Содержание

1 Введение

В тематическом моделировании рассматриваются текстовые коллекции: документы, каждый из которых состоит из множества слов-термов. При постановке задач тематического моделирования предполагается, что термы неупорядочены - изменение порядка термов не влияет на свойства коллекции. Одной из задач тематического моделирования является описание тем: предполагается, что каждый документ соответствует некоторому множеству тем, а каждой теме - некоторое множество слов. Каждый документ и какждое слово соотносятся с каждой темой с некоторой, возможно нулевой, вероятностью. Таким образом, задача тематического моделирования состоит в поиске двух дискретных вероятностных распределений.

Для нахождения двух распределений путем тематического моделирования составляется модель, решающая задачу приближенного матричного разложения. Задача приближенного матричного разложения ставится некорректно: она может иметь бесконечное множество решений для одной коллекции документов. Можно использовать регуляризацию для выделения из множества решений одного, удовлетворяющего требуемым свойствам. Одним из таких свойств решений, которые учитываются при построении тематической модели, является интерпретируемость тем: по словам,с наибольшей вероятностью соответствующим теме, возможно определить смысл темы.

Решение задачи тематического моделирования основано на использовании EM-алгоритма для максимизации правдоподобия. При таком подходе темы становятся равными по мощности - числа документов, соответствующих каждой из тем, получаются близки. Если мощности тем исходных коллекций равны или отличаются не более чем в 2-3 раза, использование EM-алгоритма решит поставленную задачу. В противном случае, когда в коллекции существуют темы превосходящие по мощности другие в десятки и более раз, при решении оптимизационной задачи малые темы объединятся в более крупные или сольются с уже существующими, а крупные разделятся, образовав множество схожих с исходной мелких тем.

Текстовая коллекция может быть произвольной. При использовании коллекции для подстановки в модель процесс генерации документов и баланс тем неизвестен - по используемой коллекции невозможно определить наличие темы значительно превышающей по мощности другие.

Если такая тема существует, то процессы слияния крупных тем и разделения мелких будут приводить к ухудшению интерпретируемости тем.

Для задачи тематического моделирования существуют различные методы решения. Одним из основных подходов является LDA, использующий логарифм априорного распределения Дирихле в качестве регуляризатора. Этот подход был обобщен для использования произвольных регуляризаторов в ARTM. В настоящее время ARTM является наиболее общим подходом, так как использование регуляризаторов позволяет адаптировать решение под любую конкретную задачу. Однако существование различных по мощности тем возможно для каждой задачи - это свойство обрабатываемой коллекции. Поэтому используемый набор регуляризаторов должен быть применим как к сбалансированной коллекции с равными по мощности темами, так и в несбалансированной, с темами, отличающимися по мощности в десятки и более раз.

В данной работе экспериментально показано, что проблема несбалансированности может наблюдаться на реальных коллекциях, рассмотрен регуляризатор семантической неоднородности, добавленный для её решения и проведены эксперименты, демонстрирующие интепретируемость получаемых при помощи добавленного регуляризатора тем.

2 Постановка задачи

2.1 Общая постановка задачи тематического моделирования

Пусть D - множество документов, W - множество термов. Каждый документ из D задается его длиной n_d : $\sum_{d \in D} n_d = n$ и упорядоченной последовательностью термов $\{w_i \in W\}_{i=1}^{n_d}$. Элементы этой последовательнсоти будем называть словами. Вводится вероятностная модель порождения коллекции, описывающая вероятности появляения слов в документе путем введения конечного множества тем T. Для этого ставятся следующие предположения:

- Гипотеза мешка слов: для каждого документа d представление слов в виду последовательности $\{w_i \in W\}_{i=1}^{n_d}$ эквивалентно представлению документа в виде неупорядченного множества $\cup_{i=1}^{n_d} w_i$, в котором каждое слово w встречается n_{dw} раз.

- Гипотеза условной независимости: вероятность появления слова w в документе d по теме t описывается распределением

$$p(w|d,t) = p(w|t)$$

Последнее предположение описывает независимость вероятности появления слов от документа - вероятность описывается вероятностями появления слов для каждой темы $p(w|t) = \phi_{wt}$ и вероятностями порождения слов документов из каждой темы $p(t|d) = \theta_{td}$. Задача тематического моделирования состоит в определении каждой из описанных вероятностей. Это эквивалентно:

$$F = \Phi\Theta \tag{1}$$

$$F = \left(\frac{n_{wd}}{n_d}\right)_{W\times D} \quad \Phi = (\phi_{wt})_{W\times T} \; \Theta = (\theta_{td})_{T\times D}$$

Для решения задачи матричного разложения ставится задача максимизации функции правдоподобия, которую решают с помощью EM-алгоритма.

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to max_{\Phi,\Theta} \tag{2}$$

Задача(1) поставлена некорректно: множество ее решений в общем случае бесконечно. Чтобы выделить одно решение, в оптимизируемую функцию (2) добавляют несколько слагаемых - регуляризаторов, зависящих от матриц Φ, Θ. Вид регуляризаторов зависит от задачи - различные регуляризаторы наделяют решение различными свойствами. Функция правдоподобия при этом принимает следующий вид:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi,\Theta) \to max_{\Phi,\Theta} \tag{3}$$

2.2 Проблема несбалансированности

Описанная модель дает на выходе две матрицы: матрицу Φ, показывающую вероятность появление конкретного слова при заданной теме и матрицу Θ, показывающую распределение тем между документами. Из матрицы Θ часто следует, что темы равны по мощности, а именно: при вероятностях тем, определенных как $p(t) = \sum_{d\in D} p(t|d)n_d$ выполнено $\forall t_1, t_2 \to \frac{p(t_1)}{p(t_2)} \approx 1$. При этом в обработанной коллекции мощности тем

могли быть произвольными. Такой эффект возникает из-за алгоритма решения задачи: модели выгодно использовать все свои параметры. Сокращение доли отдельной темы приводит к неполному использованию или, в пределе, уменьшению числа параметров. Чтобы выделять не равные по мощности темы, в модель предлагается добавить регуляризатор. Для построения тематической модели была использована гипотеза условной независимости, эквивалентная формулировка которой записывается как $p(w, s|t) = p(w|t)p(d|t)$. В данном виде для темы t гипотеза условной независимости проверяется статистикой семантической неоднородности темы

$$S_t = KL(\hat{p}(w, d|t)||p(w|t)p(d|t))$$

Здесь и далее $\hat{p}$ обозначает частотные оценки вероятностей

$$\hat{p}(w, d|t) = \frac{n_{dw}p(t|d, w)}{n \cdot p(t)},\ \hat{p}(w|d) = \frac{n_{dw}}{n_d}$$

Статистику семантической неоднородности также можно записать в виде

$$S_t = \sum_{d \in D} \sum_{w \in d} \hat{p}(w, d|t) \ln \frac{\hat{p}(w, d|t)}{p(w|t)p(d|t)} \tag{4}$$

Если гипотеза условной независимости верна, значение статистики семантической неоднородности должно быть мало, а распределения $p(w|d, t)$ близки к $p(w|t)$. Можно представить каждую тему как кластер размерности $|W|$, центром которого является $p(w|t)$. Статистика семантической неоднородности показывает удаленность $p(w|d, t)$ от центра кластера. Чтобы учитывать статистику семантической неоднородности, в модель можно добавить регуляризатор, учитывающий значение статистики. Будем суммировать статистику семантической неоднородности темы по всем темам:

$$\sum_{t \in T} S_t = \sum_{d \in D} \sum_{w \in d} \left(\sum_{t \in T} \hat{p}(w, d|t) \right) \ln \frac{\hat{p}(w, d|t)}{p(w|t)p(d|t)} \to \min_{\Phi, \Theta} \tag{5}$$

Воспользуемся формулой $\hat{p}(w, d|t) = \frac{p(t|d,w)\hat{p}(w|d)p(d)}{p(t)}$ для преобразования логарифма и домножим на постоянное для конкретной задачи n. Таким образом предыдущая формула преобразуется и вставляется в исходную постановку задачи в качестве регуляризатора R:

$$R = \sum_{d \in D} \sum_{w \in d} \beta_{dw} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \tag{6}$$

$$\beta_{dw} = \sum_{t \in T} \frac{p(t|d,w)}{p(t)}$$

Заметим, что выражение (6) это в точности выражение (**??**), домноженное на весовые множители β_{dw}. По смыслу эти множители увеличивают вес маломощных тем $p(t) << 1$

Цель данной работы - исследовать влияние регуляризатора на обработку тематической моделью реальных текстовых коллекций в зависимости от баланса тем и параметров модели.

3 Вычислительный эксперимент

3.1 Генерация реальной коллекции

Эксперименты прводились на коллекциях, полученных путем предобработки одной из коллекций: датасета 20newsgroups(20news), коллекции записей ted talks(TED), коллекции новостей BBC(BBC) и коллекции рецензий на книги(BOOK). Основные параметры этих коллекций приведены в таблице:

	20news	TED	BBC	BOOK
Число документов				
Число слов, среднее				
Число слов, медиана				
Число уникальных слов				

Для генерации несбалансированной коллекции использовалась комбинация одного из следующих подходов:

1. Берется подвыборка документов исходной коллекции. Исходные коллекции были сбалансированными: степень несбалансированности не превосходит 5. Для таких коллекций тематическая модель без регуляризаторов способна корректно соотнести большинство документов с темами так, чтобы малые темы не объединялись, а крупные не разбивались на несколько более мелких. Поэтому для получения несбалансированной коллекции из исходных удаляется часть документов, относящаяся к более мелким темам.

2. Отбираются только монотематические документы. Для этого на исходной коллекции строится тематическая модель без регуляризаторов. Для каждого документа рассматриваются две темы, порождающие его с наибольшими вероятностями. Если отношение вероятностей рассматриваемых тем больше 2, документ считается монотематическим.

3. Генерируются новые документы. Для этого выбирается число слов генерируемого документа и его тема. Пока в документе не сгенерировано достаточное количество слов, выбира.тся случайные n_{group} подряд идущих слов из случайного документа выбранной темы. Здесь n_{group} - параметр алгоритма. Эксперименты проводились при $n_{group} \leq 10$

3.2 Оценивание качества восстановления тем

Пусть для эксперимента используются коллекции с заранее известными темами документов. При генерации новых документов на основе уже имеющихся тема нового документа совпадает с темой исходных. Введем известное распределение тем Θ: матрицу размера $T \times D$ в столбце i и строке j которой стоит 1, если документ d_j принадлежит теме i и 0 в противном случае. Для того, чтобы такой способ примерно соответствовал действительности при генерации коллекции использовался второй подход. Таким образом,чтобы оценить качество восстановления тем достаточно сравнить известное распредление тем по документам с матрицей Θ_0, получаемой из модели. Чтобы оценить сходство полученных матриц Θ_0 с известными Θ в данной работе считается количество взаимно ближайших по некоторой метрике столбцов матриц Θ, Θ_0, то есть пар строк $\Theta[i]$,$\Theta_0[j]$:

$$\arg\min_k \ (dist(\Theta[i], \Theta_0[k])) = j \tag{7}$$

$$\arg\min_k \ (dist(\Theta[k], \Theta_0[j])) = i \tag{8}$$

Здесь $dist$ - расстояние по заданной метрике. В экспериментах будет использоваться косинусное расстояние и расстояние Йенсена-Шеннона.

Будем полагать, что если вышеописанная оценка сходства низкая, модель не смогла описать порождающие документы темы и наоборот, если оценка сходства высока, то модель смогла описать порождение коллекции. Тогда идеальный результат достигается, когда каждой исходной

теме сопоставляется ровно одна из тем на выходе модели. На практике такое достигается крайне редко и существуют темы $\Theta_0[j]$, которые не являются ближайшими ни для какой темы из Θ, то есть для любого i для пары $\Theta[i], \Theta_0[j]$ не выполнено (7). Такие темы будем называть невосстановленными. Аналогично определим ложные темы: такие темы $\Theta_0[i]$, что для любого j (8) не выполнено для пары $\Theta[i], \Theta_0[j]$.

3.3 Демонстрация проблемы несбалансированности

Ниже приведена демонстрация влияния баланса тем в реальной коллекции на качество восстановления тем на примере 20news. Был использован первый способ генерации несбалансированной коллекции. Были построены коллекции с различной степенью несбалансированности. Было рассмотрено значение описанных ранее метрик в зависимости от степени несбалансированности. Результат приведен на графике

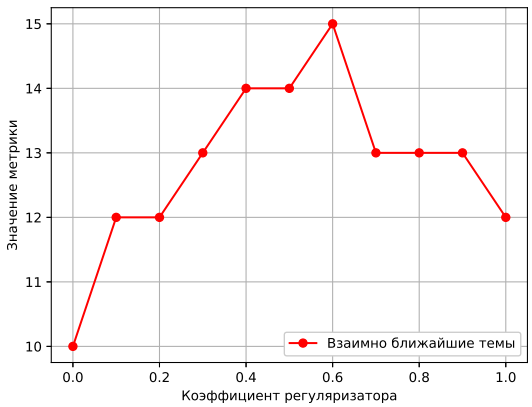

Рис. 1: Эксперимент на несбалансированной коллекции, полученной подвыборкой из 20newsgroups. Зависимость между значением метрики и коэффициентом регуляризации

Как можно увидеть из графика, при отсутствии регуляризации модель смогла выделить только 10 взаимно близких с исходными тем. В то же время в зависимости от коэффициента регуляризации модель с регуляризатором выделяла больше взаимно ближайших тем при максимуме

для коэффициента 0.6. Заметим, что оптимальное значение коэффициента регуляризации в данном эксперименте оказалось таким же, как и в экспериментах на синтетических коллекциях.

3.4 Подбор параметра генерации коллекции

Проверим, что параметр генерации коллекции не влияет на основные свойства решения задачи тематического моделирования. Для этого для коллекции 20news построим десять коллекций с высокой степенью несбалансированности ($\tau = 50$). Каждая из этих коллекций строилась третьим способом, при этом для генерации коллекции с номером i использовался параметр $n_{group} = i$. Для всех построенных коллекций проводился эксперимент с добавлением регуляризатора семантической неоднородности и без него. Коэффициент регуляризации выбирался как оптимальный из предыдущего пункта. Результат приведен на графике

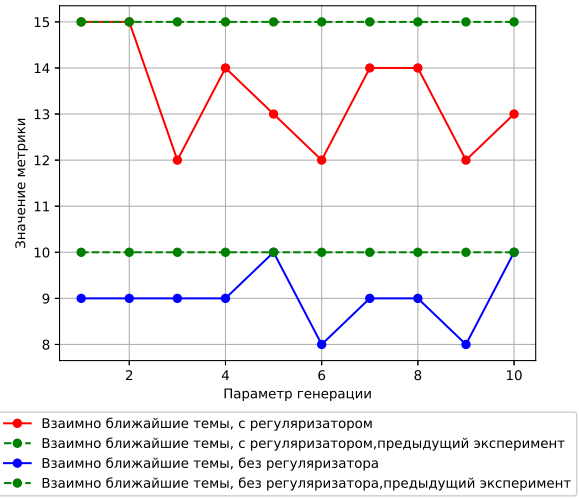

Рис. 2: Эксперимент по подбору параметров генерации коллекции

На графике и красная и синяя линии лежат под соответствующими зелеными, но разница незначительна. Это означает, что такой способ генерации можно использовать для построения коллекций. При этом, поскольку на графике отсутствует явная зависимость значения метрики от параметра его можно выбирать любым от 1 до 10. Верхняя граница

определена пределами эксперимента. Однако красная и синяя линии расположены ниже соответствующих зеленых, что означает, пусть и незначительную, потерю некой изначальной структуры коллекции.

3.5 Добавление разреживания матриц

Из приведенных параметров для описания коллекций можно оценить, что матрицы счетчиков n_{dw} могут оказаться разреженными как при обработке исходной коллекции, так и при обработке сгенерированных документов. Чтобы улучшить работу модели для таких данных в модель была добавлена стратегия разреживания на основе работы Глушаченкова. Было проведено два эксперимента. Для первого проводилось сравнение результатов работы модели с использованием стратегии и без для различных коэффициентов регуляризатора семантической неоднородности. Для этого использовалась коллекция 20newsgroups. Результаты приведены на графике

Из графика можно увидеть, что добавление указанной стратегии не ухудшило, но и не улучшило результаты модели. Второй эксперимент проверял, что результаты оказались устойчивыми в зависимости от начальной инициализации. Эксперимент проводился на коллекции новостей BBC. Строились тематические модели с добавлением стратегии разреживания и без нее. Модели запускались несколько раз с различными перестановками строк и столбцов матрицы n_{dw} Проверялось среднее отклонение результатов одной модели друг от друга в зависимости от количества тем модели. результаты приведены на графике.

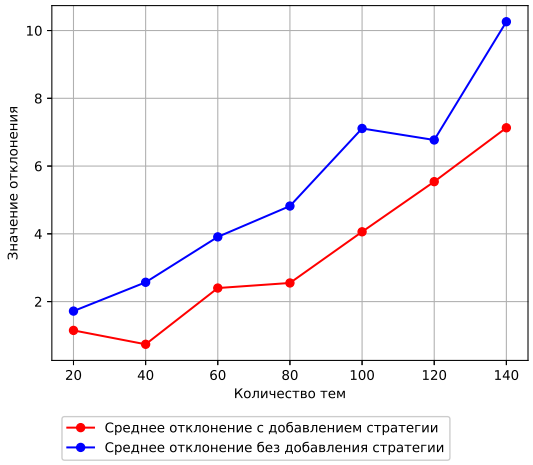

Рис. 3: Измерение отклонения между запусками модели в зависимости от числа тем

Заметим, что модель с добавлением стратегии оказалась более устойчивой. При этом отклонение растет в зависимости от числа тем. Следующий эксперимент демонстрирует, что это связано со свойствами используемой коллекции. Для всех используемых коллекций построим тематическую модель для каждого числа тем из $[20, 40, 60, 80, 100]$. Добавим в модели регуляризатор семантической неоднородности с параметром 0.6 и рассмотрим среднее число взаимно близких тем при последовательных запусках. Результаты приведены на графике.

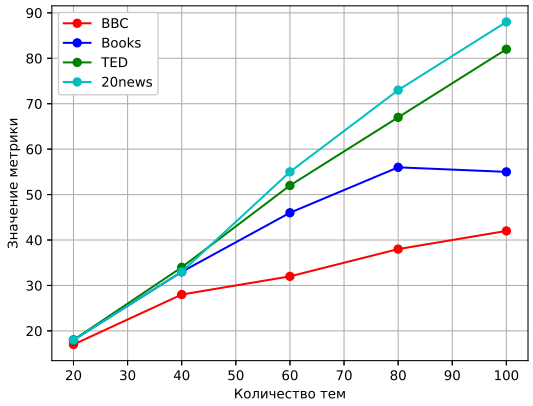

Рис. 4: Среднее число взаимно близких тем между повторными запусками одной модели на одном датасете

Для коллекции BBC модель неспособна выделить больше чем 40 тем. В то же время, из предыдущего графика видно, что при выделении числа тем меньшего, чем 40, отклонение результатов при использовании стратегии разреживания получается достаточно низким. В случае когда число тем заведомо превышает реальное и модель не может корректно восстановить темы, отклонение результатов моделей с добавленим стратегии разреживания ниже чем без неё. Таким образом, стратегия разреживания добавлена во все последующие эксперименты.

3.6 Добавление регуляризаторов

Проверим, как изменяется качество восстановления тем при добавлении регуляризатора разреживания. Для этого сгенерируем несколько коллекций со степенью несбалансированности 50 на основе 20news третьим подходом. Исследовалась зависимость метрик от коэффициента регуляризатора разреживания при постоянном коэффициенте регуляризатора семантической неоднородности. Результаты обработки различных коллекций усреднялись и приведены на графике ниже.

Аналогичный эксперимент был повторен для регуляризатора декоррелирования.Результаты представлены на графике ниже

Как видно из графиков, добавление регуляризатора декоррелирования при подборе коэффициента регуляризации повышает качество восстановления тем. В то же время добавление регуляризатора разрежива-

ния не дает улучшений качества и поэтому в дальнейшем не добавляется в модель.

3.7 Добавление итеративной балансировки тем

Проверим, как изменится качество восстановления тем если добавить в модель итеративную балансировку тем. Алгоритм балансировки возьмем из работы Веселовой. Он заключается в добавлении регуляризатора, поощряющего несбалансированность тем путем приближения его к априорному балансу по формуле

$$R_{TopicPrior} = \sum_t \sum_w \beta_t \log \phi_{wt}$$

Поскольку β_t изначально неизвестны, вектор β генерируется из симметричного распределения Дирихле с заданным параметром. На графике ниже представлена зависимость качества восстановления тем от параметра. Видно, что ни при каком значении параметра модель не улучшила результат

4 Заключение

- На примере коллекции 20newsgroups и сгенерированных на её основе коллекций показано, что тематическая несбалансированность коллекции приводит к дроблению крупных и слиянию мелких тем
- Предложен алгоритм устранения проблемы несбалансированности путем добавления регуляризации на основе семантической однородности тем
- Проведены эксперименты, демонстрирующие возможные модификации модели и показывающие, что модель утсраняет проблему несбалансированности