УДК 519.67+004.942

# Машинные методы тематического моделирования коллекции учебных текстов на естественном языке

*Н.С. Бабкина[1], А.Б. Нугуманова[2], Н.М. Оскорбин[1], О.Н. Половикова[1], Л.Л. Смолякова[1]*

[1]Алтайский государственный университет (Барнаул, Россия)
[2]Восточно-Казахстанский государственный университет им. С. Аманжолова (Усть-Каменогорск, Казахстан)

# Computer Methods for Thematic Modeling of Textbooks Collection in Natural Language

*N.S. Babkina[1], A.B. Nugumanova[2], N.M. Oskorbin[1], O.N. Polovikova[1], L.L. Smolyakova[1]*

[1]Altai State University (Barnaul, Russia)
[2] Sarsen Amanzholov East Kazakhstan State University (Ust-Kamenogorsk, Kazakhstan)

Представлены результаты разработки информационных технологий машинных методов анализа текстовых данных и оценки погрешностей классификации на этапах тематического моделирования. Исследование проводится на примере задачи обработки текстовых данных коллекции выпускных квалификационных работ кафедры информатики, которые подготовлены и защищены в последние годы студентами ФМиИТ АлтГУ.

Основные результаты работы состоят в следующем: выбраны актуальные направления использования машинных методов и задач тематического моделирования в учебном процессе; проведено обоснование общего алгоритма решения задачи тематического анализа коллекций учебных материалов; разработаны информационные технологии тематического моделирования и получены оценки погрешностей анализа на совокупности тестовых документов. Показано, что использование машинных методов тематического моделирования и информационных технологий их поддержки возможно как в учебном процессе, так и при подготовке учебно-методических материалов.

***Ключевые слова:*** анализ текстовых данных, тематическое моделирование, тестирование информационных технологий.

The paper presents the development results of computer methods for analyzing text data and assessing classification inaccuracies at the stages of thematic modeling. This study uses as an example the task to process textual data of a collection of graduate qualification works prepared and defended by students of Altai State University, Faculty of Mathematics and IT in recent years.

The main results obtained in the paper are listed as follows. Relevant application areas and directions for computer methods and thematic modeling in the educational process are identified. Justification of the general algorithm for solving the problem of the thematic analysis of collections of educational materials is carried out. Information technologies for thematic modeling are developed, and estimation of analysis errors on a set of test documents is obtained. It is shown that computer-based methods of thematic modeling and information technology to support them can be used both in the educational process and in the development of educational and methodological documents.

***Key words***: text data analysis, thematic modeling, information technology testing.

**DOI 10.14258/izvasu(2020)1-10**

**Введение.** В статье рассматривается тематическое моделирование как одно из эффективных направлений в области автоматизированной обработки текстов на естественном языке [1]. Актуальность проблем автоматизированного анализа текстов обоснована как отечественными учеными [2–4], так и в трудах зарубежных авторов [5, 6].

Тематические модели — это модели со скрытыми переменными, для выявления которых лучше всего подходит нечеткая кластеризация, при которой любое слово или документ с некоторой вероятностью относится к нескольким темам [7, 8]. В работе [9] выделены проблемные области тематического моделирования текстовых материалов на естественном языке, которые следует учитывать при анализе и интерпретации учебно-методических материалов.

В данной работе проведено исследование машинных методов тематического моделирования коллекций учебных текстовых документов, разработаны алгоритм поиска документов коллекции по заданным темам и информационные технологии его поддержки.

**Постановка задачи и схема алгоритма анализа коллекций текстовых документов**

Рассматриваем коллекцию текстовых документов, которые следует классифицировать по заданным темам. Начальные этапы подготовки файлов текстовых документов и списков ключевых слов, включая технологии выполнения лемматизации текстов, представлены в работах [8, 10].

Заметим, что лемматизацию рекомендуют проводить для всех текстов коллекции. Однако можно воспользоваться операцией противоположного направления, а именно расширить состав ключевых слов, который обычно задан в нормальной форме. Тогда поиск ключевых слов и оценку их суммарной значимости можно проводить по исходному тексту исследуемых документов.

Далее по этапам классификации производят оценку значения классифицирующей функции принадлежности или непринадлежности каждого из исследуемых документов коллекции к выделенным темам. В качестве простой функции можно принять число ключевых слов в рассматриваемом документе по каждой теме и определить пороговые критерии принадлежности документа к теме.

Возможны следующие результаты классификации [10]:

1) «Прав»: документ («Свой») правильно определился в свою рубрику.

2) «Чуж»: действительно «Чужой» документ определился как «Чужой».

3) «Ошиб»: документ определился не в свою рубрику.

4) «Св_чуж»: «Свой» документ ошибочно определился как «Чужой».

5) «Чуж_св»: «Чужой» документ ошибочно попал в какую-то рубрику, т.е. ошибочно определился как «Свой».

Исходы 1 и 2 соответствуют правильной работе классификатора, исходы 3, 4, 5 являются ошибками классификации.

Схема алгоритма классификации текстовых документов на русском языке по шагам запишется в следующем виде:

**Шаг 1.** Формируем коллекцию текстовых документов, из которой необходимо найти документы по заданным темам.

**Шаг 2.** Для заданных базовых тем с помощью экспертов составляем список ключевых слов в нормальной форме.

**Шаг 3.** Проводим операцию, обратную лемматизации, в автоматизированном режиме, т.е. расширяем список ключевых слов для характеристики заданных базовых тем.

**Шаг 4.** Выполняем в диалоговом режиме предварительную обработку документов коллекции, включая их очистку, уменьшение размерности, выделение информативных разделов. Формируем сравниваемый документ в текстовом формате.

**Шаг 5.** Подсчитываем значения классифицирующей функции принадлежности или непринадлежности каждого из исследуемых документов коллекции к выделенным темам и выделяем искомые документы коллекции, которые относятся к заданным темам.

Анализ текстовых документов можно проводить последовательно или в процессе классификации уточнять решения предыдущих этапов. Это касается как составов коллекции документов, ключевых слов, так и пороговых значений классифицирующей функции.

**Информационная технология анализа коллекции текстовых документов**

Используемые коллекции текстовых документов предварительно обрабатываются и преобразовываются в файлы (обычно с расширением .txt). Дополнительной очистки файлов не требуется, т.к. при обработке исключаются таблицы, формулы и другие объекты. Статистический анализ текстов осуществляется с использованием приложения Wordstat, которое позволяет подсчитать количество вхождений каждого слова в отдельности.

Рассмотрим шаг 5 алгоритма. Обозначим индексом $j$ номер документа в коллекции ($j = 1,...,N$); $T$ — множество базовых тем $t$ классификации, $t \in T$; $I_t$ — множество расширенных индексов для каждой темы; $w_i^t$ — вес ключевого слова $i \in I_t$ для темы $t \in T$. Считаем, что веса ключевых слов неотрицательны и в сумме равны единице [8].

Пусть для документа $j$ с числом слов $M_j$, очищенного на шаге 4 алгоритма, посчитан общий вес $V_j^t$ ключевых слов по следующей формуле:

$$V_j^t = \sum_{i \in I_t} w_i^t \cdot m_{i,j}^t, j = 1,...,N, \quad (1)$$

где $m_{i,j}^t$ — число ключевых слов $i$ по теме $t$ в $j$ документе.

Введем с учетом размера текста анализируемого документа следующее нормирование функции (1):

$$\overline{V}_j^t = V_j^t \cdot \frac{M}{M_j}, j = 1,...,N;\ t \in T, \quad (2)$$

где $M$ — размер эталонного документа.

Пусть $V_j^P$ пороговое значение функции $\overline{V}_j^t$. Тогда правило классификации имеет следующую логическую формулу:

*Если* $\overline{V}_j^t \geq V_j^P$, *то документ «свой» для темы* $t \in T$, *иначе «чужой».* (3)

Идентификацию параметров формул (1), (2), (3) следует проводить по обучающей выборке, а качество классификации проверять по контрольной выборке документов, как это принято в системах искусственного интеллекта, в частности — в нейросетевых классификаторах.

**Тестирование информационных технологий тематической классификации**

Рассмотрим методику и результаты тестирования алгоритмических и программных средств на примере тематического моделирования выпускных квалификационных работ (ВКР). Коллекция включает 15 ВКР бакалавров и магистров, которые защищены на кафедре информатики АлтГУ в период 2016–2018 гг. В соответствии с шагом 2 алгоритма выделим базовые темы и для каждой из них сформируем начальный список ключевых слов (табл. 1).

*Таблица 1*

Базовые темы ВКР для тестирования алгоритма классификации

| Номер темы | Наименование темы | Ключевые слова |
|---|---|---|
| Тема 1 | Машинное обучение и анализ данных | Метод, обучение, машина, технология, статистика, данные, алгоритм, нейросеть, искусственный, интеллект |
| Тема 2 | Информационные технологии | Система, информация, хранение, поиск, база, обработка, проектирование, классификация, автоматизация |
| Тема 3 | Комплексы программ | Архитектура, модель, код, язык, схема, структура, программирование, тестирование |

В соответствии с шагом 3 алгоритма проводим операцию, обратную лемматизации, в автоматизированном режиме. После указанной обработки данных таблицы 1 получаем по каждой теме расширенный в 3–4 раза список ключевых слов и их преобразований. Значения классифицирующей функции при базовом размере текста ВКР, равным 1000, представлены в таблице 2, в том числе визуально оцененные пороговые значения $V_j^P$ в формуле (3).

*Таблица 2*

Значения классифицирующей функции и классификация ВКР

| Номер ВКР | Значение функции (2) | | | Экспертные оценки тем ВКР | | | |
|---|---|---|---|---|---|---|---|
| | тема 1 | тема 2 | тема 3 | тема 1 | тема 2 | тема 3 | чужой |
| 1 | 42,76 | 17,18 | 7,31 | 1 | 0 | 0 | 0 |
| 2 | 108,68 | 55,61 | 50,03 | 1 | 1 | 1 | 0 |
| 3 | 27,17 | 20,82 | 14,47 | 0 | 0 | 0 | 1 |
| 4 | 75,67 | 30,73 | 24,89 | 1 | 0 | 0 | 0 |
| 5 | 60,94 | 24,63 | 29,46 | 1 | 0 | 0 | 0 |
| 6 | 32,76 | 16,25 | 6,60 | 1 | 0 | 0 | 0 |
| 7 | 5,08 | 21,08 | 8,38 | 0 | 0 | 0 | 1 |
| 8 | 9,65 | 5,84 | 13,97 | 0 | 0 | 0 | 0 |
| 9 | 62,21 | 27,17 | 30,98 | 1 | 0 | 1 | 0 |
| 10 | 16,51 | 26,92 | 37,07 | 0 | 0 | 1 | 0 |
| 11 | 11,93 | 15,24 | 13,97 | 0 | 0 | 0 | 1 |
| 12 | 93,45 | 44,18 | 66,02 | 1 | 0 | 1 | 0 |
| 13 | 30,98 | 68,31 | 9,14 | 1 | 1 | 0 | 0 |
| 14 | 6,60 | 26,92 | 5,59 | 0 | 0 | 0 | 1 |
| 15 | 49,26 | 48,76 | 83,04 | 1 | 0 | 1 | 0 |
| $V_j^P$ | 30 | 55 | 30 | | | | |

Результаты классификации ВКР представлены в таблице 3 по выделенным 5 классам, анализом данных таблицы 2.

*Таблица 3*

Результаты тематической классификации ВКР

| Классы оценок | Результаты классификации, оценки точности и полноты | Число ВКР | Итог, % |
|---|---|---|---|
| 1 | «Прав»: документ («Свой») определен правильно | 8 | 53,3% |
| 2 | «Чуж»: документ («Чужой») определен правильно | 3 | 20,0% |
| 3 | «Ошиб»: документ определился не в свою рубрику | 1 | 6,7% |
| 4 | «Св_чуж»: документ «Свой» определился как «Чужой» | 2 | 13,3% |
| 5 | «Чуж_св»: документ «Чужой» определился как «Свой» | 1 | 6,7% |
| Значение показателя точности (пп.1+2) / (пп. 1+2+3+5) | | 11/13 | 84,6% |
| Значение показателя полноты (пп.1+2) / (пп. 1+2+3+4) | | 11/14 | 78,6% |

Итоговые результаты классификации, полученные оценки точности и полноты решений показывают работоспособность предложенных в данной работе метода и информационных технологий тематического моделирования.

**Заключение.** В работе представлены результаты разработки информационных технологий машинных методов анализа текстовых данных и оценки погрешностей классификации на этапах тематического моделирования. Показано, что использование машинных методов *тематического моделирования и информационных технологий их поддержки возможно как в учебном процессе, так и при подготовке учебно-методических документов.*

Выделены актуальные направления исследований машинных методов анализа учебных текстов на естественном языке:

1. Формирование ключевых слов информационно-поисковых систем с использованием электронного онлайн-корпуса русских текстов.

2. Тематическое моделирование с использованием гистограмм распределения по тексту ключевых слов.

3. Разработка информационных технологий тематического моделирования многопрофильных текстов на естественном языке.

4. Информационные технологии тематического моделирования текстов на естественном языке с использованием ключевых фраз.

5. Применение нейросетевых технологий структурно-параметрической идентификации классифицирующих функций в тематическом моделировании.

## Библиографический список

1. Ерланова Р.Е. и др. Тематическое моделирование текстовых учебных материалов по информатике средствами языка R // Известия АлтГУ. 2018. № 4(102). DOI: 10.14258/izvasu(2018)4-12.

2. Махина Е. Д., Пальчунов Д. Е. Программная система для определения речевых действий в текстах естественного языка // Вестник НГУ. Серия : Информационные технологии. 2018. Т. 16. № 4. DOI: 10.25205/1818-7900-2018-16-4-95-106.

3. Коляда А.С. и др. Применение латентного размещения Дирихле для анализа публикаций из наукометрических баз данных // Pratsi. 2014. № 1 (43).

4. Леонова Ю. В., Федотов А.М. Извлечение знаний и фактов из текстов диссертаций и авторефератов // Системный анализ и информационные технологии : Тр. V Меж. конф. Красноярск, 2013. Т. 1.

5. Dezhao S., Schilder F., Smiley C., Brew C., Zielund T., Bretz H., Martin R., Dale C., Pomerville S., Duprey J., Miller T., and Harrison J. TR Discover: a natural language interface for querying and analyzing interlinked datasets. Proc. 14th Intern. conf. on the Semantic Web: ISWC 2015. Springer Intern. Publ., 2015.

6. Chen F. Topic Modeling of Document Metadata for Visualizing Collaborations over Time / P. Chiu, S. Lim // Proc. of the Int. Conf. on Intelligent User Interfaces (IUI), 2016. DOI: 10.1145/2856767.2856787.

7. Бабкина Н.С., Смолякова Л.Л. Проблемы реализации тематического моделирования в учебном процессе : сб. научн. ст. Межд. конф «Ломоносовские чтения на Алтае: фундаментальные проблемы науки и техники». 2018. URL: https://sites.google.com/ site/lomchten/.

8. Федотов А.М., Прозоров О.В., Федотова О.А., Бапанов А.А. О подходе к тематической классификации документов // Вестник НГУ. Серия : Информационные технологии. 2017. Т. 15. № 1.

9. Половикова О.Н., Бабкина Н.С., Смолякова Л.Л. Анализ проблематики тематического моделирования // МАК : «Математики — Алтайскому краю» : сб. трудов Всерос. конф. по математике с междунар. участием. Барнаул, 2018.

10. Леонова Ю.В., Федотов А.М., Федотова О.А. О подходе к классификации авторефератов диссертаций по темам // Вестник НГУ. Серия : Информационные технологии. 2017. Т. 15. № 1. DOI: 10.25205/1818-7900-2017-15-1-47-58.