УДК 004.853

Проблемы разработки алгоритмов для определения качества ансамблей тематических моделей для построения рубрикаторов

А.П. Ширяев, А.Р. Федоров, П.А. Федоров, Л.Г. Гагарина, Е.М. Портнов
Национальный исследовательский университет «МИЭТ», г. Москва, Россия
e-mail: af123@yandex.ru

Аннотация. Интеллектуальный анализ данных — одно из самых актуальных направлений исследований в современном мире. Спектр его применения чрезвычайно широк и охватывает практически все научные дисциплины. Весьма актуальна задача анализа текстовых коллекций с целью установления тематических рубрик, к которым должны быть отнесены отдельные статьи с соблюдением принципа систематизации «от общего к частному» и формированием перечня «ядерных» рубрик. Одним из методов интеллектуального анализа текстовой информации является кластеризация и, в частности, тематическое моделирование.

Решение задачи кластеризации текстовых коллекций принципиально неоднозначно, и тому есть несколько причин. Во-первых, не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд достаточно разумных критериев, но все они могут давать разные результаты. Во-вторых, число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием. В-третьих, результат кластеризации существенно зависит от метрики расстояния, выбор которой, как правило, также субъективен и определяется экспертом.

В настоящее время среди методов интеллектуального анализа данных все большее распространение получают ансамбли моделей, позволяющие значительно повысить точность результатов моделирования.

Цель данного исследования — повышение эффективности кластеризации текстовой информации при использовании ансамбля тематических моделей.

В статье рассмотрено использование алгоритма голосования на основе группы из различных оценочных алгоритмов, что позволяет выбрать наиболее

Problems of Algorithms Development to Determine Quality of Topic Models Ensembles for Make Rubricators

A.P. Shiryaev, A.R. Fedorov, P.A. Fedorov, L.G. Gagarina, E.M. Portnov
National Research University of Electronic Technology, Moscow, Russia
e-mail: af123@yandex.ru

Abstract. Intelligent data mining is one of the most relevant areas of research in the modern world. The spectrum of its application is extremely wide and covers practically all scientific disciplines. The task of analyzing text collections with the purpose of establishing thematic headings, which should be classified as separate articles with observance of the principle of systematization "from the general to the particular" and the formation of the list of "nuclear" categories, is very actual. Clustering and, in particular, topic modeling is one of the methods of intelligent text analysis.

The solution of the problem of clustering text collections is fundamentally ambiguously, and there are several reasons. Firstly, there isn't known clearly the best criterion of quality of clustering. There are a lot of reasonable criteria, but they all can give different results. Secondly, the number of clusters is usually unknown in advance and determined according by some subjective criterion. Thirdly, clustering result depends significantly on the distance metric, the choice of which is usually subjective and set by the expert.

Nowadays ensembles of models are becoming more widespread among the data mining techniques. They can significantly improve the accuracy of modeling results.

The main purpose of this research is to increase the clustering effectiveness of textual information by using the ensemble thematic models.

This article describes the usage of a voting algorithm, which is based on a group of different evaluation algorithms. Voting algorithm allows you to select the most appropriate solution, to accurately assess the quality of the topic model and to generate a set of relevant topics. Computational experiment demonstrates coincidence with the results of expert assessments and the evaluations of formal criteria in general. The concept for quality evaluation of thematic

подходящее решение, достаточно точно оценить качество тематических моделей и сформировать набор релевантных тем. В данной работе проведено исследование и предложена концепция оценки качества ансамбля тематических моделей с помощью использования простого голосующего алгоритма. Вычислительный эксперимент использования оценочного алгоритма, анализирующего поисковые запросы, демонстрирует в общем случае совпадение с результатами экспертного оценивания.

Ключевые слова. Кластерный анализ, голосующий алгоритм, качество тематических моделей, перплексия.

models ensemble, which uses the simple voting algorithm, was explored and proposed for further researches.

Keywords. Cluster analysis, voting algorithm, quality of topic models, perplexity.

DOI: 10.31432/1994-2443-2018-13-3-53-58

Интеллектуальный анализ данных — одно из самых актуальных направлений в современном мире. Одним из методов интеллектуального анализа данных является тематическое моделирование [1]. Существует большое количество разнообразных алгоритмов в данной области, но они обладают рядом недостатков:

1) не учитываются лингвистические обоснования, к тому же распределение документов по темам носит субъективный характер. Эксперт может отнести документ к разным темам в зависимости от области применения, своей квалификации и т.п.;

2) сложность и высокая стоимость интерпретации результатов экспертами;

3) выделяют также проблему устойчивости группировочных решений [2]. В классических алгоритмах решения задач кластерного анализа и тематического моделирования результаты группировки могут сильно меняться в зависимости от выбора начальных условий, порядка объектов, параметров работы алгоритмов и т.п.

Автоматизация оценки качества тематических моделей, несвязанной с перплексией (величина перплексии зависит, во-первых, от данных, а во-вторых, от количества тем) [1] и коррелирующей с мнениями экспертов, является актуальной задачей.

Для решения перечисленных проблем предлагается использование ансамбля тематических моделей и применение алгоритма, позволяющего оценить качество моделей. Ансамбль или комитет — группа моделей, предназначенная для решения одной задачи. Широко известны ансамблевые алгоритмы для решения задачи классификации, например, бустинг [3, 4], бэггинг [3, 4], метод случайных подпространств [4], стекинг [3], алгоритмы вычисления оценок Ю.И. Журавлева [6] и др.

Точность результата ансамбля выше, чем точность любой отдельно взятой модели. При этом могут использоваться как различные модели, так и одна, но с разными настроечными параметрами. Одним из достоинств ансамблевых моделей является возможность выполнения распределенной обработки информации, каждая модель может быть построена на своем вычислительном узле. По итогам оценки качества выбирается результирующая модель-победитель (с лучшими показателями). Особенно это актуально при работе с трудоемкими моделями, пример — латентное размещение Дирихле (LDA) [6].

Метод LDA основан на вероятностной модели [7]:

$$p(w \mid d) = \sum_{t \in T} p(t \mid d) p(w \mid t), \qquad (1)$$

где t — темы (t∈T); d — документы (d∈D); w — термины (w∈W); p(t|d) и p(w|t) — соответствующие условные распределения вероятностей.

Введем векторы документов Θ_d, векторы тем φ_t и параметры α и β:

$$\begin{gathered} \Theta_d = (\Theta_{td}) \in R^{|T|}, \\ \varphi_t = (\varphi_{tw}) \in R^{|W|}, \\ \alpha \in R^{|T|}, \\ \beta \in R^{|W|} \end{gathered} \qquad (2)$$

В латентном размещении Дирихле используется дополнительное предположение, что векторы документов и тем порождаются распределениями Дирихле [8]:

$$Dir(\Theta_d; \alpha) = \frac{\tilde{A}(\alpha_0)}{\prod_t \tilde{A}(\alpha_t)} \prod_t \Theta_{td}^{\alpha_t - 1}$$

$$\alpha_t > 0, \alpha_0 = \sum_t \alpha_t,$$

$$\Theta_{td} > 0, \sum_t \Theta_{td} = 1, \quad (3)$$

$$Dir(\varphi_t; \beta) = \frac{\tilde{A}(\beta_0)}{\prod_w \tilde{A}(\beta_w)} \prod_w \varphi_{tw}^{\beta_w - 1},$$

$$\beta_w > 0, \beta_0 = \sum_w \beta_w,$$

$$\varphi_{tw} > 0, \sum_w \varphi_{twt} = 1 \quad (4)$$

где Γ(z) — гамма-функция; векторы α и β — гиперпараметры. Чем меньше значения гиперпараметров α и β, тем сильнее разрежено распределение Дирихле, и тем дальше отстоят друг от друга порождаемые им векторы. Чем меньше α_0, тем сильнее различаются документы θ_d. Чем меньше β_0, тем сильнее различаются темы φ_t.

Стандартным инструментом для обучения модели LDA является EM-алгоритм (Expectation-Maximization) [8]. EM-алгоритм сначала инициализирует модель какими-то начальными значениями, а затем повторяет шаги алгоритма до сходимости или заданного количества итераций:

- на E-шаге определяются вероятности того, что каждый документ принадлежит разным категориям, оцениваются математические ожидания (expectation) скрытых переменных (тем). Темы не могут быть измерены в явном виде, а могут быть только выведены через математические модели с использованием наблюдаемых переменных;
- на M-шаге фиксируются вероятности принадлежности и оптимизируют α и β. Переменные считаются равными своим ожиданиям, найденным на E-шаге для максимизации правдоподобия (maximization). Оценка параметров, альтернативная оценке максимума правдоподобия вычисляется следующим образом:

$$\varphi_{tw} = \frac{\sum_{d \in D} n_{dwt} + \beta_w}{\sum_{d \in D} \sum_{w \in d} n_{dwt} + \beta_0},$$

$$\theta_{td} = \frac{\sum_{w \in d} n_{dwt} + \alpha_t}{\sum_{w \in W} \sum_{t \in T} n_{dwt} + \alpha_0},$$

$$n_{dwt} = n_{dw} \, p(t \mid d, w), \quad (5)$$

где n_{dwt} — число вхождений термина w в документ d, связанных с темой t.

Вычислительная сложность алгоритма LDA равна $O(N \cdot N_T \cdot i)$, где N — число ненулевых элементов терм-документной матрицы (математическая матрица, описывающую частоту терминов, которые встречаются в коллекции документов), N_T — число тем, i — число итераций EM-алгоритма.

Алгоритм оценки качества моделей можно рассматривать как задачу принятия решения. Среди методов принятия решений наибольший интерес представляют методы на основе контроля большинства (голосования) между экспертами. Выбор стратегии голосования является одной из важных проблем для ансамблей моделей.

Представим группу экспертов в виде отдельных оценочных алгоритмов. Каждый алгоритм может допустить ошибку, но при голосовании ошибки отдельных экспертов компенсируют друг друга.

Выделяют следующие стратегии голосования [9]:

- простое голосование, при котором группа алгоритмов выбирает модель по наибольшему количеству голосов;
- алгоритм взвешенного голосования — взвешенное голосование из смеси экспертов. В этом случае голос каждого из алгоритмов T_i, входящих в группу T, имеет свой вес α_i;
- голосование по старшинству (машина покрывающих множеств). Первый алгоритм ансамбля оценивает модели и должен выбрать лучшую. Если он отказывается от выбора, то модели передаются второму алгоритму, который может выбрать наилучшую. Если этого не произошло, объект передается к третьему алгоритму и т.д., пока один из алгоритмов не примет решения.

Для получения значимых результатов оценочные алгоритмы должны существенно отличаться друг от друга.

Схема оценки качества ансамбля моделей представлена на Рис. 1. Для ансамбля выбирается n тематических моделей, которые генерирует набор из k тем каждая. Далее полученные темы оцениваются с помощью голосующего алгоритма, включающего в себя, например, набор экспертов, использующий различные методы анализа: поисковых запросов по ключевым словам в теме, на основе модели Word2vec [10], с использованием ресурса Wikipedia [11]. В качестве основной стратегии голосования выбрано простое голосование. Взвешенное голосование и голосование по старшинству предполагают ранжирование оценочных алгоритмов, что является нетривиальной задачей.

Проведено исследование на коллекции из 1070 статей по тематике информационной безопасности, полученной с IT-ресурса «Habrahabr.ru». Данные статьи написаны на русском языке с использованием англоязычных терминов и фрагментов исходного

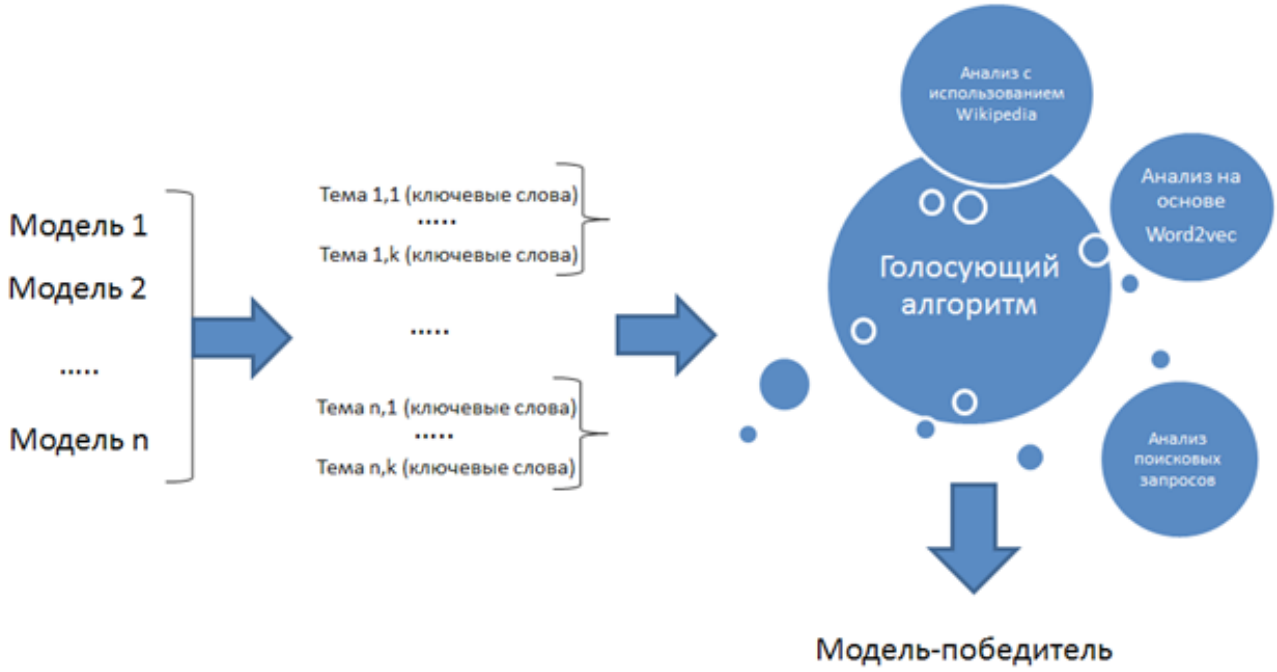

Рис. 1. Схема оценки качества ансамбля моделей

кода. В качестве ансамбля моделей использовались модели LDA с разными входными параметрами. Для статей выполнена предобработка, также каждому слову или лексеме установлен соответствующий весовой коэффициент.

В качестве оценочного алгоритма использован анализ средней конкурентности поисковых запросов по отдельным словам внутри каждой темы. Базовая формула для расчета конкурентности:

$$KEI = \frac{\text{Кол}-\text{во страниц в выдаче по запросу}}{\text{Кол}-\text{во поисковых запросов}} \quad (6)$$

Как правило, к этой формуле добавляется ряд дополнительных параметров (число внешних ссылок, количество главных страниц в запросе и др.).

В данной работе также проведено экспертное оценивание специалистами в области IT-технологий и информационной безопасности. С помощью оценочного алгоритма была выбрана одна из моделей в качестве модели-победителя.

На Рис. 2 представлен график средней конкурентности поисковых запросов по отдельным словам для первых 12 тем одной из моделей ансамбля. Незакрашенными кружочками выделены темы, однозначно определенные всеми экспертами. Эти темы имеют самую высокую среднюю конкурентность по отдельным поисковым запросам темы. Закрашенным кружочком выделен отдельный выброс, содержащий фрагменты исходного кода: echo, nul, temp, vault, key, код, 9db8b89a, 61231f25. Тема с индексом 4, обладающая наибольшей средней конкурентностью среди первых 12-ти тем, содержит следующие ключевые слова: сертификат, ssl, домен, сайт, проверка, центр, сервер, tls, браузер, https. Ключевые слова данной темы коррелируют между собой и имеют отношение к криптографическим протоколам, сопутствующим им понятиям. Чем выше средняя конкурентность запросов по отдельным словам в теме, тем тема ближе к экспертной оценке.

В данной работе проведено исследование и предложена концепция оценки качества ансамбля тематических моделей с помощью использования простого голосующего алгоритма. Вычислительный эксперимент использования оценочного алгоритма, анализирующего поисковые запросы, демонстрирует в общем случае совпадение с результатами экспертного оценивания. Проведенный эксперимент имеет погрешности: а) предобработки; б) отсутствия убедительных лингвистических обоснований LDA.

Использование дополнительных оценочных алгоритмов позволит делать более точный анализ текстовых коллекций и формировать более качественный перечень «ядерных» рубрик с помощью модели word2vec за счет учета лингвистической обоснованности моделей. Таким образом, можно сделать вывод о целесообразности дальнейшего развития алгоритмов оценки качества ансамблей тематических моделей.

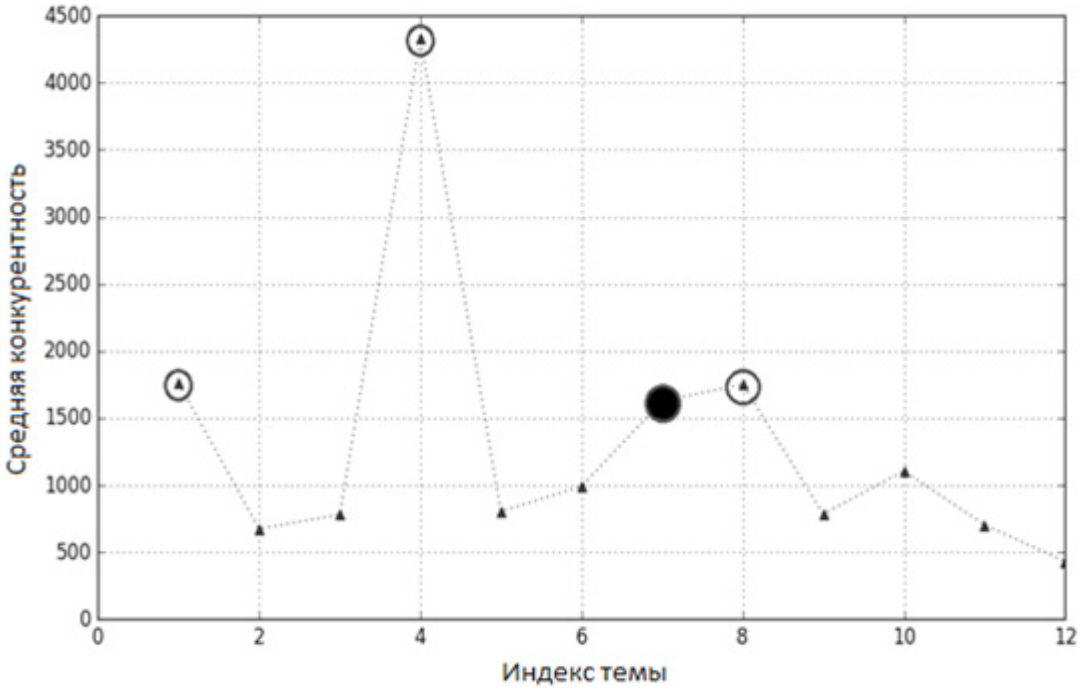

Рис. 2. График средней конкурентности поисковых запросов по ключевым словам по первым 12 темам одной из моделей ансамбля

ЛИТЕРАТУРА

1. Воронцов К.В. Вероятностное тематическое моделирование. URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (дата обращения 26.09.2018).

2. Бериков В.Б., Лбов Г.С. Современные тенденции в кластерном анализе. URL: https://docplayer.ru/26851064-Sovremennye-tendencii-v-klasternom-analize-v-b-berikov-g-s-lbov.html (дата обращения 26.09.2018).

3. Кашницкий Ю.С., Игнатов Д.И. Ансамблевый метод машинного обучения, основанный на рекомендации классификаторов // Интеллектуальные системы. Теория и приложения. 2015. Т. 19. № 4. С. 37–55.

4. Skurichina M., Duin R. P. W. Limited bagging, boosting and the random subspace method for linear classifiers // Pattern Analysis Applications. — 2002. — Pp. 121–135.

5. Журавлев Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. — М: Фазис, 2005 г. , 159 стр.

6. Blei D., Ng A., and Jordan M. Latent Dirichlet allocation // Journal of Machine Learning Research. — 2003. — vol. 3. — Pp. 993–1022.

7. Thomas Hofmann. Probabilistic latent semantic analysis // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. 1999.

8. Vorontsov K.V., Potapenko A.A. EM-like algorithms modification for probabilistic topic modeling // Machine learning and data analysis — 2013. — vol. 1, № 6. — Pp. 657–686.

9. Воронцов К.В. Лекции по алгоритмам кластеризации и многомерного шкалирования URL: http://www.cs.ru/voron/download/Clustering.pdf (дата обращения 26.09.2018).

10. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space// ICLR Workshop. — 2013.

11. David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence // In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Association for Computational Linguistics, 2010. — Pp. 100–108.

REFERENCE

1. 1. Voroncov K.V. Veroyatnostnoe tematicheskoe modelirovanie. URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (data obrashcheniya 26.09.2018).

2. Berikov V.B., Lbov G.S. Sovremennye tendencii v klasternom analize. URL: https://docplayer.ru/26851064-Sovremennye-tendencii-v-klasternom-analize-v-b-berikov-g-s-lbov.html (data obrashcheniya 26.09.2018).

3. Kashnickij Yu. S., Ignatov D. I. Ansamblevyj metod mashinnogo obucheniya, osnovannyj na rekomendacii klassifikatorov // Intellektual'nye sistemy. Teoriya i prilozheniya. 2015. T.19. № 4. S. 37-55.

4. Skurichina M., Duin R. P. W. Limited bagging, boosting and the random subspace method for linear classifiers // Pattern Analysis & Applications. — 2002. — Pp. 121–135.

5. ZHuravlev YU.I., Ryazanov V.V., Sen'ko O.V. Raspoznavanie. Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya. — M: Fazis, 2005 g. , 159 str.

6. Blei D., Ng A., and Jordan M. Latent Dirichlet allocation // Journal of Machine Learning Research. — 2003. — vol. 3. — Pp. 993–1022.

7. Thomas Hofmann. Probabilistic latent semantic analysis // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. 1999.

8. Vorontsov K.V., Potapenko A.A. EM-like algorithms modification for probabilistic topic modeling // Machine learning and data analysis — 2013. — vol. 1, № 6. — Pp. 657–686.

9. Voroncov K.V. Lekcii po algoritmam klasterizacii i mnogomernogo shkalirovaniya

URL: http://www.cs.ru/voron/download/Clustering.pdf (data obrashcheniya 26.09.2018).

10. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space// ICLR Workshop. — 2013.

11. David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence // In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Association for Computational Linguistics, 2010. — Pp. 100–108.