕНИЯ ЧИСЛА ТЕМ[1]

Воронцов К. В. (voron@forecsys.ru)

Вычислительный центр им. А. А. Дородницына РАН;
Московский Физико-Технический Институт, Москва, Россия

Потапенко А. А. (anya_potapenko@mail.ru)

Московский Государственный Университет
им. М. В. Ломоносова, Москва, Россия

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематики коллекций документов. Задача построения тематической модели имеет бесконечно много решений, что приводит к неустойчивости и плохой интерпретируемости тем. Для решения этих проблем применяется новый многокритериальный подход — аддитивная регуляризация тематических моделей (ARTM). Вводятся четыре регуляризатора: для выделения слов общей лексики в отдельные фоновые темы, для повышения разреженности и различности основных предметных тем, для удаления незначимых тем. В экспериментах показывается, что комбинирование этих регуляризаторов улучшает разреженность, когерентность, чистоту и контрастность тем без значимого ухудшения правдоподобия модели.

Ключевые слова: вероятностная тематическая модель, латентное размещение Дирихле, вероятностный латентный семантический анализ, регуляризация

REGULARIZATION OF PROBABILISTIC TOPIC MODELS TO IMPROVE INTERPRETABILITY AND DETERMINE THE NUMBER OF TOPICS

Vorontsov K. V. (voron@forecsys.ru)

Dorodnicyn Computing Centre of RAS;
Moscow Institute of Physics and Technology, Moscow, Russia

Potapenko A. A. (anya_potapenko@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований, проекты 14-07-00847, 14-07-00908, 14-07-31176.

Probabilistic topic modeling is a rapidly developing branch of statistical text analysis. The topic model uncovers a hidden thematic structure of the text collection. Learning a topic model from a document collection has an infinite set of solutions. The nonuniqueness results in a weak interpretability and instability of the solution. To tackle these problems we use a new multi-objective approach — Additive Regularization of Topic Models (ARTM). ARTM is a non-Bayesian framework free of redundant probabilistic assumptions, which dramatically simplifies the inference of topic models and makes topic models easy to design, infer, and explain. With ARTM we combine four regularizers to concentrate common vocabulary words in background topics, to make domain topics sparse and distinct, and to eliminate insignificant topics. In our experiments the combination of the regularizers improves sparsity, coherence, purity, and contrast criteria at once almost without any loss of the perplexity.

Keywords: probabilistic topic modeling, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, regularization, EM-algorithm

1. Введение

Вероятностное тематическое моделирование (probabilistic topic modeling) — это современный мощный инструментарий статистического анализа текстов, предназначенный для выявления латентных тем в коллекциях документов [Blei 2012]. *Вероятностная тематическая модель* (ВТМ) определяет тему (topic) как совокупность слов, которые часто употребляются совместно в документах коллекции. Например, в коллекциях научных публикаций темы могут соответствовать явлениям, теориям, методам, при описании которых используется устоявшаяся терминология. В коллекциях новостных сообщений темы могут соответствовать событиям, странам, компаниям, персонам и т. д.

ВТМ осуществляет «мягкую» кластеризацию слов и документов по кластерам-темам. «Мягкость» означает, что слово или документ могут относиться к нескольким кластерам-темам с различными вероятностями. Тем самым выявляется тематическая структура документов, а также решаются проблемы синонимии и омонимии, возникающие при обычной «жёсткой» кластеризации. Синонимы, часто употребляющиеся в схожих контекстах, с большой вероятностью попадают в одну тему. Омонимы, употребляющиеся в разных контекстах, распределяются между несколькими темами соответственно частоте их употребления.

Вероятностный латентный семантический анализ (probabilistic latent semantic analysis) PLSA [Hofmann 1999] и *латентное размещение Дирихле* (latent Dirichlet allocation) LDA [Blei 2003] считаются стандартными методами вероятностного тематического моделирования и часто используются в прикладных исследованиях. В литературе описаны сотни их обобщений и модификаций [Daud 2010], имеются доступные реализации. Несмотря на интенсивный поток исследований в этой области, многие проблемы, в частности, проблемы неустойчивости и слабой интерпретируемости тем, пока не имеют окончательного решения.

Интерпретируемость тем является плохо формализуемым требованием. Предполагается, что, увидев список наиболее частотных слов и документов темы, человек сможет понять, о чём эта тема, дать ей адекватное название, определить более общие, более частные или близкие темы. Интерпретируемость тем важна для приложений тематического моделирования — информационного поиска, категоризации, аннотирования, сегментации текстов. Интерпретируемость тем позволяет систематизировать, визуализировать, объяснять результаты, выдаваемые пользователю информационной системы.

Однако темы, найденные с помощью ВТМ, часто оказываются непонятными, содержат слишком много слов, включают слова общей лексики, кажутся смесью нескольких слабо связанных тем, оказываются слишком похожими на другие темы. Более того, многократное обучение модели по одной и той же коллекции может давать совершенно разные темы в зависимости от случайного начального приближения. Исследователи либо мирятся с этими недостатками, не добиваясь понятности латентных тем, либо отказываются от применения ВТМ, не находя достойных альтернатив доступным реализациям PLSA и LDA.

Фундаментальная причина этих недостатков в том, что задача построения ВТМ по коллекции документов имеет бесконечно много решений, лишь малая доля которых интерпретируемы. Алгоритм оптимизации ВТМ выдаёт некоторое произвольное решение из этого множества.

Задачи, решение которых не существует, не единственно или не устойчиво, в математике принято называть *некорректно поставленными (по Адамару)*. Известен общий подход к их решению, называемый *регуляризацией*. Он заключается в том, что для выбора наилучшего решения задаются дополнительные критерии оптимальности, учитывающие специфические требования решаемой задачи и называемые *регуляризаторами*. Если вводится несколько критериев, то задача оптимизации становится многокритериальной. В данной работе рассматриваются требования интерпретируемости и предлагается их формализация в виде набора из четырёх регуляризаторов.

К сожалению, возможности гибкого введения регуляризаторов в PLSA и LDA не предусмотрены ни в теории, ни в реализациях. Современные ВТМ основаны на байесовском подходе, в котором комбинирование регуляризаторов вызывает структурные изменения модели и наталкивается на значительные технические трудности. Попытки построения многоцелевых ВТМ на основе LDA и генетических алгоритмов [Khalifa 2013] представляются громоздкими и вычислительно неэффективными. Большинство байесовских моделей, начиная с LDA, используют в качестве основного регуляризатора априорное распределение Дирихле. Это довольно сильное вероятностное допущение, которое не имеет убедительных лингвистических обоснований, не улучшает интерпретируемость и устойчивость модели, и затрудняет комбинирование регуляризаторов.

В данной работе для комбинирования регуляризаторов применяется новый подход, альтернативный байесовскому — *аддитивная регуляризация тематических моделей*, ARTM [Vorontsov 2014]. Он свободен от избыточных вероятностных допущений, не требует введения распределений Дирихле и позволяет использовать регуляризаторы, вообще не имеющие вероятностной

интерпретации. Включение ещё одного регуляризатора в модель выполняется «в одну строку» по готовым формулам, намного проще, чем в байесовском подходе.

Предлагаемый подход к повышению интерпретируемости основан на предположении, что если тема интерпретируема, то в ней имеется *ядро* — множество характерных слов, являющихся терминами определённой предметной области, которые с большой вероятностью употребляются в данной теме и практически не употребляются в других темах. Отсюда вытекают требования разреживания и повышения различности тем, переноса слов общей лексики в отдельные «фоновые» темы, и удаления незначимых тем. В данной работе эти требования формализуются с помощью комбинации четырёх регуляризаторов.

Большинство существующих методов оценивания интерпретируемости основано на привлечении экспертов-асессоров. В исследовании [Newman, 2009] экспертам предлагалось непосредственно оценивать полезность тем по трёхбалльной шкале. В методе word intrusion [Chang 2009] для каждой найденной темы составляется список из 10 наиболее частотных слов, в который внедряется одно случайное слово. Тема считается интерпретируемой, если подавляющее большинство экспертов правильно указывают лишнее слово. Экспертные подходы необходимы на стадии исследований, но затрудняют создание полностью автоматических технологий построения ВТМ. В серии работ [Newman 2009, Newman 2010, Mimno 2011] удалось найти величину, которая хорошо коррелирует с экспертными оценками интерпретируемости, и при этом вычисляется по коллекции автоматически. Это *когерентность* (coherence), оценивающая, насколько часто наиболее вероятные слова темы встречаются рядом в данной коллекции или в Википедии. Когерентность на сегодняшний день остается основной мерой интерпретируемости, вычисляемой автоматически.

В данной работе для оценивания тематической модели используются стандартные меры качества (*контрольная перплексия, когерентность*) и предлагаются новые меры интерпретируемости тем (*размер ядра, чистота и контрастность*), не требующие привлечения асессоров.

Эксперименты на коллекции англоязычных статей научной конференции NIPS показывают, что комбинирование регуляризаторов позволяет строить сильно разреженные модели с лучшими показателями интерпретируемости, без значимого ухудшения правдоподобия (перплексии) модели.

2. Тематическая модель PLSA

Вероятностная тематическая модель (ВТМ) описывает процесс пословного порождения документов. Пусть каждая тема t задана условным распределением $\phi_{wt} = p(w|t)$ на множестве всех слов w, каждый документ d задан условным распределением $\theta_{td} = p(t|d)$ на множестве всех тем. Моделируется процесс порождения коллекции. Для получения очередного слова сначала выбирается тема из распределения $\theta_{td} = p(t|d)$, затем выбирается само слово из распределения . В результате каждый документ d описывается распределением $\phi_{wt} = p(w|t)$.

$$p(w|d) = \sum_t p(w|t)p(t|d) = \sum_t \phi_{wt}\theta_{td} \tag{1}$$

Построение ВТМ является обратной задачей по отношению к описанному порождающему процессу. По заданной коллекции документов требуется найти параметры модели ϕ_{wt}, θ_{td} и определить число тем. Это задача стохастического матричного разложения. Заданную матрицу вероятностей слов в документах $F = \|p(w|d)\|$ требуется представить в виде произведения двух матриц меньших размеров — матрицы вероятностей слов в темах $\Phi = \|\phi_{wt}\|$ и матрицы вероятностей тем в документах $\Theta = \|\theta_{td}\|$. Данная задача является некорректно поставленной, поскольку, если пара матриц (Φ, Θ) является её решением, то пары матриц вида $(\Phi S, S^{-1}\Theta)$ при некоторых S также будут её решениями.

В данной работе принимается гипотеза «мешка слов» — предположение, что порядок слов не важен для определения тем документа. Коллекция задаётся частотами n_{dw} слов w в документах d. Заметим, что многие известные ВТМ отходят от гипотезы «мешка слов», учитывая порядок слов, синтаксис предложений, выделяя вместо слов коллокации или словосочетания. Поэтому нельзя говорить, что гипотеза «мешка слов» является ограничением для всех ВТМ.

Для оценивания параметров модели $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$ решается задача максимизации логарифмированного правдоподобия:

$$L(\Phi, \Theta) = \sum_d \sum_w n_{dw} \log \sum_t \phi_{wt}\theta_{td} \ \to \max_{\Phi,\Theta}.$$

Обычно для этого применяется итерационный процесс, называемый *EM-алгоритмом*. Его легко объяснить, не прибегая к строгим выкладкам. Процесс начинается со случайной инициализации параметров модели ϕ_{wt} и θ_{td}. Каждая итерация состоит из двух шагов, «E» (expectation) и «M» (maximization).

На E-шаге по формуле Байеса оценивается вероятность $p(t|d, w)$ и число вхождений n_{dwt} каждого слова w в каждый документ d, связанных с темой t:

$$n_{dwt} = n_{dw}p(t|d, w); p(t|d, w) = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}} \tag{2}$$

На M-шаге параметры ϕ_{wt} и θ_{td} вычисляются как частотные оценки соответствующих условных вероятностей. Значение ϕ_{wt} пропорционально числу раз n_{wt}, когда употребление слова w было связано с темой t. Значение θ_{td} пропорционально числу слов n_{dt} в документе d, относящихся к теме t:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \quad n_{wt} = \sum_d n_{dwt}, \quad n_t = \sum_{d,w} n_{dwt};$$

$$\theta_{td} = \frac{n_{dt}}{n_d}, \quad n_{dt} = \sum_w n_{dwt}, \quad n_d = \sum_{w,t} n_{dwt}.$$

Для краткости эти формулы записывают через знак пропорциональности $\propto$, позволяющий опускать нормировочный множитель:

$$\phi_{wt} \propto n_{wt}\,; \quad \theta_{td} \propto n_{dt}\,. \tag{3}$$

Известно, что основные алгоритмы обучения моделей PLSA и LDA можно рассматривать как EM-подобные алгоритмы [Asuncion 2009], различающиеся порядком применения формул E-шага (2) и M-шага (3), модификациями M-шага в результате регуляризации, способами распределения частоты n_{dw} по темам. Детали реализации и отличия этих алгоритмов обсуждаются в [Vorontsov 2013].

3. Аддитивная регуляризация тематической модели

Пусть наряду с правдоподобием $L(\Phi,\Theta)$ требуется максимизировать регуляризатор $R(\Phi,\Theta)$ зависящий от параметров модели. Будем максимизировать сумму двух критериев $L(\Phi,\Theta)+R(\Phi,\Theta)$. Решение данной задачи находится EM-алгоритмом с модифицированной формулой M-шага [Vorontsov 2014]:

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+; \quad \theta_{td} \propto \left(n_{dt} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+; \tag{4}$$

где $(x)_+ = \max\{x, 0\}$. К функции $R(\Phi,\Theta)$ предъявляется требование непрерывной дифференцируемости. Она может быть суммой нескольких регуляризаторов, взятых с весами, называемыми *коэффициентами регуляризации*. Таким образом, алгоритм многокритериальной оптимизации ВТМ, независимо от числа критериев, может быть получен из обычных EM-подобных алгоритмов PLSA или LDA простой заменой формулы M-шага.

Для повышения интерпретируемости тем будем опираться на предположение, что хорошо интерпретируемая тема должна иметь ядро, состоящее из терминов предметной области, отличающих её от других тем. Такие темы будем называть *предметными*. Слова общей лексики должны концентрироваться в отдельных *фоновых* темах. Формализуем эти гипотезы с помощью регуляризаторов.

Регуляризатор для разреживания предметных тем основан на предположении, что каждая предметная тема состоит из небольшого числа слов словаря, вероятности остальных слов в распределении ϕ_{wt} равны нулю. Предполагается также, что каждый документ относится к небольшому числу предметных тем, вероятности остальных тем в распределении θ_{td} равны нулю. Вводится регуляризатор, максимизирующий расстояние между распределениями ϕ_{wt} и распределением слов в коллекции β_w, а также между распределением θ_{td} и заданным распределением α_t на множестве предметных тем. Если в качестве расстояния между распределениями взять дивергенцию Кульбака-Лейблера, то формула M-шага (4) примет простой вид:

$$\phi_{wt} \propto (n_{wt} - \beta_0\beta_w)_+; \quad \theta_{td} \propto (n_{dt} - \alpha_0\alpha_t)_+,$$

где β_0, α_0 — коэффициенты регуляризации. Чем они больше, тем больше вероятностей ϕ_{wt} и θ_{td} обращаются в нуль на каждой итерации. Разреживание позволяет достигать 95–99% нулевых значений без значимого ухудшения

правдоподобия модели [Vorontsov 2013]. На ранних итерациях EM-алгоритма коэффициенты регуляризации лучше оставлять равными нулю, затем, по мере сходимости, постепенно увеличивать. Стратегия постепенного разреживания позволяет избежать преждевременного обнуления вероятностей.

Регуляризатор для сглаживания фоновых тем формализует требование, чтобы предметные темы не содержали слов общей лексики. Для описания этих слов в модель вводятся *фоновые* темы, распределения которых должны быть похожи на распределение слов во всей коллекции β_w. Регуляризатор минимизирует дивергенции Кульбака-Лейблера между распределениями ϕ_{wt} фоновых тем и распределением β_w. Кроме того, фоновые темы должны присутствовать в каждом документе. Поэтому вводится второй регуляризатор, минимизирующий дивергенции Кульбака-Лейблера между θ_{td} и α_t для фоновых тем t. В результате формула М-шага (4) даёт оценки параметров, аналогичные модели LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0 \beta_w; \quad \theta_{td} \propto n_{dt} + \alpha_0 \alpha_t,$$

где β_0, α_0 — коэффициенты регуляризации. Эффектом данного регуляризатора является сглаживание (увеличение) малых значений параметров ϕ_{wt} и θ_{td} для фоновых тем за счёт незначительного уменьшения больших значений. Фоновые темы собирают слова общей лексики, стоп-слова и редкие слова, которые исключаются из предметных тем в результате разреживания. Сглаживание фоновых тем является обобщением робастных моделей [Potapenko 2012], в которых фактически использовалась только одна фоновая тема.

Отметим, что сглаживание и разреживание описываются общей формулой и отличаются только знаком параметров β_w, α_t. Это позволяет одновременно сглаживать фоновые темы и разреживать предметные. В байесовском подходе такая возможность до сих пор оставалась незамеченной из-за ограничения неотрицательности параметров распределения Дирихле. Модель LDA описывает только сглаживание; для построения разреженных моделей до сих пор приходилось использовать довольно сложные вероятностные конструкции.

Регуляризатор для декоррелирования тем. Ещё одно требование к предметным темам состоит в том, чтобы они как можно сильнее различались. Данное требование формализуется регуляризатором, который минимизирует сумму ковариаций между распределениями ϕ_{wt} и ϕ_{ws} для всех пар тем t, s [Tan 2010]. Формула регуляризованного М-шага (4) в этом случае принимает вид

$$\phi_{wt} \propto (n_{wt} - \tau \phi_{wt}(\phi_w - \phi_{wt}))_+; \quad \phi_w = \textstyle\sum_t \phi_{wt},$$

где τ — коэффициент регуляризации. Декоррелирование приводит к разреживанию тем и к более чёткому выделению ядер тем, состоящих из слов w с сильно доминирующей вероятностью $p(t|w)$. Декоррелирование, как и разреживание, хорошо сочетается со сглаживанием фоновых тем.

Регуляризатор для сокращения незначимых тем формализует требование, чтобы в модели не было тем, к которым относится слишком мало слов. Такие темы имеют маломощное ядро из редких слов. Чтобы исключать эти темы

из модели, вводится требование разреженности распределения тем во всей коллекции $p(t) = \sum_d \theta_{td} p(d)$. Регуляризатор максимизирует дивергенцию Кульбака-Лейблера между $p(t)$ и равномерным распределением. Формула регуляризованного М-шага (4) в этом случае принимает вид

$$\theta_{td} \propto \left(n_{dt} - \tau\theta_{td}\frac{n_d}{n_t}\right)_+.$$

где τ — коэффициент регуляризации. Согласно этой формуле, если число слов n_t, отнесённых к теме t во всей коллекции, мало, то вероятности этой темы понижаются для всех документов, вплоть до обнуления t-й строки матрицы Θ. Данный регуляризатор позволяет оптимизировать число тем, если начинать итерации с заведомо избыточного числа тем.

4. Оценки качества и интерпретируемости модели

Многокритериальная оптимизация требует также и многокритериального подхода к оцениванию качества ВТМ. При комбинировании регуляризаторов предлагается изменять коэффициенты регуляризации в ходе итерационного процесса и следить за изменениями различных показателей качества модели.

Перплексия является общепринятой мерой качества ВТМ. Она показывает, насколько хорошо модель (1) приближает наблюдаемые частоты появления слов в документах. Качество модели тем выше, чем меньше перплексия. Перплексия измеряется по контрольной выборке документов, не используемых для построения модели. Это позволяет избежать занижения оценки в результате переобучения.

Разреженность модели — доля нулевых значений среди параметров ϕ_{wt} и θ_{td}, только для предметных тем.

Число тем может уменьшаться при обнулении строк матрицы Θ в результате действия регуляризаторов разреживания или сокращения незначимых тем.

Доля фоновых слов $\frac{1}{n}\sum_{d,w} n_{dwt}$, где n — длина коллекции в словах. Принимает значения от 0 до 1. Значения, близкие к 1, свидетельствуют о вырождении тематической модели, например, в результате чрезмерного разреживания.

Когерентность темы определяется как средняя совместная встречаемость двух слов по всем парам k наиболее вероятных слов темы [Newman 2010, Mimno 2011]. Совместная встречаемость оценивается как *поточечная взаимная информация (PMI)* по документам, в которых встречаются оба слова. Число k в большинстве работ полагают равным 10. Для получения более глубоких оценок мы также вычисляем ещё две оценки когерентности: при $k = 100$ и по ядрам тем.

В данной работе предлагаются новые меры интерпретируемости тематической модели, не требующие асессорских оценок, как и когерентность. Будем относить слово w к *ядру* темы t, если $p(t|w) > \delta$. В наших экспериментах $\delta = 0.25$. Обозначим ядро темы t через W_t и определим три показателя интерпретируемости темы.

Размер ядра $|W_t|$ должен быть не слишком мал, но и не слишком велик. Ядра, содержащие ориентировочно от 20 до 200 слов, представляются адекватными.

Чистота темы $\sum_{w \in W_t} p(w|t)$ определяется как суммарная вероятность слов ядра. Принимает значения от 0 до 1; чем выше, тем лучше.

Контрастность темы $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$ равна средней вероятности встретить слова ядра именно в данной теме. Принимает значения от 0 до 1; чем выше, тем лучше. Показывает, насколько хорошо ядро темы отличает её от остальных тем.

Интерпретируемость модели тем лучше, чем выше когерентность, чистота и контрастность всех её тем. Поэтому мы определяем соответствующие показатели качества всей модели путём их усреднения по всем предметным темам.

5. Эксперименты

Исходные данные. Эксперименты проводились на коллекции NIPS, содержащей 1700 текстов статей научной конференции Neural Information Processing Systems на английском языке. Суммарная длина коллекции $2.3 \cdot 10^6$, объём словаря $1.3 \cdot 10^4$. Предварительная обработка текстов включала приведение к нижнему регистру, удаление пунктуации, удаление стоп-слов с помощью библиотеки BOW toolkit [McCallum 1996]. Во всех экспериментах общее число тем равно 100, для сглаженно-разреженных моделей среди них выделяется 90 предметных и 10 фоновых тем.

На рисунках 1 и 2 приведены зависимости показателей качества модели от номера итерации. На каждом рисунке сравниваются результаты модели PLSA и регуляризованной модели. Показатели качества выведены на четырёх графиках друг под другом с одинаковой горизонтальной осью. Верхний график: по левой оси перплексия, по правой — разреженности матриц параметров Φ, Θ. Второй график: по левой оси число тем, по правой — доля фоновых слов. Третий график: по левой оси размер ядра, по правой — контрастность и чистота. Нижний график: по левой оси когерентность ядра, по правой когерентности top-10 и top-100.

Такие графики предлагается использовать на этапе построения тематической модели для мониторинга показателей качества модели в ходе итерационного процесса. В частности, эти графики дают понимание, какие эффекты производит каждый регуляризатор в отдельности, как они взаимодействуют в комбинации, как выбирать стратегию изменения коэффициентов регуляризации. Не имея возможности привести здесь результаты всех протестированных комбинаций регуляризаторов, перечислим только основные выводы.

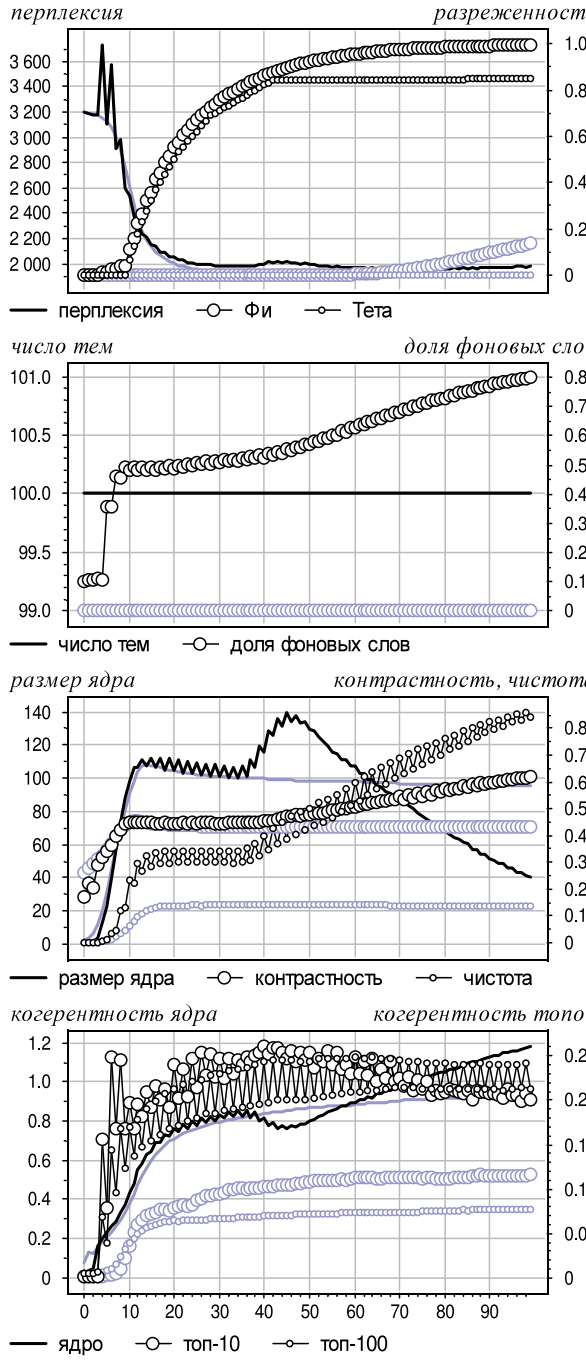

Рис. 1. Сравнение PLSA (серый) с комбинацией разреживания, сглаживания и декоррелирования (чёрный)

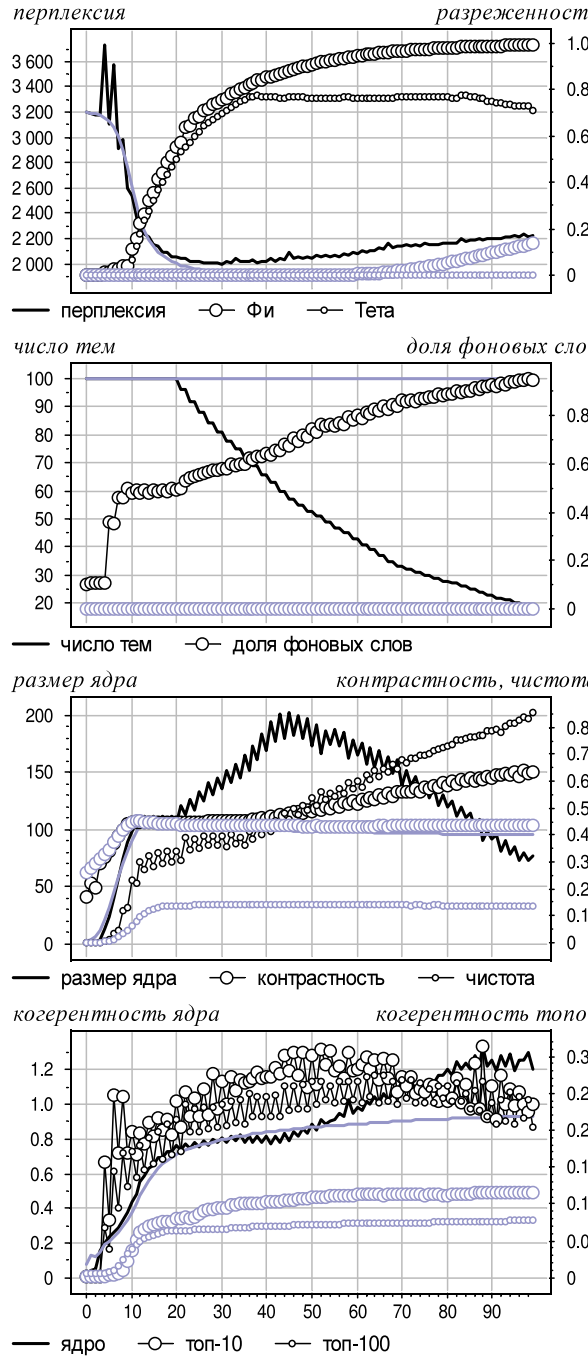

Рис. 2. Сравнение PLSA (серый) с комбинацией разреживания, сглаживания, декоррелирования и удаления тем (чёрный)

1. Разреживание предметных тем возможно до 98 % в матрице и до 90 % в матрице Θ практически без потери перплексии. Разреживание матрицы по равномерному распределению β_w приводит к удалению редких слов и улучшению контрастности. Если же в качестве β_w брать распределение слов в коллекции, то разреживание улучшает когерентность и чистоту. Коэффициенты регуляризации для разреживания рекомендуется плавно увеличивать после 10–15 итераций, когда итерационный процесс уже почти сошёлся, или хотя бы появилась определённость в том, какие именно элементы в матрицах Φ, Θ являются наименьшими.

2. Декоррелирование в несколько раз увеличивает чистоту и когерентность тем, но слабо разреживает матрицу Φ и вообще не разреживает матрицу Θ. Комбинация декоррелирования с разреживанием позволяет достичь сильной разреженности без уменьшения чистоты и когерентности. Декоррелирование рекомендуется включать с первой итерации, с максимально возможным коэффициентом регуляризации.

3. Сглаживание фоновых тем способствует переходу слов общей лексики из предметных тем в фоновые. Для этого достаточно одной фоновой темы. Сглаживание лучше включать с первой итерации, с фиксированным коэффициентом регуляризации. Комбинирование сглаживания фоновых тем с разреживанием и декорреляцией предметных тем достигает наилучших результатов по всей совокупности показателей (рис. 2).

4. Сокращение незначимых тем разреживает строки матрицы целиком, определяя минимальное необходимое число тем. Этот регуляризатор, так же, как и разреживающий, лучше включать постепенно, на фоне устойчивой сходимости процесса. Возрастание перплексии, уменьшение размера ядер или приближение доли фоновых слов к 1 могут свидетельствовать о вырождении тематической модели и нецелесообразности дальнейшего сокращения числа тем. В таком случае коэффициент регуляризации необходимо снова положить равным нулю. В наших экспериментах уменьшение числа тем ниже 60 ведёт к вырождению (рис. 2).

6. Выводы

Данная работа иллюстрирует применение нового подхода в тематическом моделировании, *аддитивной регуляризации тематических моделей* (ARTM), для построения сильно разреженной модели с интерпретируемыми темами. Предложены регуляризаторы разреживания предметных тем, сглаживания фоновых тем, декоррелирования и сокращения числа тем. Показано, что их комбинирование улучшает совокупность критериев качества практически без ухудшения перплексии модели. Предложена методика визуального мониторинга качества модели и подбора коэффициентов регуляризации. Поиск оптимальных *траекторий регуляризации* в пространстве коэффициентов регуляризации пока остаётся открытой проблемой.

Для оценивания интерпретируемости, наряду с когерентностью, предложены критерии чистоты и контрастности тем. Они основаны на гипотезе о том, что в интерпретируемой теме должно хорошо выделяться ядро — множество слов, отличающих данную тему от остальных.

Литература

1. *Asuncion A., Welling M., Smyth P., The Y. W.* (2009), On smoothing and inference for topic models, Proceedings of the International Conference on Uncertainty in Artificial Intelligence.
2. *Blei D. M., Ng A. Y., Jordan M. I.* (2003), Latent Dirichlet allocation, Journal of Machine Learning Research, Vol. 3, pp. 993–1022.
3. *Blei D. M.* (2012), Probabilistic topic models, Communications of the ACM, Vol. 55, No 4, Pp. 77–84.
4. *Chang J., Boyd-Graber J., Gerrish S., Wang C., Blei D. M.* (2009), Reading Tea Leaves: How Humans Interpret Topic Models, Advances in Neural Information Processing Systems, pp. 288–296.
5. *Daud A., Li J., Zhou L., Muhammad F.* (2010), Knowledge discovery through directed probabilistic topic models: a survey, Frontiers of Computer Science in China, Vol. 4, no. 2, pp. 280–301.
6. *Hofmann T.* (1999), Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA: ACM, pp. 50–57.
7. *Khalifa O., Corne D., Chantler M., Halley F.* (2013), Multi-objective topic modelling, Proceedings of 7th International Conference Evolutionary Multi-Criterion Optimization (EMO 2013), Springer LNCS, pp. 51–65.
8. *McCallum A. K.* (1996), Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, http://www.cs.cmu.edu/_mccallum/bow
9. *Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A.* (2011), Optimizing semantic coherence in topic models, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP ’11, pp 262–272.
10. *Newman D., Karimi S., Cavedon L.* (2009), External evaluation of topic models, Australasian Document Computing Symposium, December 2009, Pp. 11–18.
11. *Newman D., Lau J. H., Grieser K., Baldwin T.* (2010), Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, HLT ’10, pp. 100–108.
12. *Potapenko A. A., Vorontsov K. V.* (2013), Robust PLSA performs better than LDA, 35th European Conference on Information Retrieval, Moscow, Russia, 24–27 March 2013, Lecture Notes in Computer Science, Springer Verlag-Germany, pp. 784–787.
13. *Tan Y., Ou Z.* (2010), Topic-weak-correlated latent Dirichlet allocation. 7th International Symposium Chinese Spoken Language Processing (ISCSLP), pp. 224–228.
14. *Vorontsov K. V., Potapenko A. A.* (2013), EM-like algorithms for probabilistic topic modeling [Modifikacii EM-algoritma dlja verojatnostnogo tematicheskogo modelirovanija], Machine Learning and Data Analysis [Mashinnoe obuchenie i analiz dannyh], Vol. 1, no. 6, pp. 657–686.
15. *Vorontsov K. V.* (2014), Additive Regularization for Topic Models of Text Collections, Doklady Akademii Nauk, Vol. 455, no. 3.