УДК 004.032.26

А. А. Алимаев [1]***, Н. П. Плотникова*** [1]

[1] *Мордовский государственный университет имени Н. П. Огарева, г. Саранск, Российская Федерация*

КЛАСТЕРИЗАЦИЯ НОРМАТИВНО–СПРАВОЧНОЙ ИНФОРМАЦИИ С ПОМОЩЬЮ АЛГОРИТМА LDA

Аннотация. В статье рассматривается способ кластеризации нормативно-справочной информации с помощью алгоритма Latent Dirichlet Allocation (LDA) - порождающей вероятностной модели. Постоянный рост объемов информации усложняет возможность обработки такого количества данных человеком. Задача обработки нормативно-справочной информации в автоматическом режиме является актуальной на данный момент, потому что позволит освободить человека от выполнения монотонных задач и снизить количество ошибок. Особенностью данной задачи является то, что нормативно-справочная информация в основном представляет собой текст, написанный человеком. Это означает, что в тексте могут содержаться опечатки или ошибки. Также возможна ситуация, когда близкие наименования находятся в разных категориях нормативно-справочной информации. Использование кластеризации позволит избежать данной проблемы. Рассмотрен процесс подготовки данных для кластеризации - токенизация и удаление стоп-слов. Осуществлена фильтрация полученного набора токенов, основывающаяся на частоте появления токенов в документах. Кластеризация может быть осуществлена двумя методами – используя bag-of-words или TF-IDF. Произведена оценка результатов кластеризации. Сформулирован вывод о применимости данного способа кластеризации, а также рассмотрена возможность дальнейшего усовершенствования кластеризации с использованием иерархического подхода.

Ключевые слова: LDA, Latent Dirichlet Allocation, кластеризация, вероятностная модель, тематическое моделирование.

A. A. Alimaev [1]***, N. P. Plotnikova*** [1]

[1] *Ogarev Mordovia State University, Saransk, Russian Federation*

CLUSTERIZATION OF REGULATORY - REFERENCE INFORMATION USING THE LDA ALGORITHM

Abstract. The article discusses a method for clustering normative and reference information using the Latent Dirichlet Allocation (LDA) algorithm, a generative probabilistic model. The ever-increasing volume of information makes it difficult for humans to process this amount of data. The task of processing normative and reference information in automatic mode is relevant at the moment, because it will free a person from performing monotonous tasks and reduce the number of errors. A feature of this task is that the normative and reference information is mainly a text written by a person. This means that the text may contain typographical errors or errors. A situation is also possible when similar names are in different categories of normative and reference information. Using clustering will avoid this problem. The process of preparing data for clustering is considered - tokenization and removal of stop words. The received set of tokens was filtered based on the frequency of occurrence of tokens in documents. Clustering can be done in two ways - using bag-of-words or TF-IDF. Clustering results are assessed. Conclusions on the applicability of this clustering method are obtained, and the possibility of further improving clustering using a hierarchical approach is considered.

Keywords: LDA, Latent Dirichlet Allocation, clustering, probabilistic model, topic modeling.

Целью данной работы является изучение кластеризации нормативно-справочной информации с помощью алгоритма LDA, а также оценка точности полученных кластеров.

Для осуществления поставленной цели необходимо решить следующие задачи:

- подготовка данных для обучения модели,
- фильтрация полученных данных,
- обучение модели,
- рассмотрение точности полученных кластеров,
- выводы о возможности использования данного метода кластеризации.

Введение. Тематическая модель – модель набора документов, содержащих текстовые данные, которая определяет к какой теме отнести отдельный документ коллекции. На вход

алгоритма подается набор текстовых документов, а на выходе для каждого документа получается числовой вектор, содержащий оценки принадлежности документа каждой из тем.

Тематическое моделирование – процесс построения тематической модели.

Вероятностные тематические модели позволяют документу относиться одновременно к нескольким темам с различными вероятностями. Тематическое моделирование традиционно применяется в обработке естественного языка.

Одним из алгоритмов тематического моделирования является вероятностный латентно-семантический анализ или ВЛСА [1].

В задачах на понимание естественного языка существует несколько уровней, из которых можно получать основную мысль. От слов к предложениям, к абзацам, к текстам. На уровне текста основным способом понять основную мысль является *тема*. Процесс получения заголовков из набора текстов называется *тематическим моделированием*.

Все модели основываются на одинаковых предположениях:

- каждый документ состоит из нескольких смешанных тем,
- каждая тема состоит из набора слов,
- порядок документов в коллекции не имеет значения,
- порядок слов в документе не имеет значения, документ – мешок слов,
- слова, часто встречающиеся в большинстве документов, не важны для определения тематики, их называют *стоп-словами*.

Алгоритм ВЛСА имеет ряд недостатков:

- модель нужно переобучать каждый раз, когда возникает необходимость работать с новым документом,
- количество параметров для ВЛСА растет линейно с количеством документов.

Вероятностный латентно-семантический анализ на практике используется редко. В основном применяется алгоритм Latent Dirichlet Allocation или LDA.

Латентное размещение Дирихле. LDA (Latent Dirichlet allocation) применяется для задач тематического моделирования. Под такими задачами подразумеваются задачи кластеризации текстов – таким образом, что каждый кластер содержит в себе тексты со схожими темами. LDA можно использовать для извлечения тем из естественных текстов, например в [2] его применяли для анализа газет. При обучении LDA есть возможность скрывать конфиденциальные данные, которые присутствуют в обучающей выборке [3].

В LDA каждый документ может рассматриваться как набор различных тематик. Подобный подход схож с вероятностным латентно-семантическим анализом (ВЛСА) с той разницей, что в LDA предполагается, что распределение тематик имеет в качестве априори распределения Дирихле. На практике в результате получается более корректный набор тематик.

К примеру, модель может иметь тематики, классифицируемые как «относящиеся к кошкам» и «относящиеся к собакам», тематика обладает вероятностями генерировать различные слова, такие как «мяу», «молоко» или «котёнок», которые можно было бы классифицировать как «относящиеся к кошкам», а слова, не обладающие особой значимостью (к примеру, служебные слова), будут обладать примерно равной вероятностью в различных тематиках.

Для того, чтобы применять к корпусу текстов LDA, необходимо преобразовать корпус в терм-документную матрицу. Терм-документная матрица – это матрица, которая имеет размер N×W, где N – количество документов в корпусе, а W – размер словаря корпуса, т.е. количество уникальных слов, которые встречаются в нашем корпусе. В i-й строке, j-м столбце матрицы находится число – сколько раз в i-м тексте встретилось j-е слово. LDA строит для данной терм-документной матрицы и T заранее заданного числа тем два распределения:

1. Распределение тем по текстам. (матрица размера N×T).
2. Распределение слов по темам. (матрица размера T×W).

Значения ячеек данных матриц – это вероятности. Для матрицы распределения тем по текстам это вероятность того, что данная тема содержится в данном документе. Для матрицы распределения слов по темам значения — это соответственно вероятность встретить в тексте с темой i слово j.

Подготовка данных:

1. Токенизация: разделить текст на предложения, а предложения - на слова. Привести слова в нижний регистр и удалить знаки препинания.
2. Слова, содержащие менее 3 символов, удаляются.
3. Все стоп-слова удаляются.

Для демонстрации примеров используемых далее алгоритмов будет использован язык Python 3.

Пример алгоритма подготовки данных представлен на рисунке 1.

```
def preprocess(text):
    result = []
    for token in gensim.utils.simple_preprocess(text):
        if token not in
gensim.parsing.preprocessing.STOPWORDS and len(token) > 3:
            result.append(token)
    return result
```

Рис. 1. Алгоритм подготовки данных

Алгоритм обработки данных применяется к каждому документу в корпусе. Полученный корпус обработанных документов используется далее для составления Bag Of Words – корпуса словарей.

Примеры работы алгоритма подготовки данных представлены в таблице 1.

Таблица 1.

Сравнение исходных и подготовленных данных

Исходное наименование	Наименование после обработки
Компрессор автомобильный Торнадо АС-580, 14 А, 150 PSI, 30 л/м, 12 В 1256467	компрессор автомобильный торнадо
Компрессор автомобильный AVS KS750D, двухцилиндровый, 75 л/мин, 12 В, 10 атм 2579421	компрессор автомобильный двухцилиндровый
Компрессор автомобильный Airline X3, 40 л/мин., 10 АТМ. 2615787	компрессор автомобильный airline
Компрессор автомобильный Airline X5, двухпоршневой, 50 л/мин., 10 АТМ. 2615788	компрессор автомобильный airline двухпоршневой

Из обработанных слов создается словарь, содержащий количество повторений каждого слова.

Необходимо отфильтровать токены, которые появляются:

1. менее чем в 15 документах,
2. более чем в половине документов всего корпуса,
3. после двух вышеуказанных шагов оставить только первые 100000 наиболее часто используемых токенов.

Для каждого документа создается словарь, в котором указано, сколько всего слов, и сколько раз эти слова появляются.

Сравнение обработанного документа и соответствующего словаря представлено в таблице 2.

На данном этапе есть выбор: можно обучить модель, используя корпус словарей, либо можно создать корпус, используя TF-IDF, и обучить модель на его основе.

Таблица 2.

Сравнение обработанных документов и соответствующих словарей

Обработанное наименование	Словарь
компрессор автомобильный торнадо	(0, 1), (1, 1), (2, 1)
компрессор автомобильный двухцилиндровый	(0, 1), (1, 1)
компрессор автомобильный airline	(0, 1), (1, 1), (3, 1)
компрессор автомобильный airline двухпоршневой	(0, 1), (1, 1), (3, 1)

TF-IDF – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью корпуса документов. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах корпуса.

Создание TF-IDF корпуса из корпуса словарей показано на рисунке 2.

```
tfidf = models.TfidfModel(bow_corpus)
tfidf_corpus = tfidf[bow_corpus]
```

Рис. 2. Создание TF-IDF корпуса

Сравнение обработанных документов и соответствующих им TF-IDF приведено в таблице 3.

Таблица 3.

Сравнение обработанных документов и соответствующих им TF-IDF

Обработанное наименование	TD-IDF
компрессор автомобильный торнадо	(0, 0.503), (1, 0.621), (2, 0.601)
компрессор автомобильный двухцилиндровый	(0, 0.63), (1, 0.777)
компрессор автомобильный airline	(0, 0.498), (1, 0.615), (3, 0.611)
компрессор автомобильный airline двухпоршневой	(0, 0.498), (1, 0.615), (3, 0.611)

При обучении в модель можно передать как корпус словарей, так и TF-IDF корпус. Обучение модели LDA с помощью gensim.models.LdaMulticore [4] представлено на рисунке 3.

```
lda_model = gensim.models.LdaMulticore(bow_corpus, \
                                       num_topics=NUMBER_OF_TOPICS, \
                                       id2word=dictionary, \
                                       passes=2, \
                                       workers=12)
```

Рис. 3. Обучение модели с использованием корпуса словарей

Используя обученную модель, кластеризуем наш набор данных. Пример кластеризации данных представлен на рисунке 4.

```
classes = [ [] for i in range(NUMBER_OF_TOPICS) ]
for document in documents:
    bow_vector = dictionary.doc2bow(preprocess(document))
    result = lda_model[bow_vector]
    if not result:
        continue
    prob = 0
    for var in result:
        if var[1] > prob:
            i, prob = var
    classes[i].append(document)
```

Рис. 4. Кластеризация набора данных

Результатом кластеризации является вектор, содержащий вероятности того, что текущий документ относится к некоторому кластеру i. Текущий документ необходимо отнести к кластеру, для которого указана наибольшая вероятность.

Пример полученных кластеров представлен в таблице 4.

Таблица 4.

Сравнение полученных кластеров с изначальными классами

Наименование	Номер кластера	Наименование изначального класса
Резинки Autovirazh, 24"/60 см, для каркасной щетки с графитом, набор 2 шт. 1187188	1	Авто и мото
Крышка расширительного бачка ВАЗ 2108-2170,Волга, Газель 3120385	1	Авто и мото
Комплект ксеноновых ламп TORSO H8, для блоков DC, 12 В, 4300 К, 2 шт. 1059328	1	Авто и мото
Термосумка с элементом холода "Охладись", 16 л 2531207	1	Авто и мото
Брелок-мультитул "Герою", 3,7 х 3,7 см 3821248	2	Инструменты
Антенна внешняя, телескоп 50-120 см, врезная, кабель 1.7 м, универсальная 3099470	2	Авто и мото
Подогреватель двигателя универсальный №2 "Старт ТУРБО", с центробежным насосом, 2 Кв 2683241	2	Авто и мото
Открытка со значком "Талисман на удачу", 3 х 5 см 1687455	2	Бижутерия
Жетон "Лучший моряк", 4 х 2,5 см 2464087	2	Бижутерия
Щетка стеклоочистителя Nova Bright, каркасная, 330 мм/13*, тефлоновое покрытие 3393739	3	Авто и мото

Полученные таким образом кластеры содержат в себе элементы разных изначальных классов. Итоговая точность кластеризации оказалась достаточно низкой, кластеры не пригодны для дальнейшего использования.

В качестве дальнейшего исследования стоит рассмотреть иерархический подход, описанный в [5]. Иерархический подход может улучшить точность кластеризации и, как следствие, качество полученных кластеров.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Hofmann, Thomas. "Probabilistic latent semantic analysis." *arXiv preprint arXiv:1301.6705* (2013).

2. Marjanen J. et al. Topic modelling discourse dynamics in historical newspapers //arXiv preprint arXiv:2011.10428. – 2020.

3. Zhao F. et al. Latent Dirichlet Allocation Model Training With Differential Privacy //IEEE Transactions on Information Forensics and Security. – 2020. – Т. 16. – С. 1290-1305.

4. Radim Řehůřek. Optimized Latent Dirichlet Allocation (LDA) [Электронный ресурс]. – Режим доступа: https://radimrehurek.com/gensim/models/ldamodel.html, свободный. – (дата обращения: 02.02.2021).

5. Rieger J. et al. Improving Reliability of Latent Dirichlet Allocation by Assessing Its Stability Using Clustering Techniques on Replicated Runs //arXiv preprint arXiv:2003.04980. – 2020.

REFERENCES

1. Hofmann, Thomas. "Probabilistic latent semantic analysis." *arXiv preprint arXiv:1301.6705* (2013).

2. Marjanen J. et al. Topic modelling discourse dynamics in historical newspapers //arXiv preprint arXiv:2011.10428. – 2020.

3. Zhao F. et al. Latent Dirichlet Allocation Model Training With Differential Privacy //IEEE Transactions on Information Forensics and Security. – 2020. – T. 16. – P. 1290-1305.

4. Radim Řehůřek. Optimized Latent Dirichlet Allocation (LDA) [Electronic resource]. – Mode of access: https://radimrehurek.com/gensim/models/ldamodel.html, свободный. – (date of access: 02.02.2021).

5. Rieger J. et al. Improving Reliability of Latent Dirichlet Allocation by Assessing Its Stability Using Clustering Techniques on Replicated Runs //arXiv preprint arXiv:2003.04980. – 2020.

Информация об авторах

Андрей Алексеевич Алимаев – студент кафедры «Автоматизированные системы обработки информации и управления», Мордовский государственный университет имени Н. П. Огарева, г. Саранск, e-mail: alimaef@yandex.ru

Наталья Павловна Плотникова – к. т. н., кафедра «Автоматизированные системы обработки информации и управления», Мордовский государственный университет имени Н. П. Огарева, г. Саранск, e-mail: linsierra@yandex.ru

Authors

Andrey Alekseevich Alimaev – student of the Subdepartment "Automated information processing and control systems", Mordovia State University named after N.P. Ogarev, Saransk, e-mail: alimaef@yandex.ru

Natalya Pavlovna Plotnikova – Ph.D., Subdepartment of Automated Information Processing and Control Systems, Mordovia State University named after N.P. Ogarev, Saransk, e-mail: linsierra@yandex.ru

Для цитирования

Алимаев А.А., Плотникова Н.П. Кластеризация нормативно-справочной информации с помощью алгоритма LDA // // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2021. – №2(10). – С. 46-51 – DOI: 10.26731/2658-3704.2021.2(10).46-51 – Режим доступа: https://ismm.irgups.ru/toma/210-2021, свободный.. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 29.04.2021)

For citations

Alimaev A.A., Plotnikova N.P. Clustering normative - reference information using the LDA algorithm // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2021. No. 2. P. 46-51. DOI: 10.26731/2658-3704.2021.2(10). 46-51 [Accessed 29/04/21]