УДК 681.3

# ВЕРОЯТНОСТНОЕ ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В КОЛЛЕКЦИЯХ ДОКУМЕНТОВ

***© 2022 Я. Е. Львович, Ю. П. Преображенский, Е. Ружицкий***

*Воронежский государственный технический университет (Воронеж, Россия)*
*Воронежский институт высоких технологий (Воронеж, Россия)*
*Панъевропейский университет (Братислава, Словакия)*

*В статье обсуждаются некоторые вопросы, связанные с тематическим моделированием в коллекциях документов. Данный подход активным образом применяется в ходе анализа различной документации. Показано, что одним из возможных методов в моделировании может быть вероятностный подход. Дан анализ основных этапов осуществляемого моделирования и возможностей используемых алгоритмов.*

*Ключевые слова: технология, тематическое моделирование, документ.*

В качестве задачи, стоящей перед вероятностным тематическим моделированием (далее ВТМ), которое использует описание текста при помощи признаков, можно выделить задачу, которая рассматривается, к примеру, в [1].

Проведем рассмотрение главных определений и поставим стандартную задачу перед ВТМ, в соответствии с [2]. ВТМ является моделью большого количества документации, которая описывает все их темы, дискретно распределяя их по множеству признаков, а любой из документов – дискретно распределяя по множеству тем. рассмотрим детальнее, что собой представляет термин «тема». Существующий контекст определяет тему как определенное количество признаков, характеризуют одну предметную область.

Это определение показывает, что один документ может иметь принадлежность к разным темам. Проведение описания документа, дискретно распределяя его по множеству тем, называют т. н. «мягкой» кластеризацией. Также, признак способен иметь принадлежность больше, чем к 2-м двум темам. Зачастую это связывают с появлением полисемии – ситуация, при которой слово может иметь различные значения.

Представим, что есть D, которое является множеством документации, а также W, являющееся множеством признаков, которые имеет коллекция. Пускай каждый из документов $d \in D$, является последовательностью признаков множества W. В этом случае, задачу ВТМ можно сформулировать таким способом. Нужно найти:

1) Q – является множеством тем у документации;

2) $p(\omega|q)$ – является распределением на W, которое задает тему $q \in Q$;

3) $p(\omega|q)$ – является профилем темы, относительно документа $q \in D$.

Стоит заметить, что ВТМ можно ассоциировать с процессом, который порождает документацию, которая входит в коллекцию, имеющий название генеративного процесса.

Избранный процесс, будет определять алгоритм, по которому получается документ с существующими распределениями 2, а также 3, и будет определять обратный процесс – оценку модельных параметров, используя заданную коллекцию.

Довольно удобно задача ВТМ рассматривается в матрице. Чтобы это осуществить, по коллекции D выставляется соответствующая матрица $F = (f_{ij}) w \| x \| D |, f_{ij} =$

Львович Яков Евсеевич – Воронежский государственный технический университет, профессор, e-mail: office@vivt.ru.
Преображенский Юрий Петрович – Воронежский институт высоких технологий, профессор, e-mail: petrovich@vivt.ru.
Ружицкий Евгений – Панъевропейский университет, канд. техн. наук, доцент, e-mail: rush_evg_br53@yandex.ru.

$p(\omega_i|d_j)$, с $i=1,...,|W|, j=1,...,|D|$.. Необходимо определить представление F как произведение 2-х матриц стохастического типа:

$$F = \Phi T, \quad (1)$$

здесь $\Phi = (\phi_{ij})|w| \times |T|$ – является матрицей признаков тем, и $T = (t_{ij})|T| \times |D|$ – является матрицей тем документации.

Практически, эту задачу можно решить при помощи нескольких методов. К примеру, можно сингулярно разложить матрицу F, а затем выбрать |Q| основных компонент. Но те из матриц, которые получаются при таком разложении, нельзя принимать как стохастические, и это не позволяет формально решить задачу вероятностного тематического моделирования, но иногда этот способ полезен при прочих ситуациях. Обычно, используют способ, который называют максимальным правдоподобием [3], вместе с EM-алгоритмом или же сэмплированием Гиббса.

Методики BTM, успешно используют при осуществлении поиска информации. Основной задачей, стоящей перед поиском, является определение документов, которые будут похожи на установленный образец, т. е. на запрос поиска. Здесь запрос на поиск изменяется на векторную форму (применяемую данными моделями [4, 5]). Далее рассчитывается степень близости запроса и разложения всей документации, которая сводится к тому, что рассчитывается косинусная мера близости, полученные значения сортируются и извлекаются документы с самым большим весом. Конечно же, при использовании разных модельных параметров (например, число тем в поиске), итоги этого процесса могут быть различными.

Необходимо заметить, что эти модели можно программно реализовывать поразному, отличие будет только в применяемых средствах разработки.

Наиболее простым инструментарием вероятностного тематического моделирования, можно считать использование вероятностного латентного семантического анализа (в английской транскрипции – Probabilistic Latent Semantic Analysis), далее – ВЛСА [1]. Данная методика была разработана в 1999 г. [6]. В ее основе лежит математическая статистика, и у нее существует ряд модификаций. Проведем описание ВЛСА при помощи генеративного процесса.

1) Выберем документ d, в соответствии с распределением P(d).

2) Выберем тему $\theta$, основываясь на распределении тем, с условием фиксирования документа: $P(q=\theta|d)$.

3) Выберем признак v, по распределению $P(\omega=\upsilon|q=\theta)$.

После этого выстроится такое выражение как:

$$P(d,\omega) = P(d)\sum_{q\in Q} P(\omega|q)P(q|d), \quad (2)$$

Оно и даст определение вероятностной модели. Заметим, что формулу 1.11, можно переписать как [12]:

$$P(d,\omega) = \sum_{q\in Q} P(q)P(\omega|q)P(d|q). \quad (3)$$

Это осуществляется при помощи существующего соотношения - P(A)P(B|A) = P(B)P(A|B). Последнее выражение можно назвать как симметричное представление модели. Обычно, чтобы оценить скрытые переменные темы, применяют максимальное правдоподобие вместе с EM-алгоритмом [8]. Но при этом подходе можно явно увидеть его определенные недостатки [1].

• Количество модельных параметров, будет линейно зависеть от того, сколько документов содержит коллекция;

• Отсутствие возможности определения документа, который не содержит коллекция. Невозможно вычислить вероятность документа, отсутствующего в коллекции;

• Отсутствуют закономерности, когда генерируются документы чтобы соединить темы.

Какие-то недостатки вероятностного латентного семантического анализа, устраняются моделью, которую называют скрытое размещение Дирихле (в английском переводе Latent Dirichlet Allocation, или LDA), которая была разработана в 2003-м г. [8]. Процесс генерации в модели показан далее [1].

1) Производится выбор распределения относительно документа d, по темам $q_d$.

2) По всем признакам в d:

(a) осуществляется выбор темы $\phi_t \in q_d$;

(b) осуществляется выбор значения признака в распределении признаков по избранной теме $\phi_t$ .

Стандартная постановка задачи, заранее залает число тем (K). Принято, что параметры $\theta_d$ b $\phi_t$ , будут с распределением Дирихле. Данное предположение является основой модели, оно позволяет намного легче оценивать параметры [1]. Рассмотрим выражение, которое описывает как в общем, распределяются рассматриваемые переменные и параметры [1]:

$$P(\phi_{1:K}, \theta_{1:D}, \omega_{1:D}) = \prod_{i=1}^{K} P(\phi_i) \prod_{d=1}^{D} P(\theta_d) \left( \prod_{n=1}^{N} P(q_{dn}|\theta_d) P(\omega_{dn}|\phi 1:K, q_{dn}) \right), \quad (4)$$

Здесь, вместе с рассмотренными ранее параметрами, есть D, которое является числом документов, содержащихся в коллекции. Чтобы рассчитать оптимальные значения модельных параметров, рассчитывают распределение апостериорного типа [9].

Прочие модели, осуществляющие тематическое моделирование

В данный момент методики, которые описывают тексты по признакам, стремительно развиваются. Поэтому можно наблюдать появление новых моделей, содержащих коллекции документации, и моделей, которые дают возможность добавления новой документации в коллекции, не производя перерасчет модельных параметров. Такие модели называют онлайн-алгоритмами [1]. Также есть ряд моделей, которые могут определить число тем, и т. н. методики робастного типа.

Еще одно направление, представляет собой модернизацию рабочих алгоритмов, которые используются в уже существующих моделях.

Стоит отметить, что есть и другие подходы к созданию тематических моделей. Вместе с представленными методиками, применяют методику, предполагающую максимальное правдоподобие [8], методику моментов, а также алгоритм, который использует неотрицательную матрицу факторизации (или NMF). Также распространены методики, при которых используется сингулярное разложение матриц (или SVD) [1].

## СПИСОК ИСТОЧНИКОВ


1. Айвенс К. Администрирование Microsoft Windows Server 2003: учебное пособие / К. Айвенс. – 3-е изд. – Москва: Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 2021. – 486 с. – ISBN 978-5-4497-0853-3. – Текст: электронный // Цифровой образовательный ресурс IPR SMART: [сайт]. – Доступно по: https://www.iprbookshop.ru/101986.html.

2. Кузнецова И. В. Конфиденциальное делопроизводство и защищенный электронный документооборот: учебное пособие для бакалавров / И. В. Кузнецова, Г. А. Хачатрян. – Москва: Ай Пи Ар Медиа, 2020. – 192 с. – ISBN 978-5-4497-0588-4. – Текст: электронный // Цифровой образовательный ресурс IPR SMART: [сайт]. – Доступно по: https://www.iprbookshop.ru/97083.html.

3. Борзова А. С. Особенности построения системы принятия решений при многовариантной оптимизации структуры цифрового управления логистическим процессом в организационной системе на основе имитационного моделирования / А. С. Борзова, В. В. Муха // Моделирование, оптимизация и информационные технологии. – 2021. – Т. 9. – № 3 (34). – С. 15-16.

4. Львович Я. Е. Об управлении работой распределенных энергетических систем / Я. Е. Львович, И. Я. Львович, А. П. Преображенский, Ю. А. Клименко, О. Н. Чопоров // XIII Всероссийское совещание по проблемам управления ВСПУ-2019. Сборник трудов XIII Всероссийского совещания по проблемам управления ВСПУ-2019. Институт проблем управления им. В. А. Трапезникова РАН. – 2019. – С. 2473-2478.

5. Клименко Ю. А. Анализ структуры распределённых и изолированных энергетических систем на основе применения объектов малой энергетики / Ю. А. Клименко, А. П. Преображенский // Грозненский естественнонаучный бюллетень. – 2019. – Т. 4. – № 2 (16). – С. 99-104.

6. Машков В. Г. Предварительная оценка вероятности принятия правильного решения в автоматизированных системах управления / В. Г. Машков, В. А. Малышев,

Ю. В. Никитенко // Моделирование, оптимизация и информационные технологии. – 2021. – Т. 9. – № 3 (34). – С. 12-13.

7. Lvovich I. Optimization of the subsystem for the movement of electronic documents in educational organization / I. Lvovich, A. Preobrazhenskiy, Y. Preobrazhenskiy, Y. Lvovich, O. Choporov // Proceedings – 2021 1st International Conference on Technology Enhanced Learning in Higher Education, TELE 2021. – 1. – 2021. – С. 328-332.

8. Печенкин В.В. Моделирование динамики серверной нагрузки стохастическими сетями Петри с приоритетами (на примере системы видеоконференцсвязи) / В. В. Печенкин, А. Т. Х. Аль-Хазраджи, С. С. Гельбух // Моделирование, оптимизация и информационные технологии. – 2021. – Т. 9. – № 1 (32). – С. 10-11.

9. Преображенский Ю. П. Некоторые проблемы автоматизации процессов / Ю. П. Преображенский // Техника и технологии: пути инновационного развития. Сборник научных трудов 8-й Международной научно-практической конференции. Юго-Западный государственный университет. – 2019. – С. 62-64.

## PROBABILISTIC THEMATIC MODELING IN DOCUMENT COLLECTIONS

***© 2022 Ya. E. Lvovich, Yu. P. Preobrazhensky, E. Ruzhitsky***

*Voronezh State Technical University (Voronezh, Russia)*
*Voronezh Institute of High Technologies (Voronezh, Russia)*
*Pan-European University (Bratislava, Slovakia)*

*The paper discusses some issues related to topic modeling in document collections. This approach is actively used in the analysis of various documentation. It is shown that one of the possible methods in modeling can be a probabilistic approach. The analysis of the main stages of the ongoing modeling and the capabilities of the algorithms used is given.*

*Keywords: technology, thematic modeling, document.*