Оценка оптимального количества тематик в тематической модели: подход на основе качества кластеров

Ф.В.Краснов

Аннотация— **Несмотря на то, что тематические модели используются для построения кластеров документов уже более 10 лет до сих пор существует проблема выбора оптимального количества тематик. Авторами проанализирован ряд ключевых исследований, предпринятых на эту тему за последнее время. Основная проблема состоит в отсутствии стабильной метрики качества тематик, полученных в ходе построения тематической модели. Авторами проведён анализ внутренних метрик тематической модели: Когерентность, Контрастность и Чистота ядра тем для определения оптимального количества тем и сделано заключение об их неприменимости для решения этой задачи. Авторами проанализирован подход к выбору оптимального количества тем на основе полученных кластеров. Для этого были рассмотрены поведения метрик валидации кластеров Davies Bouldin Index, Silhouette Coefficient и Calinski-Harabaz Index в зависимости от количества тематик. В основу предлагаемой авторами новой методики определения оптимального количества тематик легли следующие принципы: настройка тематической модели с последовательной регуляризацией (ARTM) для отделения шумовых тематик; качестве векторного представления слов, входящих в тематики, авторы предложили использовать плотных представления (embedded) векторов (GloVe, FastText, Word2Vec); для оценки расстояний авторы предложили использовать косинусную меру, которая на векторах с большой размерностью работает лучше, чем Евклидова мера расстояния.**

Разработанная авторами методика получения оптимального количества тем была опробована на коллекции научных статей из библиотеки *OnePetro,* отобранным определенным тематикам рубрикатора. Эксперимент показал, что предложенная авторами методика позволяет точно оценить оптимальное количество тематик для тематической модели, построенной по небольшой коллекции англоязычных документов.

Ключевые слова— **кластеризация, тематическая модель, ARTM, метрики, валидация кластеров.**

I. Введение

Тематические модели успешно используются для кластеризации текстов уже на протяжении многих лет. Один из наиболее распространённых подходов к тематическому моделированию на основе LDA [4] моделирует выбранное в качестве параметра фиксированное количество тем на основе распределения Дирихле, для слов и документов. В результате получается плоская, мягкая вероятностная кластеризация терминов по темам и документов по темам. Все полученные тематики равноправны, они сами по себе не создают каких-либо характерных признаков, которые могли бы помочь исследователю определить наиболее полезные темы, то есть выбрать подмножество тем, которые лучше всего подходят для интерпретации человеком. Проблема нахождения метрики, характеризующей такую интерпретируемость, является предметом изучения многих исследователей [1, 2, 3, 21].

Тематическая модель не умеет читать мысли исследователя и поэтому должна иметь параметры настройки на задачу, которую собирается решать исследователь. Тематические модели на основе LDA обладают следующими параметрами согласно исследованиям [5,6]:

- alpha: параметр априорного распределения Дирихле для документов-тем,
- beta: параметр априорного распределения Дирихле для тем-слов,
- tn: Количество тем,
- b: Количество отбрасываемых начальных итераций при семплировании по Гиббсу,
- n: Количество присемплов,
- si: Интервал семплирования.

В исследовании [6], опубликованном в 2018 году, предпринята попытка нахождения оптимальных значений вышеприведённых параметров с помощью алгоритма *Differential Evolution* [7]. В качестве кост-функции (метрики) была выбрана модифицированная метрика Jaccard Similarity. В результате был создан новый алгоритм LDADE, в котором появились свободные параметры от алгоритма Differential Evolution, которые тоже нужно будет оптимизировать.

Существует разница между оценкой полного набора тем и оценкой отдельных тем для фильтрации нежелательной информации (шума). Для оценки полного набора тем исследователи обычно смотрят на метрику перплексия [8] для корпуса документов. Такой подход не очень хорошо работает по результатам исследований [9,10] потому что метрика перплексия не имеет явного минимума, а с ростом итераций выходит

Статья получена 15 января 2018.

Ф.В.Краснов, к.т.н., эксперт, ООО «Газпромнефть НТЦ», 190000 г. Санкт-Петербург, набережная реки Мойки д.75-79,. krasnov.fv@gazprom-neft.ru, orcid.org/0000-0002-9881-7371, РИНЦ 8650-1127

на асимптоту [12]. Наиболее распространённое использование метрики перплексия состоит в том, чтобы обнаружить «эффект локтя», то есть когда характер роста упорядоченности модели принципиально изменяется. Перплексия зависит от мощности словаря и распределения частот слов в коллекции, отсюда получаем её недостатки:

- невозможно оценивать качество удаления стоп-слов и нетематических слов
- нельзя сравнивать методы разреживания словаря
- нельзя сравнивать униграммные и n-граммные модели.

Сами авторы LDA сделали исследование качества тематик с помощью Байесовского подхода в работе [25]. Следует отметить, что вопрос оптимального количества тематик решён с помощью иерархического процесса Дирихле (HDP) [17], не для документов, а для коллекции в целом. Поясним разницу между Latent Dirichlet Allocation (LDA), иерархическими процессами Дирихле (HDP) и иерархическими распределениями Дирихле (hLDA) [18,19], так как это разные модели. LDA создаёт плоскую, мягкую вероятностную кластеризацию терминов по темам и документам по темам. В модели HDP вместо фиксированного количества тем для документа количество тем генерируется процессом Дирихле, что приводит к тому, что количество тем также является случайной величиной. «Иерархическая» часть имени относится к другому уровню, добавляемому процесс Дирихле, создающий количество тем, а самим темы по-прежнему являются плоскими кластерами. Модель hLDA является адаптацией LDA, которая моделирует темы как распределение нового, заранее определённого количества тем, взятых из распределения Дирихле. Модель hLDA по-прежнему рассматривает количество тем как гиперпараметр, то есть независимо от данных. Разница в том, что кластеризация теперь иерархическая: модель hLDA изучает кластеризацию первого набора тем, предоставляя более общие абстрактные отношения между темами (а, следовательно, словами и документами). Отметим, что все три описанные модели (LDA, HDP, hLDA) добавляют новые свободные параметры, которые требуют оптимизации, как отмечено в исследовании [28].

Одним из основных требований к тематическим моделям является интерпретируемость человеком [20]. Другими словами, содержат ли темы слова, которые, согласно субъективным суждениям человека, являются репрезентативными для единой когерентной концепции. В работе [11] Ньюмен показал, что человеческая оценка интерпретируемости хорошо коррелирует с автоматизированной мерой качества, называемой когерентностью.

В исследовании [13] 2018 года предлагается минимизировать энтропии Реньи и Цаллиса для нахождения оптимального количества тем в тематическом моделировании. В этом исследовании тематические модели, полученные из больших коллекций текстов, рассматриваются как неравновесные сложные системы, где количество тем рассматривается как эквивалент температуры. Это позволяет вычислять свободную энергию таких систем - значение, через которое легко выражаются энтропии Реньи и Цаллиса. Полученные на основе энтропий метрики позволяют найти минимум в зависимости от количества тем для больших коллекций, но на практике небольшие коллекции документов встречаются так же достаточно часто.

В исследовании [16], опубликованном в 2018 году, года предложен матричный подход к повышению точности определения тематик без использования оптимизации. Но с другой стороны в исследовании [15] отмечено, что повышение точности модели противоречит с интерпретируемостью человеком. В частности, в исследовании [22], завершённом в 2018 году, создан фреймворк VisArgue, предназначенный для визуализации процесса обучения модели с целью определения наиболее интерпретируемых тематик.

Использование статистической меры TF-IDF в качестве метрики количественной оценки качества тематик изучено в работе [24]. Так же есть ряд исследований совмещения преимуществ тематических моделей и плотных представлений векторов слов [14, 23, 26, 27].

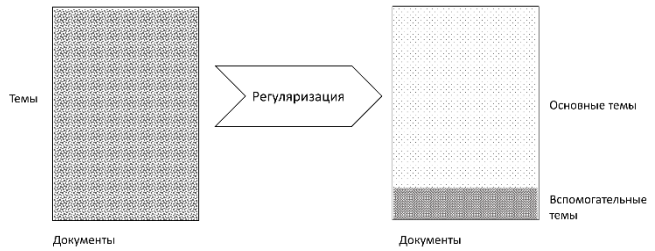
Мотивацией проведённого авторами исследования стал тот факт, что изучение стабильной метрики для качества тематик продолжаются. И использование кластерного анализа является одним из инструментов для анализа стабильности тематик [47] и оптимального количества тем [42], но при этом не рассматриваются преимущества от возможностей специальной подготовки тематической модели с последовательной регуляризацией и плотного представления векторов слов.

Для валидации качества кластеров разработано достаточно много метрик. Например, метрики Partition Coefficient [30], Dunn Index [29], а также DPI [31] и её модификации [32, 33], Silhouette [41], которые задействованы в алгоритмах кластеризации. Но в случае тематической модели мы уже получаем кластеры тематик и не нуждаемся в алгоритме кластеризации, а только в оценке полученных кластеров. Для валидации кластеров необходимо их рассмотрение в пространстве обладающим понятиями близости и удалённости. Для слов такими пространством является векторное представление слов. Значительные результаты в этом направлении получены в исследованиях [34, 35, 36]. Слова, представленные в виде плотных векторов, отражают смысловое представление и обладают свойствами близости и удалённости. Таким образом, представив тематики в виде плотных векторов авторы создали новую вариацию метрики *DPI* для тематик, которую авторы назвали *cDPI*.

Данная статья построена следующим образом: в разделе «Метод» описан предлагаемый авторами метод исследования и сформулирована исследовательская гипотеза, в разделе «Эксперимент» представлены результаты апробации новой метрики качества тематик.

II. Методология исследования

Рассмотрим способы построения тематической модели для конкретной коллекции документов. Будем называть коллекцию однородной, если она содержит документы одного типа. Например, коллекция научных статей одной конференции, созданных по единому шаблону, является однородной. В случае однородной коллекции научных статей, каждый документ обладает схожей структурой, постулируемой шаблоном конференции. Все научные статьи состоят из введения, представления результатов исследования и заключения. Таким образом, с точки зрения гипотетической тематической модели можно представить документ в виде распределения основной темы и вспомогательных



тем: введения и заключения. Конечно, основные темы в

Рисунок 1 Схема матрицы «тема-документ».

разных документах могут быть разные. Но мы можем ограничить коллекцию научных статей выбором определённых рубрик из тематического рубрикатора конференции. Тогда число тем нам будет известно. На рисунке 1 представлена матрица распределения тем по документам.

Как мы видим на левой части рисунка 1 при построении тематической модели выделены такие темы, которые распределены по документам достаточно однородно. Такая картина вероятностей матрицы «тема-документ» может быть получена с помощь, например, модели на основе алгоритма LDA [4]. А на правой части рисунка 1 показан результат работы модели с последовательной регуляризацией ARTM [1]. Основные и вспомогательные темы выделены с помощью управления процессом обучения модели. Принцип отнесения темы к вспомогательным может быть сформулирован, как наличие такой темы в подавляющем количестве документов. То есть, вероятности вспомогательной темы будут распределены по документам однородно и плотно. А основная тема будет представлена в виде разряженного вектора для каждого документа, так как каждый документ характеризуется одной основной темой.

Покажем, что существующие внутренние метрики тематической модели не подходят для определения оптимального количества тем. Для этого рассмотрим внутренние автоматизированные метрики качества тем. Введём понятие ядра тематик:

$$W_t = \{\ w \in W \mid p(t|w) \ \geq \ treshold\ \} \qquad (1)$$

На основе ядра тем могут быть рассчитаны следующие метрики качества тематической модели:

- *Чистота тем*: $Purity = \sum_{w \in W_t} p(w|t)$
- *Размер ядра:* $|W_t|$
- *Контраст тем:* $\frac{1}{|W_t|}\sum_{w \in W_t} p(t|w)$
- *Когерентность тем:* $Coh_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=1}^{k} PMI\ (w_i, w_j)$, где k — окно в котором вычисляются совместные употребления слов, поточечная взаимная информация $PMI(w_i\ ,\ w_j) = \log \frac{N\, N_{w_i w_j}}{N_{w_i} N_{w_j}}$, $N_{w_i w_j}$ — число документов, в которых слова w_i и w_j хотя бы один раз встречаются в окне k. N_{w_i} - число документов, в которых слово w_i встретилось хотя бы один раз, а N – это количество слов в словаре .

Как видно из формул для внутренних метрик тематической модели каждая из этих метрик может быть измерена для разного количества тем (tn). Рассмотрим поведение метрики *Размер ядра* в зависимости от количества тем. При увеличении количества тем размер ядра будет уменьшаться, так как при построении матриц «тема-слова» и «документ-тема» должны выполняться нормирующие условия: сумма вероятностей должна быть равна единице. Для метрик *Чистота тем и Контраст тем* характер изменений при росте количества тем также будет монотонно убывающим, так как сумма вероятностей тематик, входящих в ядро будет уменьшаться. С другой стороны, для метрики *Когерентность тем* поведение с ростом количества тем будет монотонно возрастающим, так как будет расти вклад от PMI. Конкретный характер изменения рассмотренных метрик может отличаться, поэтому целесообразно с помощью численных методов попробовать найти экстремум, если он есть.

Рассмотрение качества тематик коротких сообщениях с точки зрения кластеров было проведено в работе [37] с помощью NMF (Non-negative Matrix Factorization) и метрик отражающих энтропию кластеров. Матричный подход (LSI + SVD) к выделению кластеров тематик из программного кода был исследован в работе [38] с модифицированной метрикой близости векторов. В исследовании качества тематической модели [42] использована метрика Silhouette Coefficient [41] с Евклидовым расстоянием для разряженных векторов тематик. Таким образом, в этих работах не исследованным остаются кластеры в пространстве плотных векторов слов, составляющих тематики, и не Евклидовые расстояния в метриках.

В работах [39, 12, 40] обнаружена и исследована нестабильность тематик относительно порядка обрабатываемых документов. Поэтому для вычисления метрики качества тематик необходимо провести расчёты для корпуса документов со случайным порядком, чтобы исключить наличие зависимости от порядка документов. В работе [43] показана возможность стабилизации тематической модели с помощью регуляризации.

На основе проведённого анализа авторами был сформулирован методический каркас, изображённый в

виде схемы на рисунке 2.

Рисунок 2 Методический каркас исследования.

На рисунке 2 изображена последовательность действий, повторяемая для одного корпуса документов значительное количество раз, по порядку сравнимому с количеством документов в корпусе. Справа отображены действия, которые выполнены однократно: формирование словаря, настройка параметров регуляризации тематической модели и преобразование разряженного пространства представления тематик в плотное представление. На основании данного методического каркаса были разработаны и проведены цифровые эксперименты, описанные в следующем разделе.

III. Эксперимент

Для эксперимента был использован корпус научно-технических статей по темам, связанным с разработкой нефе-газовых месторождений. Всего было выбрано 1695 статей на английском языке по 10 направлениям исследований согласно рубрикатору. Создание словаря для выбранного корпуса подробно описано в предыдущем исследовании авторов [44].

Для построения тематической модели была использована библиотека BigARTM, позволяющая производить настройку тематической модели путём последовательной регуляризации. Выбор и настройка параметров регуляризации тематической модели сделаны авторами в предыдущем исследовании [44].

Для преобразования разряженного пространства векторов слов, составляющих тематики, была выбрана библиотека GloVe [34]. Для получения визуального представления о виде плотного представления тематик была сделана проекция на двумерное пространство с сохранением расстояний с помощью библиотеки MDS [45]. Полученный таким образом вид кластеров тематик представлен на рисунке 3.

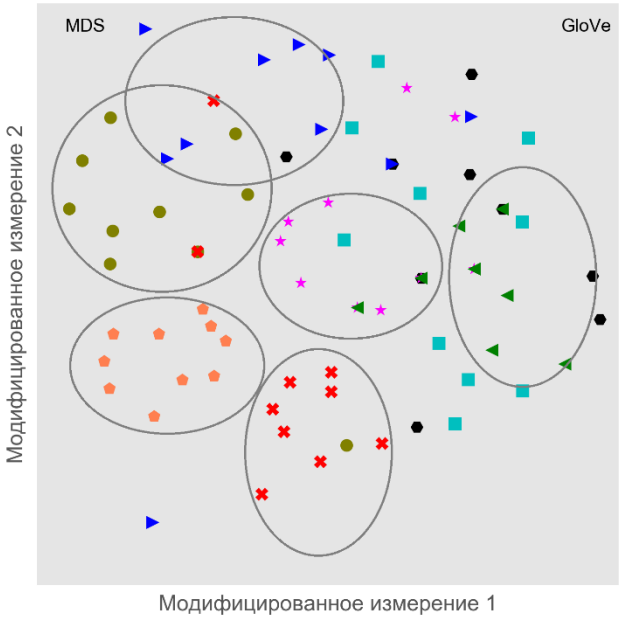


Рисунок 3 Проекция плотного представления тематик с сохранением расстояний.

На рисунке 3 различными маркерами выделены двумерные проекции слов из тематик. Овалы сделаны для того, чтобы подчеркнуть уверенную визуальную сгруппированность слов в тематиках.

На рисунке 4 представлены предварительные расчёты поведения основных метрик тематической модели, настроенной в соответствии с предложенной авторами методикой, в зависимости от количества тем.

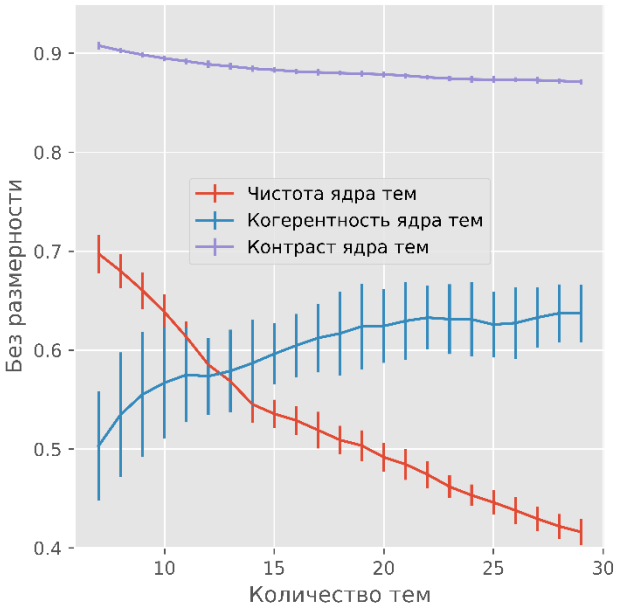


Рисунок 4 Зависимости основных внутренних метрик качества тематической модели от количества тем.

Как мы видим из рисунка 4, характер зависимостей носит монотонный характер и не позволяет определить оптимальное количество тем. Измерения основных внутренних метрик сделаны для 1000 различных случайных порядков документов. По оси *y* отложено значение одного стандартного отклонения. Мы видим, что для метрики *Контраст ядра тем* отклонения

минимальны. Для метрик *Чистота ядра тем* и *Когерентность ядра тем* большее значения характеризуют лучшее качество тематической модели. Характерной точкой можно считать количество тем равное 12, когда кривые изменения метрики *Чистота ядра тем* и *Когерентность ядра тем* пересекаются.
Рассмотрим зависимости метрик *Calinski-Harabaz Index* [46], *Silhouette Coefficient* [41], используемых для валидации количества кластеров.

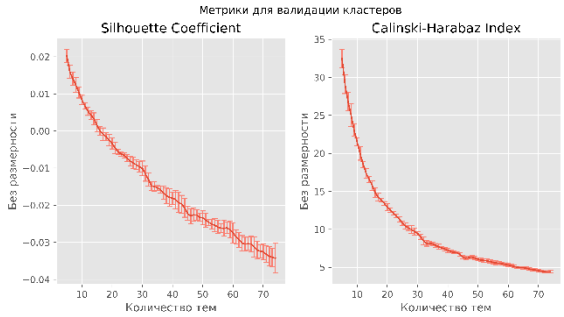


Рисунок 5 Метрики валидации кластеров.

Как можно увидеть из рисунка 5 метрики *Calinski-Harabaz Index* и *Silhouette Coefficient* не дают возможности определить оптимальное количество тем. С ростом количества тем значения этих метрик уменьшаются, что говорит о том, что с точки зрения этих метрик кластеры становятся хуже.
По-другому в зависимости от количества тем ведёт себя метрика *cDBI*, разработанная авторами и изображённая на рисунке 6.

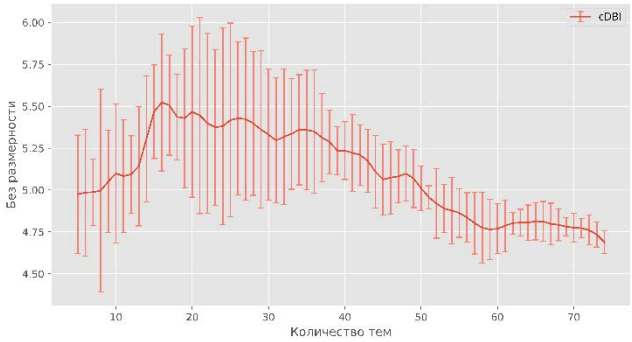


Рисунок 6 Метрика cDBI.

На рисунке 6 можно видеть явно выраженный максимум при количестве тематик равному 16.

Алгоритм для расчёта метрики *cDBI* основан на идеологии метрики Davies Bouldin Index, предложенной в работе [31] и модифицированной в работах [32,33].

$$V := GloVe\ (\ ARTM\ (\ tn, \mu, corpus\ of\ texts)\ )$$
$$for\ t \in\ W:$$
$$C_t := \sum_{i \in t} V_t^{(i)}$$
$$D_t := \frac{1}{\dim t} \sum_{i \in t} \frac{C_t \cdot V_t^{(i)}}{|\ C_t|\ |V_t^{(i)}|}$$
$$cDBI := \frac{1}{\dim W} \sum_{t \in T} \frac{D_t}{C_t}$$

Алгоритм 1. Расчёт метрики *cDBI*.

В приведённом выше Алгоритме 1 T обозначает количество выбранных тем, μ - это регуляризующие коэффициенты.
Таким образом, с помощью метрики *cDBI* возможно найти оптимальное количество тем для коллекции документов.

IV. Заключение

Авторами исследован вопрос выбора оптимального количества тематик для построения тематической модели для заданного корпуса текстов. Результатом данного исследования стала методика, позволяющая определить оптимальное количество тематик для небольшого корпуса однородных англоязычных документов.
Важным методическим приёмом авторов является подготовка тематической модели с помощью последовательной регуляризации. В предыдущих исследованиях данной коллекции документов [44] авторами были получены численные оценки коэффициентов при регуляризирующих составляющих тематической модели (μ).
При формировании коллекции текстов были заданы условия, ограничивающие число тематик научных статей согласно тематическому рубрикатору до 10. Суть эксперимента состояла в том, чтобы подтвердить выбранное число тематик с помощью оптимизационного подхода на основе разработанной авторами метрики качества тематической модели — cDBI. В результате эксперимент показал, что максимальное значение метрики cDBI достигается при количестве тем равному 16. Данный результат получен при много прогонном обучении модели, чтобы исключить эффект зависимости от порядка документов в коллекции.

Библиография

1. Vorontsov K., Potapenko A., Plavin A. Additive regularization of topic models for topic selection and sparse factorization //International Symposium on Statistical Learning and Data Sciences. – Springer, Cham, 2015. – С. 193-202.
2. Koltsov S., Pashakhin S., Dokuka S. A Full-Cycle Methodology for News Topic Modeling and User Feedback Research //International Conference on Social Informatics. – Springer, Cham, 2018. – С. 308-321.
3. Seroussi Y., Bohnert F., Zukerman I. Authorship attribution with author-aware topic models //Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. – Association for Computational Linguistics, 2012. – С. 264-269.
4. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation //Journal of machine Learning research. – 2003. – Т. 3. – №. Jan. – С. 993-1022.
5. Binkley D. et al. Understanding LDA in source code analysis //Proceedings of the 22nd international conference on program comprehension. – ACM, 2014. – С. 26-36.
6. Agrawal A., Fu W., Menzies T. What is wrong with topic modeling? And how to fix it using search-based software engineering //Information and Software Technology. – 2018. – Т. 98. – С. 74-88.
7. Storn R., Price K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces //Journal of global optimization. – 1997. – Т. 11. – №. 4. – С. 341-359.
8. Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. On smoothing and inference for topic models //Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. – AUAI Press, 2009. – С. 27-34.
9. Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. Evaluation methods for topic models //Proceedings of the 26th annual

international conference on machine learning. – ACM, 2009. – С. 1105-1112.
10. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. Reading tea leaves: How humans interpret topic models //Advances in neural information processing systems. – 2009. – С. 288-296.
11. Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. Automatic evaluation of topic coherence //Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. – Association for Computational Linguistics, 2010. – С. 100-108.
12. 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. – ACM, 2014. – С. 161-165.
13. Koltcov S. Application of Rényi and Tsallis entropies to topic modeling optimization //Physica A: Statistical Mechanics and its Applications. – 2018. – Т. 512. – С. 1192-1204.
14. Batmanghelich, K., Saeedi, A., Narasimhan, K., & Gershman, S. Nonparametric spherical topic modeling with word embeddings //arXiv preprint arXiv:1604.00126. – 2016.
15. Lipton Z. C. The mythos of model interpretability //arXiv preprint arXiv:1606.03490. – 2016.
16. Bing X., Bunea F., Wegkamp M. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics //arXiv preprint arXiv:1805.06837. – 2018.
17. Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. Sharing clusters among related groups: Hierarchical Dirichlet processes //Advances in neural information processing systems. – 2005. – С. 1385-1392.
18. Blei D. M., Griffiths T. L., Jordan M. I. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies //Journal of the ACM (JACM). – 2010. – Т. 57. – №. 2. – С. 7.
19. Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., & Blei, D. M. Hierarchical topic models and the nested chinese restaurant process //Advances in neural information processing systems. – 2004. – С. 17-24.
20. Rossetti M., Stella F., Zanker M. Towards explaining latent factors with topic models in collaborative recommender systems //Database and Expert Systems Applications (DEXA), 2013 24th International Workshop on. – IEEE, 2013. – С. 162-167.
21. Fang D, Yang H, Gao B, Li X. Discovering research topics from library electronic references using latent Dirichlet allocation //Library Hi Tech. – 2018. – Т. 36. – №. 3. – С. 400-410.
22. El-Assady, M., Sevastjanova, R., Sperrle, F., Keim, D., & Collins, C. Progressive learning of topic modeling parameters: a visual analytics framework //IEEE transactions on visualization and computer graphics. – 2018. – Т. 24. – №. 1. – С. 382-391.
23. Law, J., Zhuo, H. H., He, J., & Rong, E. LTSG: Latent Topical Skip-Gram for Mutually Learning Topic Model and Vector Representations //arXiv preprint arXiv:1702.07117. – 2017.
24. Nikolenko S. I., Koltcov S., Koltsova O. Topic modelling for qualitative studies //Journal of Information Science. – 2017. – Т. 43. – №. 1. – С. 88-102.
25. Mimno D., Blei D. Bayesian checking for topic models //Proceedings of the conference on empirical methods in natural language processing. – Association for Computational Linguistics, 2011. – С. 227-237.
26. Das R., Zaheer M., Dyer C. Gaussian lda for topic models with word embeddings //Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). – 2015. – Т. 1. – С. 795-804.
27. Nguyen, D. Q., Billingsley, R., Du, L., Johnson, M., & Fe, S. (2015). Improving Topic Models with Latent Feature Word Representations.
28. Bryant M., Sudderth E. B. Truly nonparametric online variational inference for hierarchical Dirichlet processes //Advances in Neural Information Processing Systems. – 2012. – С. 2699-2707.
29. Dunn J. C. Well-separated clusters and optimal fuzzy partitions //Journal of cybernetics. – 1974. – Т. 4. – №. 1. – С. 95-104.
30. Bezdek J. C. Cluster validity with fuzzy sets. – 1973.
31. Davies D. L., Bouldin D. W. A cluster separation measure //IEEE transactions on pattern analysis and machine intelligence. – 1979. – №. 2. – С. 224-227.
32. Halkidi M., Batistakis Y., Vazirgiannis M. Clustering validity checking methods: part II //ACM Sigmod Record. – 2002. – Т. 31. – №. 3. – С. 19-27.
33. Xie X. L., Beni G. A validity measure for fuzzy clustering //IEEE Transactions on Pattern Analysis & Machine Intelligence. – 1991. – №. 8. – С. 841-847.
34. Pennington J., Socher R., Manning C. Glove: Global vectors for word representation //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 1532-1543.
35. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. Enriching word vectors with subword information //arXiv preprint arXiv:1607.04606. – 2016.
36. Wu, L., Fisch, A., Chopra, S., Adams, K., Bordes, A., & Weston, J. Starspace: Embed all the things! //arXiv preprint arXiv:1709.03856. – 2017.
37. Bicalho, P. V., de Oliveira Cunha, T., Mourao, F. H. J., Pappa, G. L., & Meira, W. Generating Cohesive Semantic Topics from Latent Factors //Intelligent Systems (BRACIS), 2014 Brazilian Conference on. – IEEE, 2014. – С. 271-276.
38. Kuhn A., Ducasse S., Gírba T. Semantic clustering: Identifying topics in source code //Information and Software Technology. – 2007. – Т. 49. – №. 3. – С. 230-243.
39. Chuang J. et al. TopicCheck: Interactive alignment for assessing topic model stability //Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2015. – С. 175-184.
40. Greene D., O'Callaghan D., Cunningham P. How many topics? stability analysis for topic models //Joint European Conference on Machine Learning and Knowledge Discovery in Databases. – Springer, Berlin, Heidelberg, 2014. – С. 498-513.
41. Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis //Journal of computational and applied mathematics. – 1987. – Т. 20. – С. 53-65.
42. Mehta V., Caceres R. S., Carter K. M. Evaluating topic quality using model clustering //Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on. – IEEE, 2014. – С. 178-185.
43. Koltcov, S., Nikolenko, S. I., Koltsova, O., Filippov, V., & Bodrunova, S. Stable topic modeling with local density regularization //International Conference on Internet Science. – Springer, Cham, 2016. – С. 176-188.
44. Krasnov F., Ushmaev O. Exploration of Hidden Research Directions in Oil and Gas Industry via Full Text Analysis of OnePetro Digital Library //International Journal of Open Information Technologies. – 2018. – Т. 6. – №. 5. – С. 7-14.
45. Borg I., Groenen P. Modern multidimensional scaling: theory and applications //Journal of Educational Measurement. – 2003. – Т. 40. – №. 3. – С. 277-280.
46. Caliński T., Harabasz J. A dendrite method for cluster analysis //Communications in Statistics-theory and Methods. – 1974. – Т. 3. – №. 1. – С. 1-27.
47. Mantyla M. V., Claes M., Farooq U. Measuring LDA topic stability from clusters of replicated runs //Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. – ACM, 2018. – С. 49.

Evaluation of Optimal Number of Topics of Topic Model: An Approach Based on the Quality of Clusters

Fedor Krasnov

Abstract — **Although topic models have been used to build clusters of documents for more than ten years, there is still a problem of choosing the optimal number of topics. The authors analyzed many fundamental studies undertaken on this subject in recent years. The main problem is the lack of a stable metric of the quality of topics obtained during the construction of the topic model. The authors analyzed the internal metrics of the topic model: Coherence, Contrast, and Purity to determine the optimal number of topics and concluded that they are not applicable to solve this problem. The authors analyzed the approach to choosing the optimal number of topics based on the quality of the clusters. For this purpose, the authors considered the behavior of the cluster validation metrics: Davies Bouldin Index, Silhouette Coefficient and Calinski-Harabaz.**

The cornerstone of the proposed new method of determining the optimal number of topics based on the following principles: setting up a topic model with additive regularization (ARTM) to separate noise topics; using dense vector representation (GloVe, FastText, Word2Vec); using a cosine measure for the distance in cluster metric that works better on vectors with large dimensions than The Euclidean distance.

The methodology developed by the authors for obtaining the optimal number of topics was tested on the collection of scientific articles from the Onepetro library, selected by specific themes. The experiment showed that the method proposed by the authors allows assessing the optimal number of topics for the topic model built on a small collection of English-language documents.

Keywords — **clustering, additive regularization topic model, validation metrics, Davies Bouldin Index, ARTM.**

REFERENCES

1. Vorontsov K., Potapenko A., Plavin A. Additive regularization of topic models for topic selection and sparse factorization //International Symposium on Statistical Learning and Data Sciences. – Springer, Cham, 2015. – С. 193-202.
2. Koltsov S., Pashakhin S., Dokuka S. A Full-Cycle Methodology for News Topic Modeling and User Feedback Research //International Conference on Social Informatics. – Springer, Cham, 2018. – С. 308-321.
3. Seroussi Y., Bohnert F., Zukerman I. Authorship attribution with author-aware topic models //Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. – Association for Computational Linguistics, 2012. – С. 264-269.
4. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation //Journal of machine Learning research. – 2003. – Т. 3. – №. Jan. – С. 993-1022.
5. Binkley D. et al. Understanding LDA in source code analysis //Proceedings of the 22nd international conference on program comprehension. – ACM, 2014. – С. 26-36.
6. Agrawal A., Fu W., Menzies T. What is wrong with topic modeling? And how to fix it using search-based software engineering //Information and Software Technology. – 2018. – Т. 98. – С. 74-88.
7. Storn R., Price K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces //Journal of global optimization. – 1997. – Т. 11. – №. 4. – С. 341-359.
8. Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. On smoothing and inference for topic models //Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. – AUAI Press, 2009. – С. 27-34.
9. Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. Evaluation methods for topic models //Proceedings of the 26th annual international conference on machine learning. – ACM, 2009. – С. 1105-1112.
10. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. Reading tea leaves: How humans interpret topic models //Advances in neural information processing systems. – 2009. – С. 288-296.
11. Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. Automatic evaluation of topic coherence //Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. – Association for Computational Linguistics, 2010. – С. 100-108.
12. 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. – ACM, 2014. – С. 161-165.
13. Koltcov S. Application of Rényi and Tsallis entropies to topic modeling optimization //Physica A: Statistical Mechanics and its Applications. – 2018. – Т. 512. – С. 1192-1204.
14. Batmanghelich, K., Saeedi, A., Narasimhan, K., & Gershman, S. Nonparametric spherical topic modeling with word embeddings //arXiv preprint arXiv:1604.00126. – 2016.
15. Lipton Z. C. The mythos of model interpretability //arXiv preprint arXiv:1606.03490. – 2016.
16. Bing X., Bunea F., Wegkamp M. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics //arXiv preprint arXiv:1805.06837. – 2018.
17. Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. Sharing clusters among related groups: Hierarchical Dirichlet processes //Advances in neural information processing systems. – 2005. – С. 1385-1392.

18. Blei D. M., Griffiths T. L., Jordan M. I. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies //Journal of the ACM (JACM). – 2010. – Т. 57. – №. 2. – С. 7.
19. Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., & Blei, D. M. Hierarchical topic models and the nested chinese restaurant process //Advances in neural information processing systems. – 2004. – С. 17-24.
20. Rossetti M., Stella F., Zanker M. Towards explaining latent factors with topic models in collaborative recommender systems //Database and Expert Systems Applications (DEXA), 2013 24th International Workshop on. – IEEE, 2013. – С. 162-167.
21. Fang D, Yang H, Gao B, Li X. Discovering research topics from library electronic references using latent Dirichlet allocation //Library Hi Tech. – 2018. – Т. 36. – №. 3. – С. 400-410.
22. El-Assady, M., Sevastjanova, R., Sperrle, F., Keim, D., & Collins, C. Progressive learning of topic modeling parameters: a visual analytics framework //IEEE transactions on visualization and computer graphics. – 2018. – Т. 24. – №. 1. – С. 382-391.
23. Law, J., Zhuo, H. H., He, J., & Rong, E. LTSG: Latent Topical Skip-Gram for Mutually Learning Topic Model and Vector Representations //arXiv preprint arXiv:1702.07117. – 2017.
24. Nikolenko S. I., Koltcov S., Koltsova O. Topic modelling for qualitative studies //Journal of Information Science. – 2017. – Т. 43. – №. 1. – С. 88-102.
25. Mimno D., Blei D. Bayesian checking for topic models //Proceedings of the conference on empirical methods in natural language processing. – Association for Computational Linguistics, 2011. – С. 227-237.
26. Das R., Zaheer M., Dyer C. Gaussian lda for topic models with word embeddings //Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). – 2015. – Т. 1. – С. 795-804.
27. Nguyen, D. Q., Billingsley, R., Du, L., Johnson, M., & Fe, S. (2015). Improving Topic Models with Latent Feature Word Representations.
28. Bryant M., Sudderth E. B. Truly nonparametric online variational inference for hierarchical Dirichlet processes //Advances in Neural Information Processing Systems. – 2012. – С. 2699-2707.
29. Dunn J. C. Well-separated clusters and optimal fuzzy partitions //Journal of cybernetics. – 1974. – Т. 4. – №. 1. – С. 95-104.
30. Bezdek J. C. Cluster validity with fuzzy sets. – 1973.
31. Davies D. L., Bouldin D. W. A cluster separation measure //IEEE transactions on pattern analysis and machine intelligence. – 1979. – №. 2. – С. 224-227.
32. Halkidi M., Batistakis Y., Vazirgiannis M. Clustering validity checking methods: part II //ACM Sigmod Record. – 2002. – Т. 31. – №. 3. – С. 19-27.
33. Xie X. L., Beni G. A validity measure for fuzzy clustering //IEEE Transactions on Pattern Analysis & Machine Intelligence. – 1991. – №. 8. – С. 841-847.
34. Pennington J., Socher R., Manning C. Glove: Global vectors for word representation //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 1532-1543.
35. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. Enriching word vectors with subword information //arXiv preprint arXiv:1607.04606. – 2016.
36. Wu, L., Fisch, A., Chopra, S., Adams, K., Bordes, A., & Weston, J. Starspace: Embed all the things! //arXiv preprint arXiv:1709.03856. – 2017.
37. Bicalho, P. V., de Oliveira Cunha, T., Mourao, F. H. J., Pappa, G. L., & Meira, W. Generating Cohesive Semantic Topics from Latent Factors //Intelligent Systems (BRACIS), 2014 Brazilian Conference on. – IEEE, 2014. – С. 271-276.
38. Kuhn A., Ducasse S., Gírba T. Semantic clustering: Identifying topics in source code //Information and Software Technology. – 2007. – Т. 49. – №. 3. – С. 230-243.
39. Chuang J. et al. TopicCheck: Interactive alignment for assessing topic model stability //Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2015. – С. 175-184.
40. Greene D., O'Callaghan D., Cunningham P. How many topics? stability analysis for topic models //Joint European Conference on Machine Learning and Knowledge Discovery in Databases. – Springer, Berlin, Heidelberg, 2014. – С. 498-513.
41. Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis //Journal of computational and applied mathematics. – 1987. – Т. 20. – С. 53-65.
42. Mehta V., Caceres R. S., Carter K. M. Evaluating topic quality using model clustering //Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on. – IEEE, 2014. – С. 178-185.
43. Koltcov, S., Nikolenko, S. I., Koltsova, O., Filippov, V., & Bodrunova, S. Stable topic modeling with local density regularization //International Conference on Internet Science. – Springer, Cham, 2016. – С. 176-188.
44. Krasnov F., Ushmaev O. Exploration of Hidden Research Directions in Oil and Gas Industry via Full Text Analysis of OnePetro Digital Library //International Journal of Open Information Technologies. – 2018. – Т. 6. – №. 5. – С. 7-14.
45. Borg I., Groenen P. Modern multidimensional scaling: theory and applications //Journal of Educational Measurement. – 2003. – Т. 40. – №. 3. – С. 277-280.
46. Caliński T., Harabasz J. A dendrite method for cluster analysis //Communications in Statistics-theory and Methods. – 1974. – Т. 3. – №. 1. – С. 1-27.
47. Mantyla M. V., Claes M., Farooq U. Measuring LDA topic stability from clusters of replicated runs //Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. – ACM, 2018. – С. 49.