Содержание

1 Введение

Тематическое моделирование - одно из направлений обработки текстовых коллекций. В тематическом моделировании рассматриваются коллекции, состоящие из множества документов. Каждому документ состоит из некоторого множество слов, которое обычно предполагается неупорядоченным. Задачей 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 нахождении двух семейств дискретных условных вероятностных распределений.

Для нахождения двух распределений и решения задачи тематического моделирования решается задача приближенного матричного разложения. Решение задачи приближенного матричного разложения основано на использовании EM-алгоритма для максимизации правдоподобия. В процессе максимизации темы становятся близкими по мощности - числа документов, соответствующих каждой из тем, получаются близки. Если в коллекции с документами по математике и биологии мощности математической и биологической тем исходных коллекций равны или отличаются не более чем в 2-3 раза, решение задачи для нахождения двух тем выделит математическую и биологическую составляющие. В другом случае, если документов по математике тысяча, а по биологии десять, обе полученные темы окажутся математическими. Это является свойством задачи максимизации правдоподобия. При решении оптимизационной задачи для различных по мощности тем малые темы объединяются в более крупные или сливаются с уже существующими, а крупные разделяются, образовывая множество схожих с исходной мелких тем.

Задача приближенного матричного разложения ставится некорректно: она может иметь бесконечное множество решений для одной коллекции. 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. В случае, когда коллекция состоит из одной части документов по математике и другой по биологии, интерпретируемые темы будут с

наибольшей вероятностью состоять из математических и биологических терминов соответственно.

При использовании произвольной текстовой коллекции и подстановке её в модель процесс порождения документов коллекции неизвестен - по используемой коллекции невозможно определить наличие темы, значительно превышающей по мощности другие. Если такая тема существует, то эта большая по мощности тема будет разделяться, а мелкие сливаться с этой или другой крупной темой. Такой ээффект получил название проблемы несбалансированности. Проблема несбалансированности будет приводить к ухудшению интерпретируемости тем, построенных моделью. Такой темы может и не существовать - при построении модели для несбалансированной коллекции необходимо, чтобы модель могла решить задачу и для равных по мощности тем.

Для задачи тематического моделирования существуют различные методы решения. Одним из основных подходов является LDA, использующий априорное распределение 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 конкретную задачу. Однако существование различных по мощности тем возможно для каждой задачи - это свойство обрабатываемой коллекции. Для решения проблемы несбалансированности с помощью регуляризаторов необходимо, чтобы используемый набор регуляризаторов был применим как к сбалансированной коллекции с равными по мощности темами, так и в несбалансированной, с темами, отличающимися по мощности в десятки и более раз. Иначе полученнное решение не будет применимо к произвольной текстовой коллекции.

В данной работе экспериментально показано, что проблема несбалансированности может наблюдаться на реальных коллекциях, рассмотрен регуляризатор семантической неоднородности, добавленный для её решения и проведены эксперименты на реальных и синтетических коллекциях, 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 $\{w_i \in W\}_{i=1}^{n_d}$. Вводится вероятностная модель порождения коллекции, описывающая вероятности появления термов в документе. В рамках вероятностной модели вводится конечное множество тем T, а вероятности появления слов в документах связывают с условными вероятностями появления тем в документе и вероятностями появления слова для каждой пары темы и документа. При этом предполагается выполнение следующих гипотез:

- Гипотеза мешка слов: для каждого документа d представление слов в виде 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 нахождении распредений ϕ_{wt}, θ_{td}. Для нахождения ϕ_{wt}, θ_{td} ставится задача приближенного матричного разложения:

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

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

Задача состоит в максимизаци следующей функции правдоподобия:

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

Задача приближенного матричного разложения поставлена некорректно: множество ее решений в общем случае бесконечно. Чтобы наложить дополнительное ограничение на множество решений задачи, 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 $\forall t_1, t_2 \to \frac{p(t_1)}{p(t_2)} < C$ для C не превосходящих 10. При этом в обработанной коллекции мощности тем могли быть произвольными и могут существовать такие t_1, t_2, что отношение $\frac{p(t_1)}{p(t_2)}$ больше десяти, сотни или даже тысячи. 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 $p(w, d|t) = p(w|t)p(d|t)$. 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:

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

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:

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

$$\beta_{dw} = \sum_{t \in T} \frac{p(t|d, w)}{p(t)} \tag{7}$$

Заметим, что выражение (6) это в точности выражение (2), домноженное на весовые множители β_{dw}. По смыслу эти множители увеличивают вес маломощных тем $p(t) << 1$

Цель данной работы - предложить решение проблемы несбалансированности, состоящее в добавлении регуляризатора семантической неоднородности, и экспериментально проверить предложенное решение для коллекций с различным балансом тем и для моделей с разными параметрами.

3 Вычислительный эксперимент

3.1 Генерация коллекций

Эксперименты проводились на синтетических и реальных коллекциях. Для получения синтетических коллекций генерируются матрицы Φ, Θ - искомые матрицы Столбцы матриц Φ, Θ порождаются симметричными распределениями Дирихле. Параметр распределения определяется из соображений реалистичности коллекции и берется малым, чтобы получаемые матрицы были разреженными. Для матрицы Φ он берется равным 0.02, для матрицы Θ равным 0.2. Чтобы регулировать баланс тем, на этом этапе генерации наибольшие значения в столбцах Θ меняются со значениями в строках, которые соответствуют необходимым темам. После этого в обе матрицы добавляется еще одна фоновая тема, доля которой во всех документах равна 0.5, если не указано иного, порожденная несимметричным распределением Ципфа. Матрица Θ перед этим перенормируется в зависимости от желаемой доли фоновой темы в документах.

Для генерации очередного слова w_i сначала генерируется тема t_i документа из соответствующего этому документу столбцу матрицы Θ_0. Затем слово генерируется из столбца Φ_0, соответствующего теме t_i. Таким образом, процесс генерации документов описывается как

$$t_i \sim Dir(t|d) \;\; w_i \sim Dir(w|t_i), i \in 1 \dots n_d$$

Реальные коллекции были получены путем предобработки датасета 20newsgroups. Основные параметры приведены в таблице:

	20news
Число документов	18846
Число слов, среднее	
Число слов, медиана	
Число уникальных слов	42616

Для генерации несбалансированной коллекции использовалась комбинация одного из следующих подходов:

1. Берется подвыборка документов исходной коллекции. Для исходной коллекции отношения чисел документов, соответствующих двум

темам могло не превосходить 10. 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 исходной удаляются документы, соответствующие более мелким темам. Число удаленных документов зависит от требуемого баланса тем в коллекции.

2. Отбираются 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, выбираются случайные n_{group} подряд идущих слов из случайного документа выбранной темы. Здесь n_{group} - параметр алгоритма. Эксперименты проводились при $n_{group} \leq 10$

Определим понятие степени несбалансированности. Для синтетической коллекции известна вероятность появления темы в документе для каждой пары темы и документа. Это следует из процесса генерации коллекции на основе матрицы Θ - искомой вероятностью будет элемент соответствующей строки и соответствующего столбца матрицы. Будем генерировать документы с равным числом слов. Тогда мощностью темы $p(t)$ назовем сумму вероятностей её появления по документам - сумму соответствующего столбца Θ.

$$p(t) = \sum_d p(t|d)p(d) \sim \sum_d p(t|d)$$

Степень несбалансированности в таком случае определим как

$$\frac{\max\limits_{t \in T}|t|}{\min\limits_{t \in T}|t|} \tag{8}$$

Для реальной коллекции будем отбирать монотематические документы. В таком случае составим матрицу 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. Тогда определение степени несбалансированности через равенство (8) применимо и к реальным или сгенерированным на их основе коллекциям.

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. Таким образом, как для реальных так и для синтетическиз коллекций можно получить искомое распределение тем Θ. Для реальных коллекций его получение было описано в предыдущей секции. Чтобы оценить качество восстановления тем достаточно сравнить известное распределение 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 строк матриц Θ, Θ_0, то есть пар строк $\Theta[i]$,$\Theta_0[j]$:

$$\arg\min_k \ (dist(\Theta[:,i], \Theta_0[:,k])) = j \tag{9}$$

$$\arg\min_k \ (dist(\Theta[:,k], \Theta_0[:,j])) = i \tag{10}$$

Здесь $dist$ - расстояние по заданной метрике. В экспериментах используется косинусное расстояние. Заметим, что свойство взаимной близости можно ввести и для столбцов матриц Φ, Φ_0

$$\arg\min_k \ (dist(\Phi[i], \Phi_0[k])) = j \tag{11}$$

$$\arg\min_k \ (dist(\Phi[k], \Phi_0[j])) = i \tag{12}$$

Все последующие метрики также могут применяться как для строк Φ, Φ_0, так и для столбцов $\Theta[:,i], \Theta_0[:,k]$. 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 (9). Такие темы будем называть невосстановленными. Аналогично определяются ложные темы: такие темы $\Theta[i]$, что для любого j (10) не выполнено для пары $\Theta[i], \Theta_0[j]$.

3.3 Демонстрация проблемы несбалансированности

На примере синтетических коллекций проверим, что при возрастании степени несбалансированности число взаимно ближайших тем уменьшается. Для этого сгенерируем коллекции с одной большой по мощности темой и множеством малых равномощных. Общее число документов в коллекции равно 2000 для каждой сгенерированной коллекции, число тем равно 100. Результат обработки такой коллекции тематической моделью представлен на графике

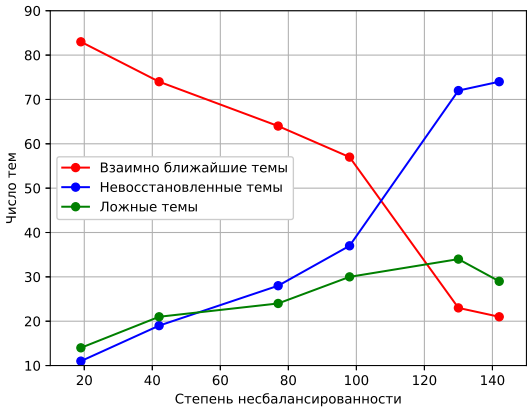

Рис. 1: Зависимость числа восстановленных взаимно близких тем от степени несбалансированности коллекции, синтетические коллекции

При малых степенях несбалансированности число взаимно близких тем модели близко к 90. Однако оно уменьшается с возрастанием степени несбалансированности и становится очень мало для степеней несбалансированности больше 100.

Проверим, что проблема несбалансированности наблюдается и для реальных коллекций. Проведем аналогичный эксперимент для коллекций, сгенерированной на основе 20newsgroups. Генерировались коллекции с одной темой, состоящей из 3000 документов и девятнадцати темами, состоящими из $\frac{3000}{x}$ документов, где x - зависящая от эксперимента степень несбалансированности. Рассматривались результаты для модели без регуляризатора семантической неоднородности и для модели с добавлением регуляризатора с коэффициентом 0.6. Результаты эксперимента приведены на графике.

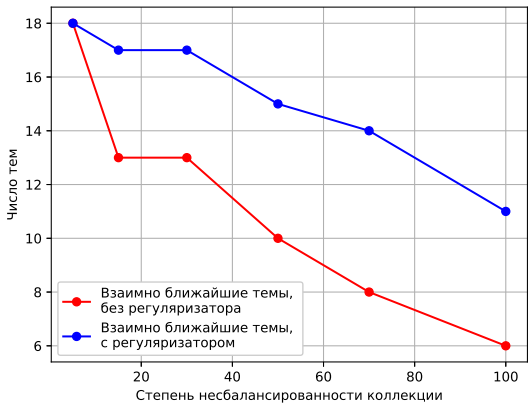

Рис. 2: Зависимость числа восстановленных взаимно близких тем от степени несбалансированности коллекции, коллекции, сгенерированные на основе 20newsgroups

Число взаимно близких тем убывает с ростом степени несбалансированности для обеих моделей, однако для модели с добавлением регуляризатора оно остается больше половины тем оказались взаимно близки при значении степени несбалансированности 100. Подберем коэффициент регуляризации модели. Рассмотрим коллекции иного вида баланса тем, где темы также делятся на равномощные крупные и равномощные мелкие, однако число крупных больше одной. Сделаем подвыборку 20newsgroups из 4 тем по 500 монотематических документов и 16 тем по 10 монотематических документов. Таким образом, степень несбалансированности полученной коллекции равна 50. Добавим в модель регуляризатор семантической неоднородности и построим зависимость количества

получившихся взаимно ближайших тем от коэффициента регуляризации

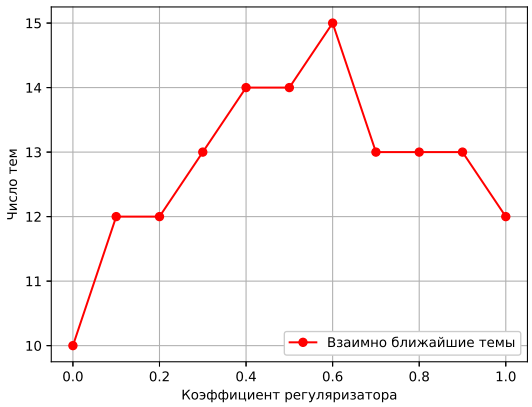

Рис. 3: Эксперимент на несбалансированной коллекции, полученной подвыборкой из 20newsgroups. Зависимость между числом взаимно близких тем и коэффициентом регуляризации

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, что соответствует использованному в предыдущем эксперименте значению. Таким образом, проблема несбалансированности наблюдается как для реальных так и для синтетических коллекций и добавление регуляризатора приводит к выявлению большего числа взаимно близких с исходными тем.

3.4 Подбор параметров генерации коллекции

Проверим, как меняется качество работы тематической модели в зависимости от состава коллекции. Для этого сгенерируем синтетические коллекции с различным числом крупных тем. Будем генерировать коллекции из нескольких крупных тем по 500 документов и оставшихся мелких тем по 10 документов. Общее число тем для каждой коллекции равно 100. Зависимость числа взаимно близких тем от числа крупных тем при генерации представлена на графике

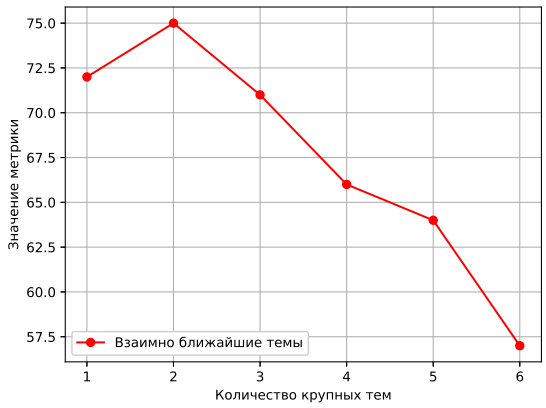

Рис. 4: Зависимость метрик от числа тем

Заметим, что чем больше число крупных тем, тем хуже модель восстанавливает искомые темы. Поэтому в процессе составления коллекции для дальнейших экспериментов необходимо генерировать несколько крупных тем. В то же время мелкие темы также необходимы для лучшей демонстрации их слияния с более крупными. Поэтому в дальнейшем будут генерироваться коллекции с 6 темами по 500 документов и остальными по 10, если не указано иное.

Проверим, что параметр n_{group}генерации коллекции не влияет на основные свойства решения задачи тематического моделирования. Для этого для коллекции 20news построим десять коллекций со степенью несбалансированности $\tau = 50$. Именно такая степень получается для коллекции, описанной в итогах прошлого эксперимента. Каждая из этих коллекций генерировалась на основе монотематических документов 20news, 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 предыдущей секции. Результат приведен на графике

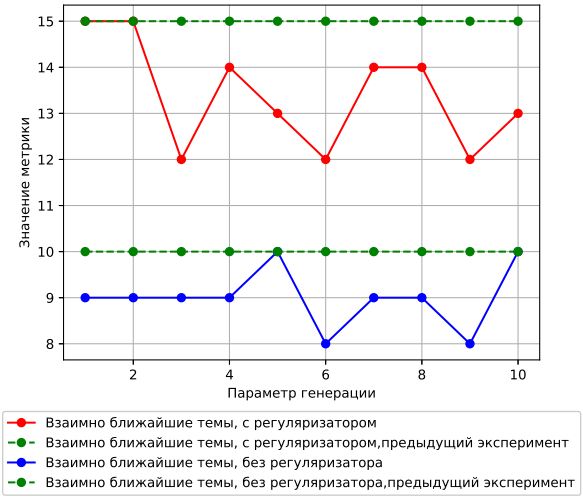

Рис. 5: Эксперимент по подбору параметра n_{group} генерации коллекции

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. Значения параметра больше 10 в экспериментах не рассматривались.

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, сгенерированная на основе 20newsgroups в первом эксперименте. Результаты работы модели с добавлением стратегии приведены на графике

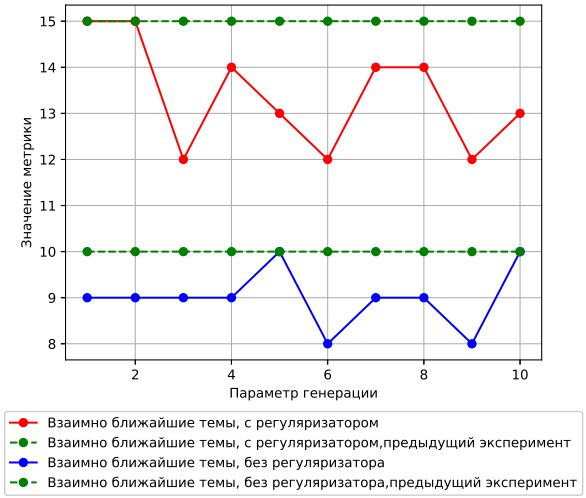

Рис. 6: Зависимость числа восстановленных взаимно близких тем от степени несбалансированности коллекции при добавлении стратегии разреживания

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.

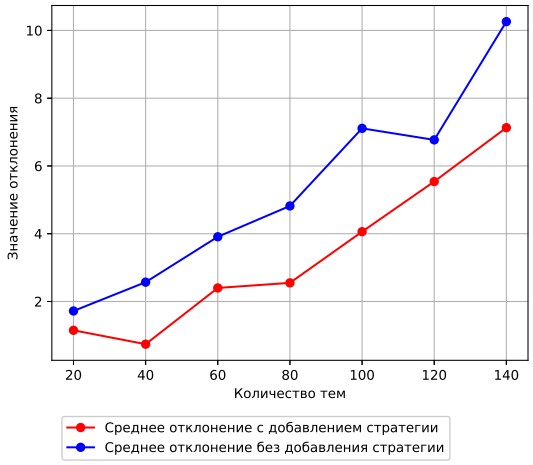

Рис. 7: Измерение отклонения между запусками модели в зависимости от числа тем

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.

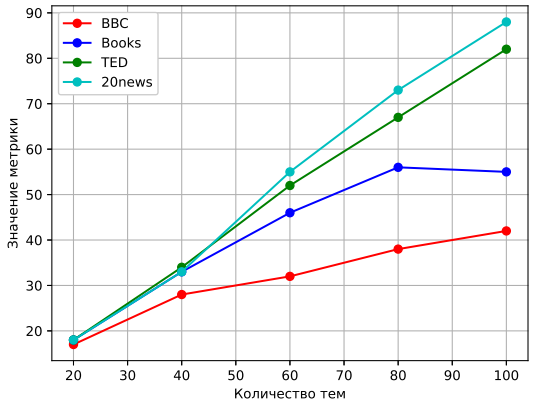

Рис. 8: Среднее число взаимно близких тем между повторными запусками одной модели на одном датасете

Для коллекции новостей 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 декоррелирования. Для этого сгенерируем коллекцию со степенью несбалансированности 50 на основе 20news описанным выше способом. Исследовалась зависимость числа найденных моделью взаимно ближайших тем от коэффициента регуляризатора декоррелирования при постоянном коэффициенте регуляризатора семантической неоднородности равном 0.6. Результаты приведены на графике ниже.

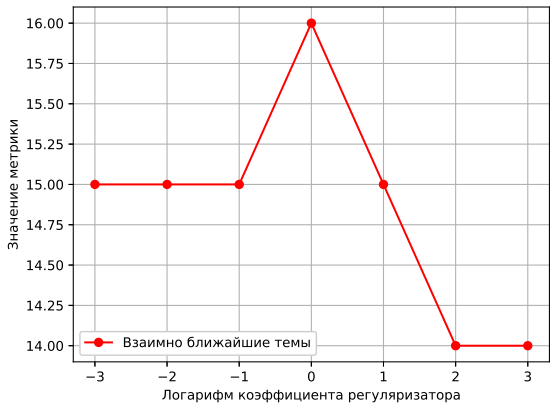

Рис. 9: Зависимость числа взаимно близких тем от десятичного логарифма коэффициента регуляризатора декоррелирования

Как видно из графиков, добавление регуляризатора декоррелирования при подборе коэффициента регуляризации повышает качество восстановления тем.

Так как регуляризатор декоррелирования улучшил результат для коэффициента регуляризатора семантической неоднородности 0.6, проверим, останется ли значение 0.6 оптимальным при его добавлении. Для этого из предыдущего графика получим наилучшее значение коэффициента регуляризатора декоррелирования. Повторим эксперимент по подбору коэффициента регуляризатора семантической неоднородности с добавлением регуляризатора декоррелирования в модель.

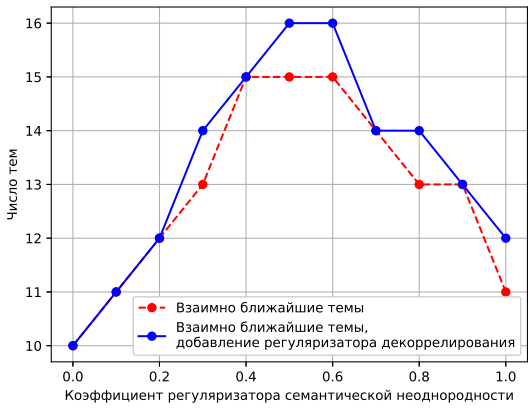

Рис. 10: Эксперимент на несбалансированной коллекции, полученной подвыборкой из 20newsgroups. Зависимость между значением метрики и коэффициентом регуляризации для модели с добавлением декоррелирования

Как и в предыдущем эксперименте, 0.6 оказалось наилучшим значением для коэффициента регуляризатора семантической неоднородности. Качество восстановления тем при добавлении декоррелятора улучшилось и число взаимно близких тем в максимуме теперь 16, а не 15.

3.7 Сравнение результатов на 20newsgroups

Покажем, как эффекты слияния и расщепления тем мешают интерпретации результатов. Для коллекции 20newsgroups построена тематическая модель на 40 темах. Полученные темы упорядочивались по числу соответствующих им документов. Для 20 тем, которым соответствуют меньше всего документов в коллекции были оставлены первые 10 документов соответствующих тем. Документы, соответствующие осстальным темам, из коллекции не удалялись. Степень несбалансированности полученной коллекции равна 80.6

На полученной коллекции строятся две модели. Первая - с добавлением регуляризатора декоррелирования. Вторая - с добавлением регуляризатора семантической неоднородности, регуляризатора разреживания и регуляризатора декоррелирования. Коэффициенты регуляризатора выбирались как указанные в экспериментах ранее. Для регуляризатора се-

мантической неоднородности коэффициент равен 0.6. Ниже в таблице приведены наиболее вероятные слова для близких тем между исходной моделью и моделью без регуляризатора.

	Исходная модель	Модель без регуляризатора
1	size frame pixel quality image	user graphic support software animation
2	size frame pixel quality image	display image library animation screen
3	spirit soul eternal life heaven	death sin church doctrine god
4	spirit soul eternal life heaven	death faith heaven sin believe
5	void include null int code	application window use get problem code
6	menu mouse button application window	application user window server program
7	firearm legal citizen law crime	wrong society person life crime
8	wrong society orientation partner relationship	wrong society person life crime

Сстроки 1-2 - одна и та же тема для первой модели. Ей соответствуют 2 похожих по смыслу темы модели без регуляризатора. Такой эффект соответствует расщеплению тем- когда одна крупная тема разделяется на 2 более маленьких по мощности. Аналогичный пример приведен в строках 3-4. В то же время мелкая темы в строке 7, у которой было оставлено только 10 документов, слилась с темой из строки 8. Имел место и другой эффект, когда 2 различные темы сливались сами с собой и с другими, как в случае исходных тем из строк 5-6. Для них можно выделить соответствующие им наиболее близкие темы, но такие темы не стали взаимно близкими, а соответствующие им темы не интерпретируемые.
Темы, построенные моделью без регуляризатора не соответствуют исходным. Однако такое могло случиться из-за некорректного выбора документов для удаления из коллекции. Проверим, что модель с добавлением регуляризатора способна восставновить исходные темы.

Исходная модель	Модель с регуляризатором
size frame pixel quality image	animation interactive graphic draw image
spirit soul eternal life heaven	soul holy god divine revelation
void include null int code	function include void return null
menu mouse button application window	button menu user command server
firearm legal citizen law crime	murder person crime moral society
wrong society orientation partner relationship	orientation wife partner relationship marriage

Темы, которые восстановила модель с добавлением регуляризатора близки по смыслу к исходным. Поэтому указанное несоответствие тем не является следствием алгоритма построения коллекции. Таким обра-

зом, модель с добавлением регуляризатора помогла устранить эффекты слияния и расщепления тем.

4 Заключение

- Показано, что тематическая несбалансированность коллекции приводит к дроблению крупных и слиянию мелких тем
- Предложен алгоритм устранения проблемы несбалансированности, заключающийся в добавлении регуляризации на основе семантической однородности тем
- Проведены эксперименты, демонстрирующие возможные модификации модели и показывающие, что модель улучшает интерпретируемость результатов обработки несбалансированных коллекций