Основы и тенденции R в поиске информации



Том. ХХ, № ХХ (2017) 1–154

c 2017 г.



DOI: 10.1561/XXXXXXXXXX

**Приложения тематических моделей**

Джордан Бойд-Грабер

Кафедра компьютерных наук, umiacs, Language Science

Университет Мэриленда 1

jbg@umiacs.umd.edu

|  |  |
| --- | --- |
| Юнин Ху | Дэвид Мимно |
| Google, Inc. 2 | Информационная наука |
| ynhu@google.com | Cornell University |
|  | mimno@cornell.edu |

1 Работа завершена во время учебы в Университете Колорадо.

2 Работа, выполненная в Yahoo!

**Содержание**

|  |  |  |
| --- | --- | --- |
| **1 Что и зачем тематические модели** | | **2** |
| 1.1 | Расскажи мне о своем стоге сена. . . . . . . . . . . . . . . . | 2 |
| 1,2 | Что такое тематическая модель. . . . . . . . . . . . . . . . . . . | 4 |
| 1,3 | Фонды . . . . . . . . . . . . . . . . . . . . . . . . . | 5 |
| 1,4 | Скрытое распределение Дирихле. . . . . . . . . . . . . . . . . | 11 |
| 1,5 | Вывод . . . . . . . . . . . . . . . . . . . . . . . . . . . | 13 |
| 1,6 | Остальная часть этого обзора. . . . . . . . . . . . . . . . . . . | 20 |

|  |  |  |
| --- | --- | --- |
| **2 Специализированный поиск информации** | | **21** |
| 2.1 | Моделирование языка документов. . . . . . . . . . . . . . . | 23 |
| 2.2 | Применение тематических моделей к языковым моделям документа. . | 25 |
| 2.3 | Расширение запроса в информационном поиске. . . . . . . . . | 27 |
| 2,4 | Применение тематических моделей для расширения запроса. . . . . . . . | 30 |
| 2,5 | Помимо релевантности — персонализация поиска . . . . . . . . . | 34 |
| 2,6 | Резюме . . . . . . . . . . . . . . . . . . . . . . . . . . . | 36 |

|  |  |  |
| --- | --- | --- |
| **3 Оценка и интерпретация** | | **38** |
| 3.1 | Отображение тем. . . . . . . . . . . . . . . . . . . . . . | 38 |
| 3.2 | Маркировка тем. . . . . . . . . . . . . . . . . . . . . . . | 40 |
| 3.3 | Отображение моделей. . . . . . . . . . . . . . . . . . . . . . | 42 |
| 3.4 | Оценка, стабильность и ремонт. . . . . . . . . . . . . . | 44 |

II

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | III |
|  | 3,5 | Резюме . . . . . . . . . . . . . . . . . . . . . . . . . . . | 47 |
| **4** | **Исторические документы** | | **48** |
|  | 4.1 | Газеты . . . . . . . . . . . . . . . . . . . . . . . . . | 49 |
|  | 4.2 | Исторические записи. . . . . . . . . . . . . . . . . . . . . . | 53 |
|  | 4.3 | Научная литература. . . . . . . . . . . . . . . . . . . . . | 55 |
|  | 4.4 | Резюме . . . . . . . . . . . . . . . . . . . . . . . . . . . | 57 |
| **5** | **Понимание научных публикаций** | | **59** |
|  | 5.1 | Понимание областей исследования. . . . . . . . . . . . . . . | 61 |
|  | 5.2 | Как меняются поля. . . . . . . . . . . . . . . . . . . . . | 63 |
|  | 5.3 | Инновации. . . . . . . . . . . . . . . . . . . . . . . . . . | 65 |
|  | 5.4 | Резюме . . . . . . . . . . . . . . . . . . . . . . . . . . . | 67 |
| **6** | **Художественная литература и литература** | | **68** |
|  | 6.1 | Тематические модели в гуманитарных науках. . . . . . . . . . . . . . | 68 |
|  | 6.2 | Что такое документ? . . . . . . . . . . . . . . . . . . . . | 70 |
|  | 6.3 | Люди и места . . . . . . . . . . . . . . . . . . . . . . | 71 |
|  | 6.4 | Помимо буквального. . . . . . . . . . . . . . . . . . . . . . | 75 |
|  | 6,5 | Сравнение со стилометрическим анализом. . . . . . . . . . . . . | 77 |
|  | 6,6 | Оперативная «Тема». . . . . . . . . . . . . . . . . . | 77 |
|  | 6.7 | Резюме . . . . . . . . . . . . . . . . . . . . . . . . . . . | 78 |
| **7** | **Вычислительная социальная наука** | | **80** |
|  | 7.1 | Тематические модели для качественного анализа. . . . . . . . . . . | 83 |
|  | 7.2 | Анализ настроений. . . . . . . . . . . . . . . . . . . . . | 83 |
|  | 7.3 | Восходящие и нисходящие модели. . . . . . . . . . . . . | 85 |
|  | 7.4 | Понимание стойки и поляризации. . . . . . . . . . . | 86 |
|  | 7,5 | Социальные сети и СМИ. . . . . . . . . . . . . . . . . | 87 |
|  | 7.6 | Резюме . . . . . . . . . . . . . . . . . . . . . . . . . . . | 90 |
| **8 Многоязычные данные и машинный перевод** | | | **91** |
|  | 8.1 | Выравнивание на уровне документа из многоязычных корпусов. . . | 93 |
|  | 8.2 | Выравнивание на уровне слов из лексических данных. . . . . . . . . . | 95 |
|  | 8.3 | Выравнивание из параллельных корпусов и лексической информации. | 97 |
|  | 8.4 | Тематические модели и машинный перевод. . . . . . . . . . . | 98 |

|  |  |  |  |
| --- | --- | --- | --- |
| IV |  |  |  |
|  | 8,5 | Компоненты статистического машинного перевода. . . . | 99 |
|  | 8,6 | Тематические модели для перевода на уровне фраз . . . . . . . . . | 101 |
|  | 8,7 | Тематические модели для языкового моделирования на уровне предложений. . . . | 105 |
|  | 8,8 | Изменение порядка с помощью тематических моделей. . . . . . . . . . . . . . . | 109 |
|  | 8,9 | Помимо адаптации домена. . . . . . . . . . . . . . . . . | 110 |
|  | 8.10 | Резюме . . . . . . . . . . . . . . . . . . . . . . . . . . . | 111 |
| **9** | **Построение тематической модели** | | **112** |
|  | 9.1 | Проектирование модели. . . . . . . . . . . . . . . . . . . . . . | 113 |
|  | 9.2 | Реализация модели. . . . . . . . . . . . . . . . . . | 116 |
|  | 9.3 | Отладка и проверка. . . . . . . . . . . . . . . . . . | 122 |
|  | 9.4 | Общение с вашей моделью. . . . . . . . . . . . . . . . . | 124 |
|  | 9,5 | Резюме . . . . . . . . . . . . . . . . . . . . . . . . . . . | 125 |
| **10** | **Вывод** | | **127** |
|  | 10.1 | Борьба с информационной перегрузкой. . . . . . . . . . . . . | 127 |
|  | 10.2 | Более глубокие представления. . . . . . . . . . . . . . . . . . . | 128 |
|  | 10.3 | Автоматический анализ текста для людей. . . . . . . . . . | 129 |
|  | 10.4 Кода. . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 130 |
| **использованная литература** | | | **132** |

**Абстрактный**

Как может один человек понять, что происходит в коллекции миллионов документов? Это становится все более распространенной проблемой: просеивание электронной почты организации, понимание газет за десятилетие или характеристика научных исследований. Тематические модели — это статистическая структура, которая помогает пользователям понимать большие коллекции документов: не только находить отдельные документы, но и понимать общие темы, представленные в коллекции.

В этом обзоре описываются недавние академические и промышленные применения тематических моделей с целью подготовки молодых исследователей, способных создавать собственные приложения тематических моделей. В дополнение к эффективному применению тематических моделей к традиционным задачам, таким как поиск информации, визуализация, статистический вывод, многоязычное моделирование и лингвистическое понимание, в этом обзоре также рассматривается способность тематических моделей открывать большие текстовые коллекции для качественного анализа. Мы рассматриваем их успешное использование исследователями, чтобы помочь понять художественную и научно-популярную литературу, научные публикации и политические тексты.

* *Приложения тематических моделей* . Основы и тенденции R в информационном поиске, том. ХХ, нет. ХХ, стр. 1–154, 2017.

DOI: 10.1561/XXXXXXXXXX.

**1**

**Что и зачем тематические модели**

Представьте, что вы бесстрашный репортер с потрясающей сенсацией: у вас есть 24 часа эксклюзивного доступа к трем десятилетиям электронной почты, отправленной коррумпированной корпорацией. Вы знаете, там есть грязь и скандалы, но они хорошо замаскированы политическими друзьями корпорации. Как вы собираетесь понять этот стог сена достаточно хорошо, чтобы объяснить его вашим преданным читателям в такие сжатые сроки?

**1.1**  **Расскажите мне о своем стоге сена**

В отличие от виньетки выше, взаимодействие с большими наборами текстовых данных часто представляется как проблема иголки в стоге сена. Бедный пользователь, столкнувшись с документами, на чтение которых уйдет десятилетие, ищет единственную иголку: документ (или, самое большее, несколько документов), который соответствует тому, что ищет пользователь: электронное письмо с «дымящимся пистолетом». , документ, который лучше всего представляет концепцию [Salton, 1968] или ответ на вопрос [Hirschman and Gaizauskas, 2001].

Эти вопросы важны. Дисциплина поиска информации построена на систематизации, решении и оценке этой проблемы. Империя Google построена на том, что пользователи вводят несколько ключевых слов.

2

|  |  |
| --- | --- |
| *1.1. Расскажи мне о своем стоге сена* | 3 |

в поле поисковой системы и видеть быстрые, последовательные результаты поиска. Однако это не единственная проблема, с которой сталкиваются те, кто взаимодействует с большими наборами текстовых данных.

Другой, но родственной проблемой является *понимание* больших коллекций документов, распространенное в научной политике [Talley et al., 2011], журналистике и гуманитарных науках [Moretti, 2013a]. В стоге сена не одна драгоценная иголка. Рискуя злоупотребить метафорой, *иногда вы заботитесь о соломинке* . Вместо того, чтобы искать предупреждение о дымящемся пистолетевам какое-то преступление, которое было совершено, возможно, вы ищете грех упущения: эта компания никогда не говорила о разнообразии в своей рабочей силе? Вместо одного ответа на вопрос, возможно, вы ищете множество ответов: как люди объясняют растущее неравенство в доходах? Вместо того, чтобы искать один документ, возможно, вы хотите предоставить статистику на уровне населения: какая доля пользователей Twitter когда-либо говорила о насилии с применением огнестрельного оружия?

Сначала может показаться, что для ответа на эти вопросы потребуется построить обширную схему онтологии или категоризации. Для каждого нового корпуса вам нужно будет определить сегменты, в которые может поместиться документ, вежливо попросить некоторых библиотекарей и архивариусов поместить каждый документ в правильные сегменты, возможно, автоматизировать процесс с помощью контролируемого машинного обучения, а затем собирать сводную статистику, когда вы сделали.

Очевидно, что такие трудоемкие процессы возможны — они проводились для маркировки выступлений в конгрессе 1 и понимания эмоционального состояния [Wilson and Wiebe, 2005] — и остаются важной частью социальных наук, информатики, библиотечного дела и машинного обучения. Но эти процессы не всегда возможны, быстры или даже оптимальны, если бы у нас были бесконечные ресурсы. Во-первых, они требуют значительных затрат времени и ресурсов. Даже создание *списка* категорий является сложной задачей и требует тщательного обдумывания и калибровки. Даже если бы это было возможно, конкретный вопрос мог бы не требовать времени или усилий: творчество второстепенного автора (интересно лишь немногим) или твиты дня (неактуальные завтра).

В этом обзоре исследуются способы, с помощью которых люди и компьютеры создают

* [www.congressionalbills.org/ \_](file:///C:\temp\Rar$DIa0.245\www.congressionalbills.org\)

4 *Что и зачем тематические модели*

смысл коллекций документов с помощью инструментов, называемых тематическими моделями. Тематические модели позволяют нам быстро, дешево и без вмешательства человека отвечать на общие вопросы. После обучения они предоставляют людям основу для понимания коллекций документов как напрямую, «читая» модели, так и косвенно, используя темы в качестве входных переменных для дальнейшего анализа. Читатели, уже знакомые с тематическими моделями, могут пропустить эту главу; мы в основном рассмотрим определения и реализации тематических моделей.

Целевой аудиторией этой книги является читатель, обладающий некоторыми знаниями в области обработки документов (например, знающий, что такое «токены» и «документы»), имеющий базовое представление о некоторой вероятности (например, что такое распределение) и интересующийся многими областями применения. . Мы обсуждаем информационные потребности каждой прикладной области и то, как эти конкретные потребности влияют на модели, процедуры курирования и интерпретации.

Мы надеемся, что к концу книги (глава 9) читатели будут достаточно воодушевлены, чтобы попытаться приступить к построению своих собственных тематических моделей. В этой главе мы углубимся в детали реализации. Читатели, которые уже являются экспертами по тематическому моделированию, скорее всего, не узнают многого в техническом плане, но мы надеемся, что наше освещение разнообразных приложений познакомит эксперта по тематическому моделированию с моделями и подходами, которых они раньше не видели.

**1.2**  **Что такое тематическая модель**

Возвращаясь к нашему мотивирующему примеру, рассмотрим электронные письма от Enron, прототипа проблемной корпорации начала века. Источник предоставил вам кучу электронных писем, и ваш редактор требует статью к вчерашнему дню. Вы знаете, что произошло правонарушение, но не знаете, кто его совершил и как оно было спланировано и осуществлено. У вас есть подозрения (например, в отношении спотового рынка электроэнергии в Калифорнии), но вас интересуют другие скелеты в шкафу, и вы очень заинтересованы в том, чтобы их найти.

Итак, вы запускаете тематическую модель данных. Верная своему названию, тематическая модель дает вам «темы», наборы слов, которые имеют смысл вместе. Электронные письма Enron раскрывают темы о газовых контрактах, регулирующих органах Калифорнии,

|  |  |
| --- | --- |
| *1.3. Фонды* | 5 |

**Таблица 1.1.** Пять тем из двадцати пяти тем подходят для электронных писем Enron. Примеры тем касаются финансовых операций, природного газа, коммунальных услуг Калифорнии, федерального регулирования и совещаний по планированию. Мы предоставляем пять наиболее вероятных слов из каждой темы (каждая тема — это распределение по всем словам).

Условия темы

* торговая цена финансового торгового продукта

6 газопроводов

9штат Калифорния Дэвис Энергосбыт

14 ferc издать приказ партии дело

План процесса команды на 22 встречи группы

и цены акций (рис. 1.1).

Первая половина тематической модели связывает темы с беспорядочным «мешком слов». Когда мы говорим, что тема посвящена *X* , мы вручную присваиваем метку *постфактум* (подробнее об этом в главе 3.1). Ответственность за то, чтобы пойти дальше и разобраться в этих грудах соломы, лежит на пользователе-потребителе тематических моделей (более подробно мы обсуждаем маркировку тем в главе 3).

Разобраться в одном из этих наборов слов может быть непросто. Вторая половина модели темы связывает темы с отдельными документами. Например, документ на Рисунке 1.1 описывает реакцию калифорнийской коммунальной службы на краткосрочный рынок электроэнергии и иллюстрирует Тему 9 на Рисунке **??** . Рассмотрение примеров документов, тесно связанных с темой, а также слов, связанных с этой темой, может дать нам более полное представление о теме. Если мы почувствуем, что Тема 9 представляет интерес, мы можем изучить ее глубже, чтобы найти другие документы.

**1.3**  **Основы**

Вы могли заметить, что мы используем общий термин «тематическая модель». Существует множество математических формулировок тематических моделей и множество алгоритмов, которые изучают параметры этих моделей на основе данных. Аль-

6 *Что и зачем тематические модели*

Вчера SDG&E подала ходатайство о принятии механизма возмещения затрат на закупку электроэнергии и о сокращении срока подачи сторонами комментариев по этому механизму. Прилагаемое электронное письмо от SDG&E содержит предложение, краткое изложение и подробное изложение их предложений и рекомендаций, регулирующих закупку чистых дефицитных энергетических потребностей для клиентов SDG&E. Утилита запрашивает 15-дневный период комментариев, что означает, что комментарии должны быть поданы до 10 сентября (8 сентября — суббота). Ответные комментарии будут поданы через 10 дней.

Тема Вероятность

* 0,42

1. 0,05

8 0,05

**Рисунок 1.1:** Пример документа из корпуса Enron и его связь с темами. Хотя в нем нет слова «Калифорния», в нем обсуждается недовольство одной из калифорнийских коммунальных служб тем, сколько она платит за электроэнергию.

М×К ×

Назначение темы

К × В

Темы

≈

М × В

Набор данных

**Рисунок 1.2:** Матричная формулировка поиска *K* тем для набора данных с *M* документами и *V* уникальными словами. Хотя этот взгляд на тематическое моделирование включает в себя такие подходы, как латентный семантический анализ (lsa, где аппроксимация основана на svd), в оставшейся части этого обзора мы сосредоточимся на вероятностных методах.

|  |  |
| --- | --- |
| *1.3. Фонды* | 7 |

хотя мы сосредоточимся на конкретных моделях и алгоритмах, мы выбираем нашу терминологию, чтобы подчеркнуть, что сходства между формулировками, моделями и алгоритмами часто больше, чем их различий.

Тематическое моделирование началось с подхода линейной алгебры [Deerwester et al., 1990], называемого скрытым семантическим анализом (LSA): найти наилучшее низкоранговое приближение матрицы терминов документа (рис. 1.2). Хотя в последние годы эти подходы возродились [Anandkumar et al., 2012, Arora et al., 2013], мы сосредоточимся на вероятностных подходах [Hofmann, 1999a, Papadimitriou et al., 2000, Blei et al., 2003]. , которые интуитивно понятны, хорошо работают и допускают легкое расширение (как мы увидим позже во многих наших последующих главах).

Двумя основными вероятностными тематическими моделями являются латентное распределение Дирихле [Blei et al., 2003, lda] и вероятностный латентно-семантический анализ [Hofmann, 1999a, plsa]. Мы подробно описываем первую в главе 1.4, но мы хотим воспользоваться моментом, чтобы рассмотреть некоторые исторические связи между этими двумя моделями.

plsa была исторически первой и положила начало lda. plsa широко использовался во многих приложениях, таких как поиск информации. Однако в этом обзоре основное внимание уделяется lda, потому что многие исследователи не просто *использовали* lda, но и *расширили* его. lda не только широко используется, но и широко модифицируется. Из-за этих многочисленных модификаций мы сосредоточимся на механике lda, которую многие исследователи использовали в качестве основы для новых моделей. Однако, как мы объясним ниже (Глава 1.5.4), сходства между plsa и lda перевешивают различия.

В любой технической области общие термины обычно имеют конкретное, конкретное значение, и это может быть источником путаницы. В тематическом моделировании слово «тема» приобретает конкретное значение распределения вероятностей по словам, но при этом указывает на более общее значение темы или предмета дискурса. Поскольку другие области информационного поиска аналогичным образом разработали определенные значения для слова «тема», мы различаем их здесь. Наиболее распространенным определением является конкретная информационная потребность, как в оценочных корпусах TREC, разработанных NIST [Voorhees and Harman, 2005]. Темы TREC, как правило, гораздо более конкретны, чем темы тематической модели, и могут быть связаны

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 8 |  |  |  |  |  |  |  |  |  |  |  |  | *Что и зачем тематические модели* | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | Пример | | | Пример | |  |
| Распределение |  |  |  |  |  |  | Плотность | | | | | | |  |  |  | Параметры | | | Розыгрыши |  |  |
| гауссовский |  |  |  | √ | | | 1 |  |  | *е* — | | ( *х* - *µ* ) 2 | |  |  |  | *µ знак* равно 2 *,σ* 2 знак равно 1 *.* 1 | | | *х* = 2 *.* 21 | |  |
|  |  |  |  |  |  | 2 *о* 2 |  |  |  |  |
|  |  |  |  | 2 |  |  |  |  |  |  |
|  |  |  |  |  |  | 2 *о* | | | *π* | | |  |  |  |  |  | *ф* = 0 *.* 6 | |  |  |  |  |
| Дискретный |  |  |  |  |  |  |  | *я φ я* 1 [ *ш знак* равно *я* ] | | | | | |  |  |  | *ш* = 2 |  |  |
|  |  |  |  |  |  |  | Вопрос | |  |  |  |  |  |  |  |  | 0 *.* 1 | |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 0 *.* 3 | |  |  |  |  |
|  |  |  | *К* | |  |  |  |  |  |  |  |  |  |  |  |  |  | 1 *.* 1 | | 0 *.* 8 |  |  |
| Дирихле |  |  |  |  |  | | |  | | |  |  | *я* =1 *я \_* | | − |  | *α* = | 0 *.* 1 |  | *θ* = 0 *.* 15 |  |  |
|  |  |  |  | *К* | | | *α я* | | | |  |  |  |
|  |  | Вопрос |  |  | *я* =1 | | |  |  |  |  |  | *К* | *α я* |  | 1 |  | 0 *.* 1 |  | 0 *.* 05 |  |  |
|  |  |  | *я* =1 | | |  | ( *α я* ) | | | | |  |  |  |  |  |  |  |  |
|  |  | п | | |  |  |  |  | Вопрос | | | |  |  |  |  |  |  |

**Рисунок 1.3:** Примеры распределения вероятностей, используемые в генеративных историях тематических моделей. В случае дискретного розыгрыша *w =* 2 означает, что выпал второй элемент (с вероятностью 0,6 ).

к конкретным аспектам или взглядам на предмет. Примером из проекта TREC Robust Track 2003 г. является «Определить положительные достижения телескопа Хаббл с момента его запуска в 1991 г.» [Voorhees, 2003]. Подобно поиску информации, связанная с этим область обнаружения и отслеживания тем также имеет конкретное техническое определение «темы» [Allan, 2002]. В tdt «тема» обычно ближе к событию или отдельной истории. Напротив, тематические модели, как правило, выявляют более абстрактные скрытые факторы. Например, тема tdt может включать землетрясение на Гаити, тогда как тематическая модель может представлять то же событие в виде комбинации таких тем, как Гаити, стихийные бедствия и международная помощь.

Была проведена некоторая работа по использованию тематических моделей для обнаружения возникающих событий путем поиска изменений в вероятности темы [AlSumait et al., 2008]. Но эти методы, как правило, определяют в основном тот факт, что событие произошло, не обязательно определяя специфические черты этого события. Другая работа показала, что более лексически специфичные методы, чем тематические модели, лучше всего подходят для идентификации мемов и вирусных фраз [Leskovec et al., 2009].

**1.3.1**  **Вероятностные строительные блоки**

В вероятностных моделях мы хотим найти значения для ненаблюдаемых переменных модели, которые хорошо объясняют наблюдаемые данные. Первый шаг в умозаключении состоит в том, чтобы повернуть этот процесс вспять и утвердить способ

|  |  |
| --- | --- |
| *1.3. Фонды* | 9 |

генерировать данные с заданными переменными модели. Таким образом, вероятностные модели начинаются с генеративной истории: рецепта, в котором перечисляется последовательность случайных событий, которая создает набор данных, который мы пытаемся объяснить. На рис. 1.3 перечислены некоторые из ключевых участников этих историй, как они параметризуются и как выглядят выборки, взятые из этих распределений. Мы кратко обсудим их, так как позже мы будем использовать их для построения самых разнообразных тематических моделей.

**Гауссово** . Если вы уже знаете какое-либо распределение вероятностей, то это (вероятно) гауссовское. Это распределение не играет роли в самых основных тематических моделях, которые мы будем обсуждать здесь, но это будет позже (например, в главе 7). Мы включаем его, потому что это полезная точка для сравнения с другими используемыми нами дистрибутивами *(* поскольку, возможно, это самый простой для понимания и самый известный). Гауссиана — это распределение по всем действительным числам (например , *0,0* , *0,5* , *-4,2* , *π* , *...* ) *.* Вы можете попросить его выдать число, и он выдаст вам некоторое действительное число между отрицательной бесконечностью и положительной бесконечностью. Но не все числа имеют одинаковую вероятность. Распределения Гаусса параметризуются средним значением *µ* и дисперсией *σ* 2 . Большинство выборок из распределения будут близки к среднему значению *µ* ; насколько близко определяется дисперсия: более высокие дисперсии приведут к тому, что выборки будут более разбросанными.

**Дискретный** В то время как гауссово распределение находится в непрерывном пространстве, документы представляют собой комбинации дискретных символов, обычно токенов слов. 2 Таким образом, нам нужно распределение по дискретным множествам.

Полезной метафорой для размышлений о дискретных распределениях является взвешенная игральная кость. Количество граней на кубике является его измерением, и каждая грань связана с определенным результатом. Каждое лицо имеет свою собственную вероятность того, насколько вероятен этот исход; эти вероятности являются параметрами дискретного распределения (рис. 1.3).

Тематические модели описываются дискретными распределениями (иногда называемыми полиномиальными распределениями), которые описывают связь между словами и темами (первая половина) и темами и документами (вторая половина).

2 Новая тенденция в исследованиях обработки естественного языка заключается в том, чтобы рассматривать слова как встроенные в непрерывное пространство. Мы обсудим эти подходы «обучения представлениям» и их связь с тематическим моделированием в главе 10, но даже в этом случае модели по-прежнему определяются на дискретном наборе слов.

10 *Что и зачем тематические модели*

половина). Распределение по словам называется распределением по темам; каждая из тем присваивает некоторым словам больший вес, чем другим (например, в теме 9 из корпуса Enron слова «штат» и «калифорния» имеют более высокую вероятность, чем другие слова). Каждый документ также имеет «распределение» по каждой теме: документы содержат небольшое количество тем, и большинство документов имеют очень низкий вес для большинства возможных тем.

**Дирихле** Хотя дискретные распределения являются главными игроками в тематических моделях, это еще не конец истории. Мы часто начинаем с распределений Дирихле. Точно так же, как гауссовы числа производят действительные числа, а дискретные распределения производят символы из конечного набора, распределения Дирихле производят векторы вероятности, которые можно использовать в качестве параметров дискретных распределений. Подобно распределению Гаусса, они имеют параметры, аналогичные среднему значению и дисперсии. Среднее значение называется «базовой мерой» *τ* и представляет собой ожидаемое значение распределения Дирихле: значения, которые вы получили бы, если бы усреднили множество розыгрышей из распределения Дирихле. Параметр концентрации *α* 0 определяет, насколько далеко отдельные розыгрыши находятся от базовой меры. Мы часто объединяем эти параметры в одно значение для каждого измерения: *α k* = *α* 0 *τ k* .

Если *α* 0 очень велико, то вытягивания из Дирихле будут очень близки к *τ* (рис. 1.4, слева). Однако если *α* 0 мало, дискретные распределения становятся разреженными (рис. 1.4, справа). Разреженное распределение — это распределение, в котором только несколько значений имеют высокую вероятность, а все остальные значения малы.

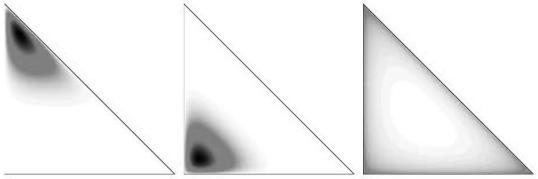
Поскольку тематические модели предназначены для отражения свойств реальных документов, важна разреженность моделирования. Когда человек садится за написание документа, он пишет лишь о нескольких темах, которые потенциально мог бы использовать. Они не пишут о каждой возможной теме, и разреженность распределений Дирихле является вероятностным инструментом, который кодирует эту интуицию.

Есть несколько важных частных случаев распределения Дирихле. Если базовая мера *τ* является равномерной, мы называем полученное распределение *симметричным* . Этот случай уместен, когда мы не ожидаем ни одного

|  |  |
| --- | --- |
| *1.4. Скрытое распределение Дирихле* | 11 |

α=10 α=10 α=0,1

* = (.8, .2, .2)τ = (.2, .8, .2) τ = (0.33, 0.33, 0.33)



**Рисунок 1.4:** При различных параметрах Дирихле распределение Дирихле может быть либо информативным (слева, посередине), либо разреженным (справа). Разреженные распределения побуждают распределения отдавать предпочтение нескольким элементам, но не важно, какие именно. Это согласуется с нашими интуитивными представлениями о том, как пишутся документы: они касаются лишь нескольких вещей, а темы содержат всего несколько слов.

элемент в среднем более вероятен, чем любой другой элемент во всех выборках из распределения. В симметричном случае распределение имеет только один параметр — концентрацию *α* 0 . Если базовая мера однородна, а параметр концентрации *α* 0 равен количеству измерений *K* (или, что то же самое, *α k* = 1 *.* 0 для всех *k* ), распределение является однородным, что делает вероятность равной для всех *K* -мерных вероятностей . дистрибутивы.

**1.4**  **Скрытое распределение Дирихле**

Теперь у нас есть все необходимые инструменты, чтобы рассказать полную историю самой популярной тематической модели: скрытого распределения Дирихле [Blei et al., 2003, lda]. Скрытое распределение Дирихле 3 постулирует «порождающий процесс» того, как появились данные. Мы собираем вероятностные фрагменты, чтобы рассказать эту историю о создании тем и о том, как эти темы используются для создания

3 Название lda — это игра с lsa, его невероятностным предшественником (латентный семантический анализ). Скрытый, потому что мы используем вероятностный вывод, чтобы вывести недостающие вероятностные части порождающей истории. Дирихле из-за разреженности кодирования параметров Дирихле. Распределение, потому что распределение Дирихле кодирует априорное распределение каждого документа по темам.

12 *Что и зачем тематические модели*

разнообразные документы.

**Создание тем** Первая часть истории заключается в создании тем. Пользователь указывает, что существует *K* различных тем. Каждая из *K* тем взята из распределения Дирихле с однородным базовым распределением и параметром концентрации *λ* : *φ k* ∼ Dir( *λ* ***u*** ). Дискретное распределение *φ k* имеет вес для *каждого* слова в словаре.

Однако, когда мы суммируем темы (как на рисунке **??** ), мы обычно используем только верхние (наиболее вероятные) слова темы. Слова с более низкой вероятностью менее релевантны теме и поэтому не показаны.

**Распределения** документов — это распределения по темам для каждого документа. Это кодирует, о чем документ; разреженность параметра концентрации распределения Дирихле *α* 0 гарантирует, что документ будет посвящен только нескольким темам. Каждый документ имеет дискретное распределение по теме: *θ d* ∼ Dir( *α* ***u*** ).

**Слова в контексте** Теперь, когда мы знаем, о чем каждый документ, мы создаем слова, которые появляются в документе. Предположим 4 , что в документе *d* имеется *N d* слов . Для каждого слова *n* в документе *d* мы сначала выбираем **тематическое задание** *z d,n* ∼ Дискретный( *θ d* ). Это одна из тем *K* , которая сообщает нам, из какой темы взят токен слова, но не что это за слово.

Чтобы выбрать, какое слово мы увидим в документе, мы снова рисуем из дискретного распределения. Учитывая назначение темы токена слова *z d,n* ,

мы берем из этой темы, чтобы выбрать слово: *w d,n* ∼ *φ z d,n* . Назначение темы говорит вам, о чем слово, а затем выбирает, какое

распределение по словам, которые мы используем для создания слова.

Например, рассмотрим документ на рис. 1.1. Чтобы сгенерировать его, мы выбираем распределение по всем темам. Это *θ* . Для этого документа предпочтение отдается теме 9 о Калифорнии . Ценность этой темы выше, чем любой другой темы. Для каждого слова в документе

4 Мы можем смоделировать это и в генеративной истории, например, с помощью распределения Пуассона. Однако нас часто не волнует *длина документа* , а только то, о чем этот документ, поэтому обычно мы можем игнорировать эту часть истории.

|  |  |
| --- | --- |
| *1.5. Вывод* | 13 |

порождающий процесс выбирает тему задания *z n* . Для этого документа теоретически возможна любая тема, но мы ожидаем, что большинство из них будет темой 9.

Затем для каждого токена в документе нам нужно выбрать, какой тип слова будет отображаться. Это происходит из распределения слов в Теме 9 (несколько тем имеют распределение слов, показанное на рисунке **??** ). Каждое из них представляет собой дискретный рисунок из распределения слов по теме, что делает более вероятными такие слова, как «Калифорния», «штат» и «Сакраменто».

Само собой разумеется, что порождающая история — это фикция [Box and Draper, 1987]. Никто не садится за кости, чтобы решить, что печатать на клавиатуре. Мы используем эту историю, потому что она *полезна* . Эта причудливая история о случайном выборе темы для каждого слова может помочь нам, потому что, если мы предположим этот порождающий процесс, мы можем работать в обратном направлении, чтобы найти темы, которые объясняют, как была создана коллекция документов: каждое слово, каждый документ связывается с этими лежащими в основе темы.

Эта простая модель помогает нам упорядочить нашу коллекцию документов: предполагая эту историю, мы можем обнаружить *темы* (которых, конечно, не существует), чтобы мы могли понять общие темы, которые люди используют для написания документов. Как мы увидим в следующих главах, небольшие изменения в этой генеративной истории позволяют нам раскрыть более сложные структуры: как авторы предпочитают определенные темы, как темы меняются или как темы могут использоваться в разных языках.

**1.5**  **Вывод**

Учитывая генеративную модель и некоторые данные, процесс обнаружения скрытых частей вероятностной генеративной истории называется *выводом* . Более конкретно, это рецепт создания алгоритмов перехода от данных к *темам, объясняющим набор данных* .

Существует множество разновидностей алгоритмов апостериорного вывода: передача сообщений [Zeng et al., 2013], вариационный вывод [Blei et al., 2003], градиентный спуск [Hoffman et al., 2010] и выборка Гиббса [Grifths and Steyvers]. , 2004]. У всех этих алгоритмов есть свои сторонники и причины, по которым вам следует их использовать. В этом обзоре мы сосредоточимся на выборке Гиббса,

14 *Что и зачем тематические модели*

который прост, интуитивно понятен и — с некоторыми хитрыми приемами, характерными для тематических моделей — быстр [Yao et al., 2009]. (Мы обсуждаем вариационный вывод в главе 9.)

Мы представляем результаты выборки Гиббса без вывода, которые — наряду с историей их возникновения в статистической физике — хорошо описаны в других источниках. 5 Мы используем разновидность выборки Гиббса, называемую *свернутой* выборкой Гиббса, которая позволяет делать выводы, обходя некоторые части порождающей истории: вместо явного представления параметров дискретного распределения, отличного от любых наблюдений, извлеченных из этого распределения, мы представлять распределение исключительно по этим наблюдениям. Затем мы можем воссоздать тему и задокументировать распределение с помощью простых формул.

**1.5.1**  **Случайные величины**

**Назначения темы** Поскольку предполагается, что каждый отдельный токен генерируется из одной темы, мы можем рассматривать назначение *темы* токена как переменную. Например, экземпляр слова «сборник» может быть в теме о компьютерах в одном документе и в теме об искусстве в другом документе. Поскольку каждый токен имеет собственное назначение темы, возможно даже, что одно и то же слово может быть назначено разным темам в одном и том же документе. Чтобы оценить *глобальные* свойства тематической модели, мы используем совокупную статистику, полученную из этих назначений тем на уровне токенов.

**Распределение документов Распределение** документов — это распределение по темам для каждого документа; другими словами, это говорит о том, насколько популярна каждая тема в документе. Если мы посчитаем, как часто в документе используется тема, это даст нам представление о популярности. Давайте определим *N d,i* как количество раз, когда документ *d* использует тему *i* . Это больше для более популярных тем; однако это не вероятность, поскольку она больше единицы. Мы можем сделать это вероятностью, разделив на количество слов в

5 Мы рекомендуем Resnik and Hardisty [2009] для получения дополнительной информации о выводе.

|  |  |  |  |
| --- | --- | --- | --- |
| *1.5. Вывод* | |  | 15 |
| документ | |  |  |
|  | *Н д, я* | *,* | (1.1) |
|  | P *k N d,k* |

но это проблематично, потому что иногда может дать нам ноль и игнорировать влияние распределения Дирихле; лучшая оценка 6

|  |  |  |
| --- | --- | --- |
| *Н д, я* + *α я* |  | (1.2) |
| *θ d,i* ≈ P *k N d,k* + *α k* | *.* |

Он никогда не должен становиться равным нулю, потому что мы не хотим, чтобы исключалась возможность использования темы в конкретном документе (следовательно, каждый *α* должен быть ненулевым). Это помогает сэмплеру исследовать больше возможных комбинаций.

**Темы** Каждая тема представляет собой распределение по словам. Чтобы понять, о чем тема, мы смотрим на профиль всех токенов, которые были назначены этой теме. Мы оцениваем вероятность слова в теме как

|  |  |  |
| --- | --- | --- |
| *В я, в* + *β v* |  | (1.3) |
| *φ i,v* ≈ P *w V i, w* + *β w* | *,* |

где *β* — параметр Дирихле для тематического распределения.

**1.5.2**  **Алгоритм**

Алгоритм свернутой выборки Гиббса для изучения тематической модели основан только на заданиях по темам, но мы будем использовать наши оценки для тем *φ k* и документов *θ d ,* обсуждавшихся выше. Начнем со случайного назначения тем: если у нас есть *K* тем, каждое слово имеет равные шансы быть связанным с любой из тем. Эти темы будут совсем плохи, выглядя как зашумленные копии общего дистрибутива корпуса. Но мы будем улучшать их по одному слову за раз.

Алгоритм продолжается, перебирая все токены слов по очереди снова и снова. На каждой итерации мы меняем назначения тем для каждого слова таким образом, чтобы они отражали лежащую в основе вероятностную модель данных. В среднем каждый проход по данным делает темы немного

6 Чтобы быть техническим, уравнение 1.1 является оценкой максимального правдоподобия, а уравнение 1.2 является максимальной *апостериорной* , которая включает влияние как априорных, так и данных.

16 *Что и зачем тематические модели*

лучше, пока модель не достигнет устойчивого состояния. Нет простого способа сказать, когда такое устойчивое состояние было достигнуто, но в конечном итоге темы «сойдутся» к разумным темам, и вы можете считать, что закончили.

Уравнение вероятности отнесения слова к определенной теме объединяет информацию о словах и о документах 7

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *p* ( *z d,n* = *i* | *. . .* ) = *θ d φ j i* = | P *k N d,k* + *α k* ! | | P *w V i,w* + *β w* ! | | | *.* (1.4) |
|  | *Н д, я* + *α я* |  |  | *В я, ш д, н* + *β v* |  |  |

Вычисление этого значения для каждой темы приведет к распределению вероятностей по назначению темы для этого токена слова с учетом всех других назначений темы. Следующим шагом является случайный выбор одного из этих индексов с вероятностью, пропорциональной значению вектора. Теперь вы назначаете

это слово в тему, обновить *N d,* · и *V* · *,w d,n* , перейти к следующему слову и повторить. Эти два термина создают два «давления» для глобального

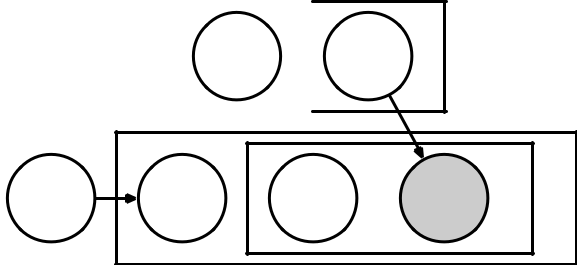
и местная согласованность. Разреженность в распределении слов по темам способствует назначению токенов одного и того же типа слов небольшому количеству тем, независимо от того, где они встречаются. Разреженность в распределениях по темам документов способствует назначению токенов в одном и том же документе небольшому количеству тем, независимо от их типа. Например, знание слова «сборник» значительно сужает количество потенциальных тем, но оставляет неясность: это компьютерная компиляция или музыкальная компиляция? Знание того, что это слово встречается в документе со многими другими словами в теме *искусства* , устраняет эту двусмысленность, оставляя тему *искусства* в качестве наиболее вероятного назначения.

В самом конце алгоритма мы можем использовать оценки каждой темы (уравнение 1.3), чтобы обобщить основные темы корпуса, и оценки распределения тем каждого документа (уравнение 1.2), чтобы начать изучение коллекции автоматически (глава 2). или с человеком в петле (глава 3).

Алгоритм, который мы набросали здесь, является основой многих более продвинутых моделей, которые мы обсудим позже в этом обзоре. Пока мы не будем подробно описывать алгоритмы,

7 Чтобы быть теоретически правильным, важно не включать в эти подсчеты счетчик, связанный с токеном, который вы выбираете, что становится более ясным, если вероятность записать как *p* ( *z d,n* = *j* | *z d,* 1 *. . . zd ,n* −1 *, zd ,n* +1 *. . . z d,N d , w d,n* ), чтобы показать зависимость от тематических назначений *всех других* токенов, но не этого токена.

|  |  |
| --- | --- |
| *1.5. Вывод* | 17 |



α

λ ф к



К

θ d z n w n



Н М

**Рисунок 1.5:** Диаграмма плиты для lda. Узлы показывают случайные величины, линии показывают (возможную) вероятностную зависимость, прямоугольники показывают повторение, а штриховка показывает наблюдение.

время от времени ссылайтесь на этот набросок, чтобы подчеркнуть проблемы или трудности в реализации тематических моделей.

**1.5.3**  **Схемы пластин**

Пластинчатые диаграммы позволяют быстро объяснить, какие случайные величины связаны друг с другом. Если вы просмотрите многие ссылки, используемые в этом обзоре, вы, вероятно, увидите диаграммы пластин (мы также используем диаграмму пластин позже на рис. 2.1b).

Начнем со схемы пластин для lda (рис. 1.5). Вы можете сравнить их с генеративной историей в главе 1.4. Все случайные величины есть, каждая в своем кругу. Линии между случайными величинами говорят больше об истории. Вы можете видеть, что если случайная величина обусловлена другой, есть линия, идущая от переменной, которая *обусловлена* **к** переменной, которая является *условно зависимой* . Например, слово зависит от назначения токена *z d,n* и темы *φ k* , поэтому мы рисуем линии из обоих.

Вы можете думать о прямоугольных коробках как о повторении. Буква в правом нижнем углу коробки показывает, как часто копируется то, что находится внутри коробки. Для каждого документа (всего *M* ) и каждого жетона (коробка слов находится внутри коробки для документов) есть коробка.

Когда переменная заштрихована, это означает, что она наблюдается. Это данные, с которых мы начинаем. Незаштрихованные переменные должны быть либо выведены

18 *Что и зачем тематические модели*

(например, темы *φ* ) или являются гиперпараметрами, которые должны быть установлены или выведены (например, параметр Дирихле *α* ).

Диаграммы пластин позволяют читателю быстро увидеть «семейное сходство» между связанными моделями, и как только кто-то полностью погрузится в тематические модели, часто становится возможным с первого взгляда понять модель по ее диаграмме пластин. Однако диаграммы пластин несовершенны; в них отсутствует часть ключевой информации, необходимой для понимания модели. Например, точная вероятностная связь между переменными недоопределена.

**1.5.4**  **Что такого замечательного в Дирихле?**

Теперь, когда мы описали, что такое lda, можно вернуться к его истории. В чем заключается инновация, которая отличает lda от plsa, ее предшественницы? Наивно, разница заключается в замене «s» на «d» (т. е. замене pl **S** a на l **D** a). Более глубокая история примерно столь же последовательна.

Вместо того, чтобы иметь априор Дирихле над *θ* , plsa предполагает, что *θ* является дискретным параметром. На практике это означает, что документы не поощряются к сосредоточению на ограниченном числе тем и часто «рассредоточены», чтобы иметь небольшой вес для многих различных тем. Теоретически это означает, что генеративная история о том, как появился документ, не может быть столь убедительной: вы не можете запускать процесс генерации с нуля, если вам необходимо иметь *θ* в качестве параметра для начала.

Эти различия относительно невелики. У lda вывод немного проще, особенно когда дело доходит до настройки модели, что сделало ее более популярной из двух моделей. Таким образом, мы сосредоточимся на сравнении моделей с lda. Это не умаляет значения plsa и его неоспоримого места в литературе, но помогает нам представить нашему читателю более целостное повествование.

**1.5.5**  **Реализации**

Надеюсь, предыдущий набросок алгоритма убедил вас в том, что реализация тематических моделей — это не геркулесова задача; наиболее опытные программисты могут выполнить разумную реализацию тематических моделей менее чем за день. Тем не менее, мы рекомендуем не пытаться реализовывать базовые lda, если вам нужен только вывод тематической модели, много надежных реализаций.

|  |  |
| --- | --- |
| *1.5. Вывод* | 19 |

могут помочь пользователям быстрее получить полезные результаты, особенно потому, что тематические модели часто требуют обширной предварительной обработки.

Mallet работает быстро и является широко используемой реализацией в Java [McCallum, 2002]. Вот с этого, пожалуй, и следует начать, по нашему предвзятому мнению. Он работает на Java, использует высокооптимизированные реализации сэмплирования Гиббса и может работать с различными текстовыми входами. Он хорошо документирован, разработан и хорошо работает на многоядерных компьютерах, что позволяет ему обрабатывать до миллионов документов. Другим важным вариантом является вариационный вывод [Blei et al., 2003; Langford et al., 2007], но часто требуется немного больше усилий для получения первого результата от новых пользователей.

Однако не всем пользователям удобно работать с Java; многие реализации доступны на других платформах и на многих языках программирования. 8 Многие из этих реализаций хорошо построены, но проверьте, обладают ли они всеми функциями зрелых реализаций, таких как Mallet, чтобы знать, что (если вообще что-то) вам не хватает.

Однако, если ваш корпус действительно большой, рассмотрите методы, которые можно распараллелить на больших компьютерных кластерах. Эти методы могут быть основаны на вариационном выводе [Narayanamurthy, 2011, Zhai et al., 2012] или на выборке [Newman et al., 2008].

В то время как эти реализации позволяют вам запускать *определенные* тематические модели, другие платформы позволяют вам указывать произвольные генеративные модели. Это позволяет быстро создавать прототипы тематических моделей и интегрировать тематические модели с другими вероятностными структурами, такими как регрессия или совместная фильтрация. Примеры этих общих фреймворков включают Stan [Stan Development Team, 2014], Theano [Theano Development Team, 2016] и Infer.net [Minka et al., 2014].

Если вы не можете найти конкретную модель, которая вам нужна, среди этих существующих программных пакетов, гибкость и простота тематических моделей и логических выводов делает относительно простой адаптацию тематических моделей для моделирования конкретных явлений (как мы опишем в следующих главах).

* Так много, что меняется так быстро; таким образом, мы неохотно одобряем конкретные здесь.

20 *Что и зачем тематические модели*

**1.6**  **Остальная часть этого обзора**

В каждой из следующих глав мы сосредоточимся на применении тематических моделей, постепенно увеличивая сложность лежащих в их основе моделей. Главы иногда ссылаются друг на друга, но читатель должен иметь возможность читать каждую из глав независимо друг от друга.

В следующей главе мы вернемся к различию между обзорами высокого уровня и поиском иголки в стоге сена. Мы показываем, как обзор высокого уровня может помочь пользователям и алгоритмам найти интересующие документы. Мы показываем, как обзор высокого уровня может помочь алгоритмам (глава 2) и пользователям (глава 3) находить интересующие документы.

Эти инструменты помогают по-новому применять тематические модели: как понимание газет (глава 4) раскрывает ход истории, как корпус писателей-фантастов (глава 6) освещает социальные нормы, как научные труды раскрывают инновации (глава 5). ), или как выступления политиков (глава 7) выявляют расколы в политических организациях.

Наконец, опрос завершается мыслями о том, как заинтересованные исследователи могут начать создавать свои собственные тематические модели (глава 9) и как тематические модели могут измениться в будущем (глава 10).

**2**

**Специальный поиск информации**

Тематические модели исследуют и резюмируют коллекции документов вне контекста какой-либо конкретной потребности в информации, когда мы не обязательно знаем, что ищем. Такой подход к поиску информации отличается от традиционных ИК-систем, которые извлекают соответствующие документы с учетом явных информационных потребностей пользователей. Там, где их системы могут искать «иголку в стоге сена», тематические модели расскажут вам об общей доле сена и иголок и, возможно, сообщат вам о мышах, о существовании которых вы не знали. Но тематические модели также могут быть полезны в ситуациях, когда у нас есть конкретная потребность в информации, но мы не совсем знаем, как ее искать. Несмотря на различия в целях, между этими двумя подходами существуют прочные математические и концептуальные связи. В этой главе мы рассмотрим использование тематического моделирования в IR, чтобы сбалансировать конкретные пользовательские запросы с более открытыми исследованиями.

В самом прямом смысле тематические модели могут использоваться как простой метод индексации. Пользователи могут находить темы, которые присваивают высокую вероятность определенному термину запроса, а затем находить документы с высокой вероятностью этих тем. Такой поиск по теме может дополнительно обеспечить некоторый уровень устранения неоднозначности запроса, поскольку он может быть понятен из тематического слова.

21

22 *Специализированный поиск информации*

рассылки о том, что та или иная тема более актуальна для информационной потребности пользователя. Более сложные подходы стирают границу между поиском на основе запросов и неконтролируемым моделированием тем. Erlin [2017] ищет отрывки, связанные с эпистемологией, в английских и немецких книгах, «засевая» тематические модели словами, которые, как считается, имеют отношение к этому предмету. Такой подход может быть успешным, но не гарантирует, что будут найдены релевантные темы или что темы будут соответствовать предполагаемой теме.

В более формальной настройке поиска *ad-hoc пользователи начинают с потребности в информации, выраженной в запросах.* Многие ИК-системы рассматривают как запросы, так и документы как «наборы слов», а затем извлекают и ранжируют документы, измеряя совпадение слов между запросами и документами. Однако возможности этого прямого и простого сопоставления всегда ограничены. Слова с похожим значением или в разных формах также должны рассматриваться как совпадающие, а не игнорироваться. **Языковое моделирование** было одной из самых популярных сред для фиксации таких семантических отношений. Но люди также хотели бы использовать базовые знания для интерпретации и понимания запросов и «добавления» пропущенных слов [Wei, 2007], что обеспечивает другой подход, называемый **расширением запроса** , для улучшения результатов поиска и ранжирования.

В обоих направлениях можно двигаться, изучая и открывая семантические отношения между словами и, далее, семантические отношения между запросами и документами. Тематические модели обеспечивают семантические отношения между словами запроса и документами [Deerwester et al., 1990; Hofmann, 1999b], описывая каждую тему с помощью вероятностно взвешенных слов и моделируя каждый документ как распределение по всем темам. Это добавляет уровень абстракции между документом и точными словами, присутствующими в этом документе. Апеллируя к генеративной «истории» модели, мы хотим восстановить слова, которые *могли* быть доступны автору на основе выбранных слов. Такие семантические отношения можно применять для сглаживания языковых моделей или введения связанных слов в расширение запроса. В этой главе основное внимание уделяется тому, как применять тематические модели в моделировании языка документов [Lu et al., 2011, Wei and Croft, 2006] и расширении запросов [Park and Ramamohanarao, 2009, Andrzejewski and Buttler, 2011] для дальнейшего улучшения результатов ранжирования информации.

|  |  |
| --- | --- |
| *2.1. Моделирование языка документа* | 23 |

поиск.

**2.1**  **Моделирование языка документа**

Подход языкового моделирования [Ponte and Croft, 1998; Song and Croft, 1999; Croft and Lafferty, 2003] является одной из основных основ использования тематических моделей в ИК-системах, поскольку он представляет собой эффективную вероятностную основу для изучения задач информационного поиска [ Понте и Крофт, 1998 г., Бергер и Лафферти, 1999 г.]. Статистическая языковая модель оценивает вероятность последовательностей слов, обозначаемых как *p* ( *w* 1 *, w* 2 *,* · · · *, w n* ). На практике модель статистического языка часто аппроксимируется моделями n-грамм. Модель униграммы предполагает, что каждое слово в последовательности независимо,

|  |  |
| --- | --- |
| *p* ( *w* 1 *, w* 2 *,* · · · *, w n* ) знак равно *p* ( *w* 1 ) *p* ( *w* 2 ) · · · *p* ( *w n* ) | (2.1) |

Модель триграммы предполагает, что вероятность текущего слова зависит только от двух предыдущих слов, и она представлена как

*п* ( *ш* 1 *,* · · · *, ш п* ) знак равно *п* ( *ш* 1 ) *п* ( *ш* 2 | *ш* 1 ) *п* ( *ш* 3 | *ш* 1 *, ш* 2 ) · · · *п* ( *ш п* | *ш п* −2 *, ш п* -1 ) *.* (2.2)

В приложении для поиска информации вероятность запросов оценивается с учетом вероятностной языковой модели, основанной на документе [Zhai and Lafferty, 2001a]. В частности, каждый документ рассматривается как образец языка, и языковая модель для каждого документа оценивается на основе терминов документа. Затем оценивается вероятность генерации запроса путем умножения вероятностей генерации каждого термина запроса с использованием языковой модели документа, и документы ранжируются на основе вероятности.

Учитывая образец документа *d* , простой способ оценить вероятность генерации слова *w* — использовать оценку максимального правдоподобия.

|  |  |  |
| --- | --- | --- |
| *р* мл ( *ш* | *д* ) = | *н д, ш* | (2.3) |
| *н д,* · |

где *n d,w* — частота встречаемости слова *w* в документе *d* , а *nd ,* · — общее количество токенов в документе *d* . Тогда вероятность

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 24 |  |  | *Специальный поиск информации* | | | | | | |
| генерация данного запроса *q* является | |  |  |  |  |  |  |  |  |
| Д | | | Д | | | *н д, ш* | |  |  |
| *п* ( *д* | *д* ) знак равно *ж* |  | *д р* ( *ш* | *д* ) знак равно |  |  |  | |  | *.* | (2.4) |
| е | *ж* | е | *д н д,* · | | |
|  |  |  |  |  |  |  |  |

Затем документы ранжируются на основе этой вероятности *p* ( *q* | *d* ). Более высокая вероятность означает, что соответствующий документ более релевантен данному запросу [Song and Croft, 1999]. Однако документ часто содержит ограниченное количество слов, и оценка максимального правдоподобия дает нулевую вероятность для этих невидимых слов. Если запрос содержит какое-либо слово, которого нет в документе, вероятность генерации всего запроса по данному документу равна нулю, что может привести к отбрасыванию совершенно хороших документов.

Эта проблема разреженности данных может быть решена путем сглаживания, которое присваивает отсутствующим членам некоторую ненулевую вероятность. Еще одно решение, дающее и другие преимущества, — тематические модели. Они обеспечивают уникальный способ извлечения вероятностей слов с учетом корпуса, который можно использовать для сглаживания языковых моделей документа. Мы суммируем два простых метода сглаживания, а затем показываем, как тематические модели вписываются в эту структуру сглаживания.

Существует два основных направления сглаживания: **интерполяция** [Je-linek and Mercer, 1980, Mackay and Petoy, 1995, Ney et al., 1994, Ponte and Croft, 1998, Zhai and Lafferty, 2001a] и **откат** [Katz, 1987, Песня и Крофт, 1999]. Метод на основе интерполяции сбрасывает количество просмотренных слов и распределяет дополнительное количество как видимых, так и невидимых слов. Альтернативная стратегия сглаживания отсрочки доверяет оценке максимального правдоподобия для слов с большим количеством слов, дисконтирует и перераспределяет массу только для менее распространенных слов [Zhai and Lafferty, 2001a].

Здесь мы рассмотрим два популярных и простых метода интерполяционного сглаживания, которые дополнительно расширяются с помощью тематических моделей для сглаживания языковых моделей документа.

**Jelinek-Mercer** Метод Jelinek-Mercer [Jelinek and Mercer, 1980] представляет собой линейную интерполяцию модели максимального правдоподобия в документе с моделью, основанной на всем корпусе, а коэффициент *λ* объединяет

|  |  |
| --- | --- |
| *2.2. Применение тематических моделей к языковым моделям документа* | 25 |
| две части: |  |
| *p* ( *w* | *d* ) = (1 − *λ* ) *p* ml ( *w* | *d* ) + *λp* ( *w* | C) *,* | (2.5) |

где C обозначает весь корпус. Эта простая смесь решает проблему разреженности данных. Для терминов, которые встречаются в документе *d* , оценка максимального правдоподобия (уравнение 2.3) не является точной, учитывая ограниченный размер документа, поэтому она сглаживается с более надежной вероятностью на уровне корпуса. Для отсутствующего термина *w* в документе *d* вероятность генерации слова *w* больше не равна нулю, а возвращается к вероятности уровня корпуса *p* ( *w* |C). Этот метод сглаживания был исследован и успешно применен в задачах информационного поиска [Ponte and Croft, 1998; Song and Croft, 1999].

**Байесовское сглаживание с использованием априорных уравнений Дирихле. Языковую** модель можно рассматривать как дискретное распределение, поэтому ее можно сгладить, применяя распределение Дирихле в качестве сопряженного априорного распределения [Mackay and Petoy, 1995]. Мы сделали аналогичное наблюдение в предыдущей главе, сравнивая уравнение 1.1 и уравнение 1.2; та же интуиция может быть расширена через несколько слоев дискретных распределений с априорными значениями Дирихле для создания модели сглаживания. Интуитивно это сглаживание добавляет дополнительный предварительный счет для каждого слова, чтобы сгладить вероятность невидимых слов.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *п* | *ж* |  | *г* | *n d,w* + *βp* ( *w* |C) | | | *,* |  |
| ( |  | | |  | знак равно |  |  |  | (2.6) |
|  |  | P *v* ∈ *V n d,v* + *β* |  |

где априор Дирихле определяется параметром концентрации *β* и вероятностями на уровне корпуса *p* ( *v* | C),

|  |  |
| --- | --- |
| ( *βp* ( *v* 1 | C) *, βp* ( *v* 2 | C) *,* · · · *, βp* ( *v n* | C)) *.* | (2.7) |

**2.2**  **Применение тематических моделей к языковым моделям документа**

Тематические модели, которые моделируют каждый документ как смесь тем и каждую тему как смесь слов, предлагают интересную основу для моделирования документов при поиске информации. Популярные тематические модели, такие как вероятностный латентный семантический анализ (plsa) и латентное распределение Дирихле (lda), были исследованы для улучшения моделей языка документов.

26 *Специализированный поиск информации*

Хофманн [1999b] вводит plsa для изучения отношений между словами запроса и документами, а условная вероятность слова запроса *w* для данного документа *d* вычисляется как маргинализация всех тем *k* ,

|  |  |
| --- | --- |
| Икс | (2.8) |
| *п* ТМ ( *ш* | *d* ) знак равно *п* ( *ш* | *k* ) *п* ( *k* | *d* ) |

*к*

Следуя этой идее, Wang et al. [2013] дополнительно добавлены регуляризации — изменение формы распределения, чтобы оно было более или менее рассредоточенным — для представлений темы документа, что полезно для поиска. Вместо использования plsa Wei и Croft [2006] применили ту же идею для изучения сглаженного по теме распределения слов документа с помощью lda. Восецкий и др. [2014] также изучают lda для языковых моделей документов в поиске в Твиттере.

Поскольку апостериорные оценки для тем сглаживаются априорными оценками Дирихле, тематические модели лучше изучают и сглаженные семантические отношения между словами документа и документами. В результате, несмотря на то, что этот подход ослабляет связь между словами запроса и документами, он является хорошим подходом для дополнения исходных языковых моделей документов. Таким образом, Wei и Croft [2006] далее предлагают объединить модель документа на основе lda с исходной моделью документа (уравнение 2.5) посредством линейной интерполяции,

*p* ( *w* | *d* ) = *λ* 0 (1 − *λ* ) *p* ml ( *w* | *d* ) + *λp* ( *w* | C) + (1 − *λ* 0 ) *p* TM ( *w* | *d* ) (2.9)

где *λ* 0 — коэффициент, объединяющий модель документа на основе LDA с общей моделью сглаженного языка.

Следуя Wei and Croft [2006], Lu et al. [2011] дополнительно оценивают эффективность применения тематических моделей в рамках модели языка документа. Вместо объединения с языковой моделью со сглаживанием Елинека-Мерсера (уравнение 2.5) Лу и соавт. [2011] сглаживание модели языка документа с помощью байесовского сглаживания (уравнение 2.7) и окончательная линейная комбинация с тематическими моделями.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *п* | *ж* | | | *г* | знак равно | *λ* | *n d,w* + *βp* ( *w* | C) | +(1- | *λ р* | ТМ ( | *ж* | | | *г* | ) | (2.10) |
|
| P *v* ∈ *V n d,v* + *β* |
| ( |  |  |  | ) |  |  |

При использовании разных стратегий сглаживания оба подхода применяют тематические модели для связи слов запроса с документами через

|  |  |
| --- | --- |
| *2.3. Расширение запроса при поиске информации* | 27 |

скрытые темы. Как показано в примере, показанном в Wei and Croft [2006], при запросе «рычаг выкупа» соответствующий документ говорит о «дефолтах подразделения Farley по ссуде выкупа Pepperell» без точного слова «леверидж», таким образом, рейтинг этого соответствующего документа очень низкий. Тем не менее, тематические модели связывают этот документ с двумя темами, которые тесно связаны с термином «кредитное плечо»: одна экономическая тема содержит такие слова, как «миллион», «компания» и «банкротство», а другая тема денежного рынка связана с « связь". Поскольку изучена лучшая семантическая связь между запросом и документом, этот релевантный документ оценивается намного выше, и производительность поиска улучшается.

**2.3**  **Расширение запроса при поиске информации**

Модели языка документов в информационном поиске [Ponte and Croft, 1998] пытаются смоделировать процесс генерации запросов на основе моделей документов. Однако большой проблемой является то, что эти модели отказываются от моделирования релевантности запроса-документа в явном виде [Lavrenko and Croft, 2001], что важно в традиционных задачах поиска информации.

На самом деле запросы, обычно краткие и использующие неформальный язык пользователей, значительно отличаются от языка документов [Мюллер, Гуревич, 2009]. Этот семантический пробел или лексический пробел может привести к кажущейся плохой релевантности запроса к документу, даже если документ является весьма релевантным с точки зрения пользователей. Например, набрав в поисковой системе «продукты Apple», разочарованный пользователь может добавить к запросу слово «компьютер», столкнувшись с экраном результатов поиска на основе фруктов.

Расширение запроса пытается автоматически имитировать аналогичный процесс, чтобы предотвратить это разочарование. Расширение запроса обычно анализирует отношения между словами запроса и другими словами и пытается найти потенциально связанные слова, чтобы исходный запрос был лучше представлен; тем самым повышая релевантность запроса к документу. Например, без особого контекста трудно понять запрос «dtd amc» [Jiang et al., 2016]. Путем расширения запроса можно построить связь между «dtd» и «диснейлендом в центре города», что более

28 *Специализированный поиск информации*

помогает в поиске документов. Далее в этом разделе рассматриваются классические структуры расширения запросов в информационном поиске, а соответствующие работы по использованию тематических моделей для расширения запросов представлены в следующем разделе.

**2.3.1**  **Изучение отношений слово-запрос для расширения запроса**

Существует два основных этапа расширения запроса. Первый шаг — найти взаимосвязь между запросами и словами и выбрать наиболее подходящие слова для расширения запроса. Второй шаг — применить расширенные запросы для ранжирования и вычислить окончательные оценки релевантности ранжирования. Начнем с первого шага. Исследовались два основных направления: модели языка запросов [Zhai and Lafferty, 2001b] и модели релевантности [Lavrenko and Croft, 2001].

**Модель языка запросов** Чтобы изучить отношение слова-запроса, Чжай и Лафферти [2001b] построили модель языка запросов для оценки вероятности *p* ( *w* | *q* ) слова *w* при заданном запросе *q* . Однако выучить хорошую модель языка запросов непросто, поскольку содержание запроса слишком ограничено.

Чжай и Лафферти [2001b] предлагают использовать как содержание запроса, так и соответствующие документы F (иногда называемые документами обратной связи).

ˆ

или выбранные документы) для оценки модели языка запросов. Пусть *θ* F будет оценочной моделью языка запросов, основанной на соответствующих документах.

ˆ

а *θ Q* — исходная модель языка запросов, оцененная исключительно на основе

ˆ

запросы, комбинированная модель запроса *θ Q* 0

|  |  |  |  |
| --- | --- | --- | --- |
| ˆ | ˆ | ˆ | (2.11) |
| *Q 0* \_ | = (1 − *λ* ) *θ Q* + *λθ* F | |
| ˆ |  | ˆ | так же может быть |
| Оценка *θ Q* очевидна на основе слов запроса, а *θ* F | | |

просто оценивается моделью языка униграмм *θ* , которая генерирует каждое слово в F независимо. Однако большинство документов содержат не только информацию, относящуюся к запросу, но и справочную информацию. В результате Чжай и Лафферти [2001b] предлагают генерировать соответствующий документ с помощью смешанной модели, которая сочетает в себе языковую модель *p* ( *w* | *θ* ) с языковой моделью коллекции *p* ( *w* | C) и логарифмическую вероятность в

|  |  |
| --- | --- |
| *2.3. Расширение запроса при поиске информации* | 29 |
| соответствующий документ, |  |
| журнал *p* (F | *θ* ) = ХХ *n d i ,w* log((1 − *λ* 0 ) *p* ( *w* | *θ* ) + *λ* 0 *p* ( *w* | C)) | (2.12) |
| *я* |  |

где *nd i* , *w* — количество слов *w* в документе *d i .* Благодаря объединению информации из соответствующих документов модель языка запросов становится более надежной, поэтому отношения между запросом и словом лучше представлены.

**Модель релевантности** В отличие от подхода, основанного на модели языка запросов, Лавренко и Крофт [2001] предполагают, что и запрос, и соответствующие документы являются случайными выборками из неизвестной модели релевантности *R* . Учитывая запрос *q* , они аппроксимируют вероятность *p* ( *w* | *R* ) на основе наблюдаемого запроса *q* как

|  |  |  |
| --- | --- | --- |
| *п* ( *ш* | *р* ) ≈ *п* ( *ш* | *q* ) знак равно | *р* ( *ш , д* ) | (2.13) |
| *р* ( *ц* ) *.* |

Чтобы оценить совместную вероятность *p* ( *w, q* ), Лавренко и Крофт [2001] предполагают, что слово *w* и запрос *q* выбираются независимо из одного и того же распределения, например, из распределения униграмм, тогда совместная вероятность равна

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *р* ( *ш , д* ) = | Х е С | | | Х е С | (2.14) |
| *p* ( *d* ) *p* ( *ш , q* | *d* ) = | | | *п* ( *d* ) *п* ( *ш* | *d* ) *п* ( *q* | *d* ) *.* |
|  | *г* | | | *г* |  |
| Тогда дискретное распределение *p* ( *w* | *q* ) для данного запроса *q* равно | | | | |  |
|  |  | *р* ( *ш , д* ) | | Х е С |  |
|  |  |  |  |  |
| *р* ( *ш* | *д* ) знак равно *р* ( *q* ) = | | | | (2.15) |
| *п* ( *ш* | *d* ) *п* ( *d* | | *q* ) *.* |
|  |  |  |  | *г* |  |

Как модель языка запросов, так и модель релевантности фиксируют взаимосвязь между запросом и другими словами, на основе которых можно выбрать наиболее релевантные слова для расширения запроса.

**2.3.2**  **Релевантность ранжирования с расширением запроса**

Учитывая родственные слова для расширения запроса, следующий вопрос заключается в том, как применять расширенные запросы для вычисления итоговой оценки релевантности ранжирования. Комбинация может произойти либо до, либо после вычисления показателя релевантности.

30 *Специализированный поиск информации*

Чжай и Лафферти [2001b] объединяют расширенный язык запросов

модель *θ* F с исходной моделью языка запросов *θ Q* как один запрос

ˆ

языковая модель *θ Q* 0 (уравнение 2.11). Учитывая запрос *q* , сгенерированный из

| ˆ

расширенная модель запроса *p* ( *q θ Q* 0 ) и сгенерированный документ *d* | ˆ

из модели документа *p* ( *d θ D* ), они сравнивают, насколько похожи распределения тем между этими двумя «документами».

В отличие от Zhai и Lafferty [2001b], Lavrenko and Croft [2001] вычисляют показатель релевантности, используя исходный запрос и расширенный запрос соответственно, а затем линейно комбинируют два показателя. Таким образом, окончательная релевантность запроса-документа *s* ˆ d ( *q* ) вычисляется как

|  |  |
| --- | --- |
| *s* ˆ d ( *q* ) *знак равно λs d* ( *e* ) + (1 − *λ* ) *s d* ( *q* ) | (2.16) |

где *sd* ( *q* ) — релевантность между исходным запросом *q* и документами *d* , а *sd (* e *) — релевантность между терминами* расширенного запроса *e* идокумент *д* .

**2.4**  **Применение тематических моделей для расширения запроса**

Тематические модели фиксируют семантические отношения слов посредством изучения скрытых тем, которые представлены в виде распределений по разным словам. Такие семантические отношения между словами обеспечивают уникальный способ сопоставления или расширения слов на семантическом уровне, а не путем прямого сопоставления правописания. Например, по короткому запросу «диабет» тематические модели могут легко найти родственные слова, такие как «инсулин», «глюкоза», «коронарный» и «метформин» и т. д., поскольку они часто встречаются в одном и том же контексте. Цзэн и др., 2012]. В результате тематические модели успешно применялись для расширения запросов [Yi and Allan, 2009; Park and Ramamohanarao, 2009; Zeng et al., 2012].

**Сглаживание модели языка запросов** Наиболее интуитивный способ использования тематических моделей для расширения запросов состоит в том, чтобы извлекать релевантность слов непосредственно из тем, как Yi and Allan [2009]. Они обучают модель темы, из которой изучается вероятность *p* TM ( *k* | *q* ) темы *k* в запросе *q* . Затем релевантность слова-запроса *p* ( *w* | *q* ) вычисляется на основе тем:

Икс

*п* ( *ш* | *q* ) знак равно *п* ТМ ( *ш* | *k* ) *п* ТМ ( *k* | *q* ) *.*  (2.17)

*к*

|  |  |
| --- | --- |
| *2.4. Применение тематических моделей для расширения запроса* | 31 |

Эта релевантность слова запроса *p* ( *w* | *q* ) из тематических моделей сглаживает исходную модель языка запросов посредством линейной интерполяции. Однако запросы, как правило, слишком короткие, чтобы изучить важные темы, поэтому качество релевантности слова запроса относительно ограничено. Чтобы улучшить качество извлеченных тем, Йи и Аллан [2009] также обучают тематические модели из соответствующих документов (например, основных документов, извлеченных с помощью запроса) и извлекают отношения запрос-слово на основе уравнения 2.17 для расширения запроса.

**Улучшение модели релевантности** В дополнение к этому прямому подходу Йи и Аллан [2009] также применяют тематические модели для улучшения модели релевантности в уравнении 2.15 для расширения запроса. В этом подходе тематические модели фиксируют отношение документ-слово *p* ( *w* | *d* ) при заданном запросе *q* как

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *pTM* ( w *| d, q* ) = | | | | Икс |  |  |  |  |  |  | (2.18) |
|  | *р* ( *ш* | *к* ) *р* ( *к* | *д, д* ) | | | | |  |  |
|  |  |  |  |  | *к* |  |  |  |  |  |  |  |
| куда, |  |  |  |  |  |  |  |  |  |  |  |  |
| *п* | *к* | *д, д* |  | *п* ( *к* | *d* ) *п* ( *д* | *к* ) | |  | *п* | *кдп* | *д* | *к* |  |  |
| ( | | |  | знак равно |  |  | ≈ | ( | | ) ( | | |  | ) | (2.19) |
|  | *п* | *qd* |  |
|  |  |  |  | ( | | ) |  |  |  |  |  |  |  |

где *p* ( *k* | *d* ) — вероятность темы в документе *d* , а *p* ( *q* | *k* ) — вероятность генерации запроса *q* по теме *k* . Затем отношение документ-слово на основе темы *pTM (w | d, q) применяется для сглаживания отношения документ-слово p* ( w *|* d *)* в *модели* релевантности *(* уравнение *2.15* ) посредством линейной интерполяции,

|  |  |  |
| --- | --- | --- |
| Х С | − *λ* ) *p* TM ( *w* | *d, q* )) *p* ( *d* | *q* ) | (2.20) |
| *p* ( *w* | *q* ) = ( *λp* ( *w* | *d* ) + (1 |

*г*

где *λ* — постоянный вес для объединения исходной модели релевантности и модели релевантности на основе темы. Поскольку тематические модели охватывают отношения слов на семантическом уровне, эта усовершенствованная модель релевантности лучше фиксирует отношения запросов и слов и улучшает расширение запроса.

**Изучение парных словесных отношений** Парк и Рамамоханарао [2009] также применяют тематические модели для расширения запросов, но в другом аспекте.

32 *Специализированный поиск информации*

путь. Они моделируют парные отношения между словами с помощью тематических моделей, а затем применяют их для расширения запроса. Более конкретно, на основе тем, извлеченных из тематических моделей, они вычисляют вероятностные отношения каждой пары слов ( *w x , w y* ),

Икс

*p* ( *w x* | *w y , α* ) = *p* ( *w x , k* | *w y , α* ) (2.21)

*к*

Икс

* *р* ( *ш х* | *k, α* ) *р* ( *к* | *ш у , α* )

*к*

где *α* — параметр концентрации априора Дирихле для распределений по темам документов, а *p* ( *w x* | *k, α* ) — вероятность слова *w x* в теме *k* , которое можно узнать из тематических моделей, а *p* ( *k* | *w y , а* ) есть

*p* ( *w y* | *k, α* ) *p* ( *k* | *α* )

*р* ( *к* | *ш у , α* ) = П *к* 0 *p* ( *w y* | *k* 0 *, α* ) *p* ( *k* 0 | *а* )

где *p* ( *w y* | *k, α* ) — вероятность слова *w y* в теме *k* . Park и Ramamohanarao [2009] также показывают, что *p* ( *k* | *α* ) = P *α k* . В результате мы

*Дж α j*

имеют,

*p* ( *w y* | *k, α* ) *α k*

*р* ( *к* | *ш у , α* ) = П *к* 0 *p* ( *w y* | *k* 0 *, α* ) *α k* 0

Окончательные вероятностные отношения каждой пары термов можно представить в виде

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *п* ( *ш Икс* | *ш у , α* ) знак равно | П *к*Р *к х* 0 *р* | ( *ш у* | *к* 0 *, α* ) | *ак* 0 *\_* | | *к* | (2.22) |
|  |  | *p* ( *wk, α* ) *p* ( *w y k, α* ) *α* |  |  |

Как только это отношение терминов получено, они выбирают наиболее связанные термины в качестве расширенных терминов *e* для данного запроса *q* , и окончательный рейтинг документа вычисляется по уравнению 2.16.

**Интерактивная обратная связь Обратная связь** по релевантности вовлекает пользователей в процесс поиска для улучшения набора результатов ранжирования [Rocchio, 1971]. Основная идея состоит в том, чтобы попросить пользователей дать отзыв о релевантности документов в начальном наборе результатов поиска, а отзывы пользователей о релевантности в дальнейшем используются для улучшения результатов ранжирования. Этот процесс может проходить через одну или несколько итераций.

|  |  |
| --- | --- |
| *2.4. Применение тематических моделей для расширения запроса* | 33 |

**Таблица 2.1:** По запросу «еврооппозиция» семь тем выбираются и показываются пользователю. Пользователь выбрал тему 79 в качестве темы для обратной связи (пример из Andrzejewski and Buttler [2011]).

|  |  |
| --- | --- |
| Тема | Условия |
|  |  |
| 196 (дебаты) | Евроскептики Тори, социальная глава, либерал-демократ, депутаты, |
|  | Труд, законопроект, Commons |
| 404 (ратификация) | ратификация Масстрихтского договора, Пол Шлютер, Пол Расмуссен, |
|  | Датский, голосование, Дания, ЕС |
| **79 (эму)** | экономический валютный союз, Масстрихтский договор, государства-члены, |
|  | Европейская, Европа, Сообщество, Эму |
| 377 (Джордж) | Президент Джордж Буш, Белый дом, г-н Клинтон, администрация |
| 115 (мощность) | законопроект о регулировании, воскресные торги, речь королевы, закон, законодательство, |
|  | правительство, действие |
| 446 (лет) | председатель исполнительный директор, управляющий директор, финансовый директор, |
|  | Сэр, правление, группа, компания |
| 431 (кабинет) | Мистер Джон Мейджор, премьер-министр, мистер Мейджор, партия, консерваторы, правительство. |
|  | мент, консерватор |

Andrzejewski и Buttler [2011] представляют новую структуру для получения и использования отзывов пользователей на уровне латентной темы. Они изучают скрытые темы из всего корпуса и строят осмысленные представления тем. Во время запроса они решают, какие скрытые темы потенциально актуальны, и представляют представления тем в результатах поиска по ключевым словам. Когда пользователь выбирает скрытую тему, исходный запрос дополняется верхними словами в этой выбранной теме, а результаты поиска уточняются. Andrzejewski, Buttler [2011] используют в качестве примера запрос «еврооппозиция»: пользователи хотят найти документы о противодействии введению единой европейской валюты. 500 тем изучаются с использованием суждений корпуса и релевантности. Семь из 500 тем выбираются для показа пользователям, как показано в таблице 2.1, и пользователь выбирает тему 79 в качестве отзыва пользователя. Используя лучшие термины в теме 79 в качестве расширенных терминов запроса, ранжирование релевантных документов улучшается.

Это направление связано с маркировкой тем и интерактивной визуализацией для тематических моделей, которые будут обсуждаться далее в главе 3.

34 *Специализированный поиск информации*

**2.5**  **Помимо релевантности — персонализация поиска**

Традиционные поисковые системы извлекают документы только на основе запросов, независимо от того, кто отправил запросы. По мере появления большего количества веб-страниц запросы обычно становятся слишком короткими, чтобы выразить потребности пользователей, и пользователи могут предпочесть разные результаты, даже если входные запросы одни и те же [Jansen et al., 2000, Dou et al., 2007]. Как показано в примерах Dou et al. [2007], запрос «мышь» для биологов может означать «грызуны», а программисты могут использовать тот же запрос для поиска компьютерной периферии. Даже для недвусмысленных запросов, например «онлайн-покупки», некоторые пользователи могут предпочесть « [www.amazon.com](file:///C:\temp\Rar$DIa0.245\www.amazon.com) », а другие — « [www.ebay.com](file:///C:\temp\Rar$DIa0.245\www.ebay.com) ».

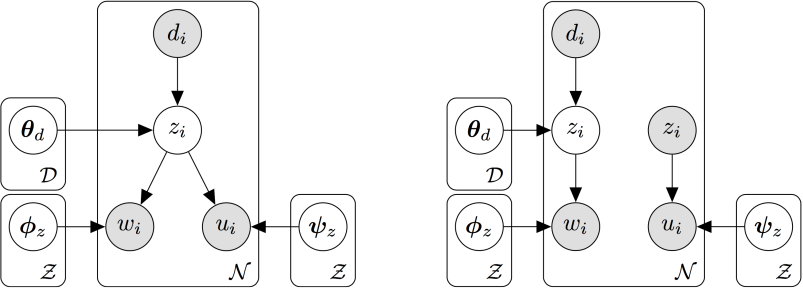
Понимание предпочтений и контекста пользователей помогает удовлетворить их информационные потребности. Таким образом, их системы должны адаптировать результаты ранжирования [Pitkow et al., 2002, Micarelli et al., 2007]. Это называется персонализированным поиском или персонализацией поиска.

Существует несколько способов персонализации поиска, и два основных направления обычно называют контекстуализацией [Melucci, 2012] и индивидуализацией [Pitkow et al., 2002]. Первый заключается в рассмотрении условий пользователей в поисковой деятельности, например времени и местоположения. Последний больше фокусируется на индивидуальных характеристиках и действиях пользователей, что также называется профилем пользователей. Тематические модели были исследованы для моделирования предпочтений пользователей в тематическом пространстве [Song et al., 2010, Carman et al., 2010].

**Моделирование предпочтений пользователей с помощью выходных тем** Song et al. [2010] применяют тематические модели для моделирования предпочтений пользователей на основе истории поиска пользователей. Идея очень похожа на сглаживание моделей языка запросов *p* ( *w* | *q* ) по тематическим моделям, как объяснено в уравнении 2.17. Оценка *p* ( *w* | *q* ) разбивается на две части *p* ( *w* | *k* ) и *p* ( *k* | *q* ).

Для запроса каждого пользователя они объединяют выбранные документы (или первые *n* документов в рейтинге, если щелчка не произошло) в большой документ предпочтений. Затем тематическая модель plsa применяется к набору предпочтений для извлечения скрытых тем в качестве предпочтений пользователей ( *p* ( *w* | *k* ) в уравнении 2.17). Вторая часть заключается в оценке распределения запросов и тем *p* ( *k* | *q* ) в уравнении 2.17. Однако запросы слишком

|  |  |
| --- | --- |
| *2.5. Помимо релевантности — персонализация поиска* | 35 |



**(а)**  **(б)**

**Рисунок 2.1:** Диаграммы пластин для персонализированного поиска на рисунке 2.1a и реальная упрощенная модель, используемая для оценки параметров на рисунке 2.1b (оба рисунка из Harvey et al. [2013]).

short для прямой оценки темы запроса. Вместо этого они сначала оценивают языковую модель *θ q* из документа с большими предпочтениями запроса *q* и сравнивают косинусное сходство по каждой теме, чтобы оценить распределение запросов и тем *p* ( *k* | *q* ),

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *р* ( *к* | *д* ) знак равно | *р* ( *к, д* ) | | ≈ |  | sim ( *θk , θq )* | | (2.23) |
|  |  |  |  | *.* |
| Р *к р* ( *к, д* ) | P *k* sim ( *θk , θq )* |

Затем это может направить пользователя к документам, которые соответствуют интересам пользователя (например, которые хорошо соответствуют этим темам).

**Кодирование пользователей в тематические модели** Carman et al. [2010] также исследуют тематические модели в больших журналах запросов для персонализации поиска и предлагают топическую модель персонализации, как показано на рис. 2.1a. Идея заключается в распространении документа по темам, будут слова, выбранные случайным образом для создания запроса, и пользователи, которые решили щелкнуть этот документ.

Как показано на рис. 2.1a (здесь используется формализм диаграммной пластины, объясненный в главе 1.5.3), эта модель имеет три наблюдаемые переменные: документ *d i* , слово *запроса wi и* пользователь *ui* . Для данной темы *k ,* выбранной из дискретного распределения *θ d* , соответствующие слова запроса *w i* выбираются из дискретного распределения слов темы *φ k* , а пользователь *u i* , подавший соответствующий запрос, выбирается из дискретного распределения пользователей темы.

36 *Специализированный поиск информации*

распределение *ψ k* .

Чтобы оценить вероятность *p* ( *k* | *w i , d i , u i* ), они дополнительно предполагают условную независимость между словом *w i* , пользователем *u i и* документом *di* с учетом темы *k* , и модель может быть упрощена. в качестве

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *п* | *к* | | | *ж* | *, д , у* | *я* ) = | *р* ( *к , ш я , и я* | *d я* ) | ∝ | *pw* | *к* | *пу* | *к* | *п* | *к* | | | *г* | *я* ) | *.* | (2.24) |
|
| *п* ( *ш я , ты я* | *d я* ) |
| ( |  | *я* | *я* | ( *я* | | ) | ( *я* | | ) | ( |  |  |  |

Путем непосредственного включения пользователя в модель темы эта модель предполагает, что тематические интересы пользователя для описания документа, который пользователь щелкнул, так же важны, как и слова для описания документа. Это предположение слишком сильное [Carman et al., 2010]. В результате Харви и соавт. [2013] также предлагают игнорировать пользователя во время вывода и упростить модель (рис. 2.1b). В этой модели темы используются для вывода распределения тема-пользователь *p* ( *u i* | *k* ) после сходимости цепи Маркова. Интуиция заключается в том, чтобы уловить идею о том, что пользователь нажимает на документ по конкретному запросу из-за его/ее интересов, выраженных в тематическом пространстве [Harvey et al., 2013].

На основе оценок этой персонализированной тематической модели документы ранжируются по вероятности с учетом запроса и пользователя следующим образом:

|  |  |  |
| --- | --- | --- |
| Д | *р* ( *ш , и* | *д* ) | (2.25) |
| *р* ( *д* | *д , и* ) ∝ *р* ( *д* ) |
| *w* ∈ *q* |  |  |
| YX | |  |
| ∝ *р* ( *д* ) | *п* ( *ш* | *к* ) *п* ( *ты* | *к* ) *п* ( *к* | *д* ) |  |
| *w* ∈ *q* | *к* |  |

Тонко включив профили пользователей в алгоритмы ранжирования, Harvey et al. [2013] значительно улучшили персонализированные ранжированные списки документов по сравнению с неперсонализированными базовыми уровнями.

**2.6**  **Резюме**

Поскольку тематические модели анализируют документы на семантическом уровне, они предлагают интересную и уникальную основу для моделирования отношений между словами, а также между документами и словами. В результате тематические модели успешно применяются для сглаживания языковых моделей, расширения запросов и персонализации поиска.

|  |  |
| --- | --- |
| *2.6. Резюме* | 37 |

Также есть работы по диверсификации результатов поиска [Dang and Croft, 2013; Santos et al., 2015]. Целью диверсификации результатов поиска является выявление различных аспектов неоднозначного запроса, извлечение документов для каждого аспекта и обеспечение того, чтобы результаты поиска содержали более разнообразные документы [Dang and Croft, 2013]. Персонализация и диверсификация ортогональны и могут сочетаться в унифицированных моделях [Vallet and Castells, 2012; Liang et al., 2014].

Эти успешные приложения для поиска информации основаны на хорошем понимании результатов тематических моделей. В следующей главе мы расскажем, как визуализировать, маркировать и оценивать темы, чтобы помочь пользователям лучше понять темы и то, как они выражены в коллекциях.

**3**

**Оценка и интерпретация**

В то время как предыдущая глава посвящена алгоритмическому использованию тематических моделей, одна из причин использования тематических моделей заключается в том, что они создают удобочитаемые сводки тем больших коллекций документов. Однако, чтобы пользователи могли использовать результаты тематических моделей, они должны быть в состоянии понимать выходные данные моделей. Это зависит от *визуализации модели* , *взаимодействия* и *оценки* .

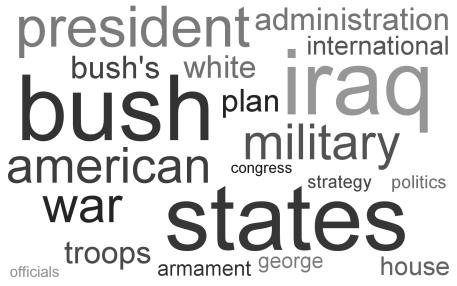
Мы начинаем эту главу с обсуждения того, как лучше показать отдельные темы пользователям. От этих основ мы переходим к тому, как мы можем отображать целые модели — со многими темами — для пользователей. Наконец, мы заканчиваем тем, как пользователи могут предоставлять обратную связь через эти интерфейсы для обнаружения ошибок и улучшения базовой модели.

**3.1**  **Отображение тем**

Вспомним из предыдущих глав, что темы — это распределения по словам; слова с наибольшим весом в теме лучше всего объясняют, о чем эта тема. Хотя самый простой ответ — просто показать наиболее вероятные слова — является распространенным решением, существуют уточнения, которые могут улучшить понимание пользователем коллекции, показывая отношения

38

|  |  |
| --- | --- |
| *3.1. Отображение тем* | 39 |



**Рисунок 3.1.** Облака слов используют двухмерный макет, чтобы показать, какие слова появляются в теме. Размер слова связан с его вероятностью в теме, показывая, какие слова более заметны.

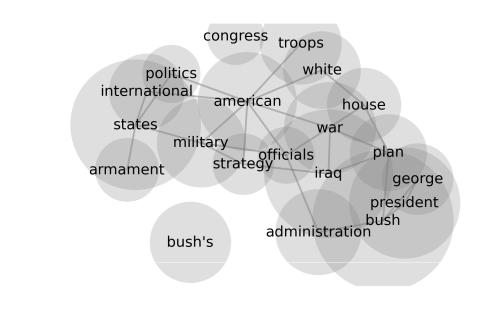
между словами или явно показывая вероятность слов.

**Списки слов** Просто показать список наиболее распространенных слов (визуализация, которую мы будем называть «списком слов») очень просто и работает хорошо. Пользователи могут быстро понять, как расположены слова, и это позволяет эффективно использовать пространство. Темы были представлены горизонтально [Gardner et al., 2010, Smith et al., 2015] или вертикально [Eisenstein et al., 2012, Chaney and Blei, 2012], с запятыми, разделяющими отдельные слова, или без них, или с использованием условных обозначений. [Чейни и Блей, 2012]. Смит и др. [2015] пошли еще дальше, добавив столбцы, представляющие вероятности слова.

**Облака** слов (например, рис. 3.1) — еще один популярный способ отображения тем. В отличие от списков слов, они также используют размер слов для передачи дополнительной информации. Облака слов обычно используют размер слов, чтобы отразить вероятность слов. Это использует большую часть заданной области визуализации для отображения темы.

Однако облака слов подвергались критике за плохую поддержку визуального поиска [Viégas and Wattenberg, 2008] и отсутствие контекстуальной информации между словами [Harris, 2011]; пользователи могут иногда проводить ложные связи между словами, которые случайно расположены рядом друг с другом в облаке слов. Другой альтернативой является использование словесных ассоциаций для определения положения слов [Smith et al., 2014]; Рисунок 3.2 размещает слова

40 *Оценка и интерпретация*



**Рис. 3.2.** Визуализация «тема в коробке» для тем — например, облако слов — показывает слова в 2D-контексте. Однако он использует локальное совместное появление (появляются ли слова вместе в предложении), чтобы решить, какие слова разместить рядом друг с другом.

которые появляются вместе рядом друг с другом в визуализации.

**3.2**  **Маркировка тем**

На протяжении всего опроса мы обращались к темам с такими *ярлыками* , как информационные технологии или искусство . Это удобные дескрипторы, но полностью удаленные из исходного распределения по словам. Таким образом, часто полезно назначать метки темам в интерфейсе.

В отличие от предыдущих подходов к *визуализации* , маркировка фокусируется на отображении не исходных слов темы, а, скорее, более четкой метки, более похожей на то, что может дать человеческое резюме данных.

Подходы к автоматической маркировке можно разделить на те, которые используют только внутреннюю информацию из тематической модели, и те, которые также используют внешние ресурсы знаний. В то время как чисто внутренние методы более надежны и соответствуют философии неконтролируемых тематических моделей, внешние ресурсы часто дают более качественные метки.

Из методов, использующих внешние ресурсы, мы далее отделяем те, которые используют непосредственный контроль для навешивания ярлыков (т. е. знание того, что представляет собой правильное навешивание ярлыков), от тех, которые используют общие ресурсы знаний, такие как Википедия или базы знаний.

**Внутренняя маркировка**  Mei et al. [2007b] предлагают внутреннюю маркировку

|  |  |
| --- | --- |
| *3.2. Маркировка тем* | 41 |

метод, который берет известные фразы из темы и сравнивает, насколько контекст фразы соответствует распределению темы. Фразы, контекст которых очень похож на тему, часто появляются в областях текста, резюмирующих документ, что делает их хорошими кандидатами на ярлыки. Мао и др. [2012] распространили эту технику на иерархии, используя понимание того, что ярлыки родителей должны соответствовать ярлыкам их детей.

**Маркировка контролируемыми этикетками** Lau et al. [2010] используют контролируемый подход для переранжирования слов в теме, чтобы гарантировать, что пользователь увидит «лучшее» слово в теме. Каждое слово-кандидат формирует вектор признаков, состоящий из следующих признаков:

* условная вероятность слова при наличии других слов в теме (что подразумевает связность темы, как обсуждалось в главе 3.4);
* является ли слово гипернимом других слов в теме (например, «собака» в теме, которая также содержит слова «терьер» и «пудель»); а также
* исходная вероятность слова в теме.

Хотя их можно использовать отдельно в качестве неконтролируемого переранжирования, Lau et al. [2010] используют выбранные пользователями лучшие тематические слова, чтобы взвесить, какие из этих функций наиболее важны для выбора лучшего тематического слова. Эти веса изучаются с помощью регрессии опорных векторов. Лау и др. [2011] расширяют свою технику, добавляя в набор кандидатов из Википедии. Слабость этого подхода в том, что Википедия может не освещать темы коллекции; если Википедия игнорирует тему, захваченную моделью темы, то она не сможет найти подходящую метку для этой темы.

**Маркировка с помощью баз знаний** Mao et al. [2012] согласовывают тематические модели с внешней онтологией ярлыков. Они утверждают, что метки должны соответствовать словам темы (как маркировка с плоскими темами); слова темы должны соответствовать дочерним элементам меток в иерархии; и метки темы должны быть уникальными.

Алетрас и др. [2014] вместо этого запросите всю сеть, а затем постройте график, включающий слова в заголовках найденных веб-страниц. Их цель — найти слова, которые являются «центральными» в графе: эти слова

42 *Оценка и интерпретация*

должно получиться хорошее название. Слова имеют границы между собой, если они кажутся близкими друг к другу больше, чем можно было бы ожидать случайно. Это свойство измеряется с помощью метрики нормализованной точечной взаимной информации (npmi). Они находят центральные слова с помощью алгоритма PageRank [Page et al., 1999], который находит слова, наиболее вероятные в теме и часто встречающиеся со многими другими словами в теме. Это тот же алгоритм, который используют поисковые системы для поиска страниц, имеющих «высокий авторитет» в Интернете.

Точно так же, как поисковая система должна возвращать simpsonsarchive.com для поиска по «Симпсонам», потому что все ссылаются на него, этот метод маркировки найдет слово в теме, за которое все остальные слова «голосуют». Для каждого слова в теме найдите все слова, которые, вероятно, также будут встречаться с этим словом, а затем возьмите победителя этих выборов в качестве ярлыка для темы. Например, по теме

клеточный ответ иммунный лимфоцитарный антиген цитокин Т-клеточный индуцирующий рецепторный иммунитет

алгоритм выбирает иммунную систему, поскольку она появляется рядом со многими другими терминами в теме [Aletras et al., 2014].

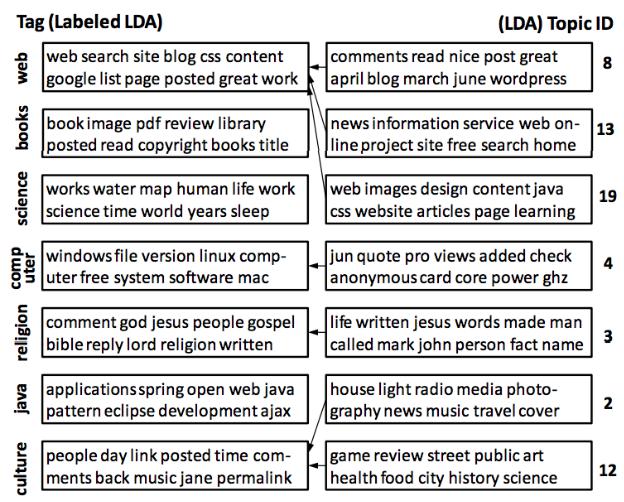
**Использование помеченных документов** Задача связывания меток с темами становится намного проще, если многие из ваших документов сами по себе имеют метки. Помеченный lda [Ramage et al., 2009] связывает темы с каждой из меток и заставляет помеченные документы использовать только темы, связанные с документом. Это ограничение заставляет темы соответствовать исходным меткам (рис. 3.3). Бакалов и др. [2012] распространяют это на иерархические наборы меток (например, темы ny Times, помещающие Россию в категорию международных ), в то время как Nguyen et al. [2014] расширили его до изучения иерархии тем от неорганизованных лейблов, узнав, что ска 1 — это вид музыки без предоставленных ссылок.

**3.3**  **Отображение моделей**

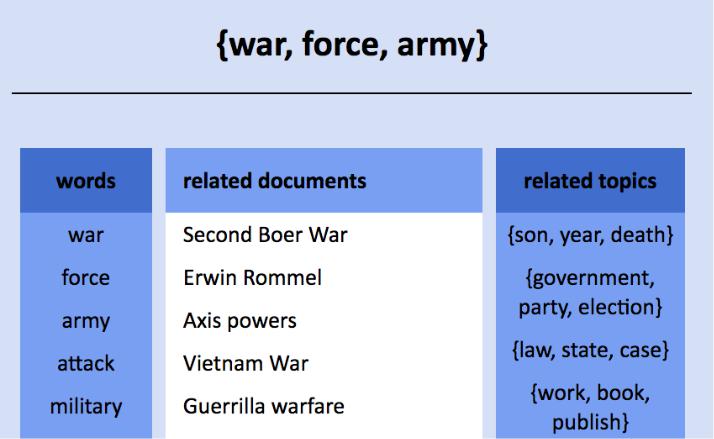
Тем не менее, темы не конец истории. Пользователи часто хотят использовать темы, чтобы найти соответствующие документы в коллекции. Возвращаясь к

1 Музыкальный жанр, знакомый любителям регги и крестовербалистам.

|  |  |
| --- | --- |
| *3.3. Отображение моделей* | 43 |



**Рисунок 3.3:** Пример тем, изученных с помощью lda (рисунок из Ramage et al. [2009]). Каждая тема в помеченном lda связана с меткой, что способствует тому, чтобы темы согласовывались с онтологией меток. lda, напротив, использует эмпирическую частоту тем для разделения коллекции, в результате чего три темы (8, 13, 19) связаны с помеченной веб- темой lda.



**Рисунок 3.4:** Механизм визуализации тематической модели [Chaney and Blei, 2012] показывает документы, наиболее связанные с темой, а также связанные темы.

44 *Оценка и интерпретация*

В нашем примере из предыдущей главы пользователь может захотеть найти «дымящийся пистолет» в корпусе Enron, а не просто использовать темы для понимания основных тем в коллекции.

Таким образом, хорошая визуализация тематической модели должна также отображать документы, связанные с темой. Механизм визуализации тематической модели [Chaney and Blei, 2012, tmve] показывает основные документы, связанные с темой (рис. 3.4). Напомним, что каждый документ имеет распределение по темам *θ d* , которое представляет собой вектор с записью для каждой темы. Мы фокусируемся на измерении, связанном с определенной темой, а затем сортируем документы на основе этой координаты темы от наибольшего к наименьшему.

Тематическое руководство [Gardner et al., 2010] расширяет этот подход, обогащая тематические представления дополнительными метаданными. Например, если в коллекции есть суммы в долларах или настроения [Pang and Lee, 2008], связанные с документом, она предоставляет гистограмму метаданных, связанных с темой. Он также предоставляет примеры тематических слов *в контексте* , позволяя увидеть, как слово используется в теме (помогает учесть предположения тематической модели о наборе слов).

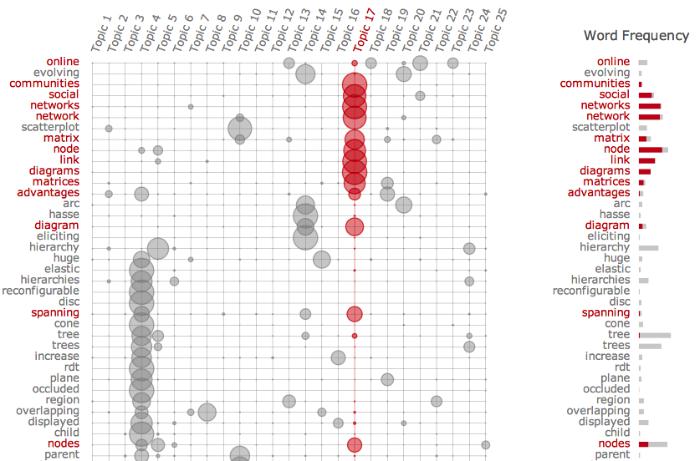
Интерактивная модель TOpic и метаданные [Eisenstein et al., 2014, Интерактивная модель TOpic и метаданные] фокусируются на конкретном типе метаданных: времени. Это позволяет пользователям просматривать эволюцию тем с течением времени, чтобы понять, например, как проблема рабства переформулируется из экономического аргумента в спор о правах человека. Он поддерживает фильтрацию по определенным темам или просмотр того, как слова используются с течением времени в разных темах.

Вместо того, чтобы показывать, как темы связаны с метаданными, Chuang et al. [2012] сосредоточиться на том, как темы связаны *друг с другом* . Их тематическая визуализация «Термит» (рис. 3.5) показывает почленное сходство между темами. Представляя вероятности тематического термина в сетке с темами в виде столбцов и терминами в виде строк, пользователи могут видеть, когда темы содержат общие слова или когда темы содержат только несколько слов.

**3.4**  **Оценка, стабильность и ремонт**

Визуализации могут помочь показать пользователям, где у тематических моделей есть проблемы. Тематические модели редко бывают идеальными, и качество может варьироваться в пределах модели. Даже в

|  |  |
| --- | --- |
| *3.4. Оценка, стабильность и ремонт* | 45 |



**Рисунок 3.5:** Визуализация тем в Termite помогает выявить, в каких темах используются похожие слова и, следовательно, они, вероятно, говорят о сходных вещах.

хороших моделей мы часто находим несколько плохо подходящих или неправильно объединенных тем.

В течение многих лет основным показателем для оценки качества тематической модели была предполагаемая вероятность модели [Wallach et al., 2009b]. Поскольку тематическая модель является генеративной вероятностной моделью, как и языковая модель [Chen et al., 1998], мы можем задаться вопросом, насколько хорошо модель может предсказать невидимый текст: запустите генеративный процесс вперед для документа и посмотрите, насколько хорошо он совпадает. с протянутым документом. Если модель хорошо справляется с использованием тем для предсказания того, какие слова появятся в новых документах, то это хорошая модель, а если нет, то это плохая модель.

В некотором смысле удержанная вероятность имеет смысл, но она неполна. Мы должны быть в состоянии обнаруживать некоторые виды сбоев: темы, которые являются просто случайным шумом, будут иметь низкую вероятность удержания. С другой стороны, сотни тем настолько специфичны, что любой выложенный документ хорошо смоделирован, что дает отличную вероятность вынесения. Тем не менее тем не хватало бы обобщаемости и интерпретируемости в глазах пользователя.

Чанг и др. [2009] показывают, что сдерживаемая вероятность, традиционная

46 *Оценка и интерпретация*

мера качества вероятностной модели, подчеркивает *сложность* , а не простоту интерпретации, которую ищут пользователи. Пользовательские оценки того, как хорошие темы отрицательно коррелируют с вероятностью задержания: более сложная модель (например, модель с более сложными уравнениями или модель с большим количеством тем) может лучше соответствовать случайному задержанному документу. Однако более сложные модели больше запутывают пользователей.

Автоматизированные измерения [Newman et al., 2010, Mimno et al., 2011, Lau et al., 2014] качества темы могут служить косвенным показателем для оценок интерпретируемости человеком. Однако эти подходы могут не сказать вам, подходит ли тематическая модель для конкретного приложения, какие части модели надежны и почему. Демонстрация взаимосвязей между несколькими моделями также может помочь отличить стабильные темы от ложных [Chuang et al., 2015], а корректировка «гиперпараметров» распределений (параметров Дирихле моделей, обсуждаемых в главе 1) может иметь большое влияние на то, какие окончательными моделями являются [Wallach et al., 2009a].

Танг и др. [2014] предоставляют руководство по диагностике того, какие свойства набора данных могут привести к сбою тематической модели: несоответствие между количеством тем и документов, темы, расположенные «слишком близко друг к другу», или несоответствие между количеством тем в документ и параметр Дирихле *α* .

Интерактивное тематическое моделирование — в сочетании с визуализацией — может помочь исправить проблемы тематических моделей. Сначала пользователь получает обзор коллекции с помощью визуализации тем и документов, а затем может увидеть и исправить случаи, в которых модель допускает ошибки.

Например, на рис. 3.6 показана тема, изученная из тезисов грантов, финансируемых Американскими национальными институтами здравоохранения (более подробно обсуждается в главе 5.1). Большинство тем были «хорошими»: они обобщали данные и рассказывали историю о связном фрагменте исследования, поддержанного NIH. Однако эта тема более проблематична; в нем сочетаются слова о центральной нервной системе со словами о мочевыделительной системе. Такая тема [Mimno et al., 2011] не дает четкого понимания, какие документы она должна представлять.

Ху и др. [2014a] решают эту проблему, позволяя пользователю добавлять в модель вероятностные ограничения [Boyd-Graber et al., 2007, Andrze-

*3.5. Резюме*

Тематические слова (до)

мочевой пузырь , наука, спинной мозг ,

спинномозговая\_травма ,

спинномозговая , мочевая , ури-

nary\_tract , уроте-

лиал , травма , двигатель ,

восстановление , рефлекс , цер-

викал , уротелий , фунц-

tional\_recovery

47

Тематические слова (после)

, спинной мозг ,

спинномозговая\_травма ,

позвоночник , травма , восстановление

эри , двигатель , рефлекс ,

уротелиальный , травмированный ,

функциональное\_восстановление ,

пластичность , двигательная активность ,

шейный отдел , движение

**Рисунок 3.6:** Примеры тем до и после интерактивного тематического моделирования из Hu et al. [2014а]. Изначально тема объединяет две темы ( мочевыделительная и центральная нервная система ), что нежелательно. Добавление ограничения на то, что слова «мочевой пузырь» и «спинной мозг» не должны встречаться в теме вместе, делает тему более последовательной и открывает понятия, которых раньше не было .

еврейский и др., 2009]. Например, пользователь может сказать, что «мочевой пузырь» и «спинной мозг» не относятся к одной и той же теме вместе. На рис. 3.6 показано, как тема становится более сфокусированной после того, как пользователь предоставил эту обратную связь. В отличие от вероятностных ограничений Choo et al. [2013] и Лунд и соавт. [2017] используют ограничения матричной факторизации, чтобы управлять изменениями в темах, что может быть намного быстрее.

**3.5**  **Резюме**

В то время как тематические модели предоставляют пользователям обзоры корпусов, тематические модели не могут сильно помочь, если пользователи не могут эффективно видеть или понимать основные темы и то, как они связаны с конкретными документами. Оценки помогают определить, каким частям модели можно доверять, а какие следует использовать с осторожностью. Интерактивные визуализации позволяют пользователям находить и уточнять информацию. В следующих главах мы поговорим о конкретных применениях этих идей, но эти идеи часто строятся на первоначальном понимании модели, предлагаемой визуализацией, обсуждаемой в этой главе.

**4**

**Исторические документы**

Тематические модели играют важную роль в анализе исторических документов. Исторические записи, как правило, обширны, и ими трудно управлять без интенсивной и трудоемкой организации. Записи сложны: они сопротивляются категоризации и могут даже не иметь стандартного правописания и форматирования. Но в истории есть нечто большее, чем управление документами. Задача историка не только впитывать содержание исторических записей, но и обобщать; найти закономерности и закономерности, которые соответствуют документам, но также и вне каких-либо отдельных доказательств. Тематические модели полезны, потому что они решают эти проблемы. Они масштабируемы, устойчивы к изменчивости и способны обобщать, оставаясь при этом основанными на наблюдениях.

Автоматизированные методы являются особенно ценным противовесом традиционным научным методам. Изучение истории — это столкновение с неожиданным, часто в контекстах, которые кажутся знакомыми. Мы не обязательно знаем, как люди в прошлом говорили о конкретных проблемах или как они организовывали свою жизнь. Возможно, более опасно то, что мы предполагаем, что знаем эти вещи и что наши предки видели мир так же, как и мы. Тематические модели дают нам интерпретируемую, но в то же время чуждую перспективу, основанную на документах, а не на наших собственных представлениях.

48

|  |  |
| --- | --- |
| *4.1. Газеты* | 49 |

собственные представления о том, как должно быть.

Время является критической переменной в изучении исторических документов. Хотя многие современные коллекции имеют значительный аспект изменения времени (см., например, наукометрию), время является определяющим элементом исторических исследований. Коллекции исторических документов обязательно относятся к другому времени, отличному от нашего, но также имеют тенденцию охватывать длительные периоды — десятилетия или даже столетия. В результате во многих примерах, приведенных в этой главе, документы организованы по временной оси. Соответствующий анализ особенно касается того, как язык, что отражается в концентрации тем и содержании тем, меняется с течением времени.

Эта глава организована вокруг различных форматов исторических документов. Повторяющийся фокус - желание построить события и дискурсы против времени. Мы начнем с исторических газет, которые относительно близки к современным новостным статьям, которые являются более привычным вариантом использования в тематическом моделировании. Затем мы рассмотрим другие формы исторических записей, такие как анналы и дневники. Они демонстрируют гибкость тематического моделирования, включая корпус не на английском языке и корпус на английском языке с неправильным написанием. Наконец, мы рассматриваем исследования исторической научной литературы.

**4.1**  **Газеты**

Newman and Block [2006] представляют пример тематического моделирования в исторических газетах 1 в сборнике статей из *Pennsylvania Gazette* с 1728 по 1800 год . 2 Эти статьи содержат 25 миллионов слов.токены в статьях и рекламных объявлениях и охватывают повседневную жизнь нескольких поколений до, во время и после основания Соединенных Штатов Америки. Авторы противопоставляют свое исследование созданным вручную индексам на основе ключевых слов, которые сосредоточены на конкретных терминах и могут непоследовательно применяться в больших корпусах. Ложные закономерности в использовании индексных терминов могут усложнить историческое исследование. Они приводят пример тега *adv* , который широко используется в первых и последних десятилетиях корпуса, но не в середине. Тематический подход привлекателен

1 Mei and Zhai [2005] представляют более ранний пример анализа *современных* новостей (т. е. когда данные уже оцифрованы). В их работе также используются тематические модели, чтобы показать эволюцию тем с течением времени.

* http://www.accessible-archives.com/

50 *исторических документов*

потому что он согласован во всей коллекции (при условии, что термины, используемые в документах, сами по себе согласуются) и потому что он абстрактен, что снижает вероятность того, что современные историки упустят ключевые термины.

Они сравнивают три метода нахождения семантических измерений, латентный семантический анализ [Deerwester et al., 1990], кластеризацию *k* -средних и топическую модель [Hofmann, 1999a]. Разницу между этими методами можно описать через экспрессивность. lsa хорошо встраивает типы слов и документы в низкоразмерное пространство, но отдельные измерения этого пространства не интерпретируются как темы. lsa слишком выразителен: он не накладывает ограничений, таких как положительность, на изученные измерения и, следовательно, дает неинтерпретируемые результаты, которые, тем не менее, хорошо соответствуют набору документов. Кластеризация k-средних больше похожа на модель тем и более успешна в поиске узнаваемых тем. Но он также склонен к повторению похожих кластеров с небольшими вариациями. Из-за предположения об одном членстве (документ может принадлежать только к одному кластеру) модель кластеризации не может представлять документы с различными комбинациями несколько независимых тем. Таким образом, модель k-средних недостаточно выразительна: она вынуждает «растрачивать» кластеры на частые комбинации более простых тем. Тематическая модель, напротив, обладает как гибкостью моделирования, так и ограничениями для поддержки интерпретируемых результатов.

Авторы считают, что изученные темы хорошо отражают динамику корпуса, хотя и не всегда прямым образом. В период, непосредственно связанный с американской революцией ( *конституция правительства штата, закон об объединении власти* ), резко возросло количество дискуссий о политике. Имеются также данные об экономических факторах: тема, касающаясяк описаниям ткани ( *шелк, хлопок, белый, черный лен* ) растет в 1750-х годах, но затем снижается, когда американцы перешли к производству домашней ткани в ответ на британскую торговую политику. Другие темы указывают на более тонкие изменения в языке. Тема, которая менее интерпретируема ( *скажем, вещь может подумать, что она сделала* ), соответствует серии длинных «публичных писем», содержащих более академические «аргументы». Это согласуется с другими результатами [Viermetz et al., 2008], которые предполагают, что темы могут быть долгосрочными или преходящими, что непосредственно зафиксировано Viermetz et al. [2008].

|  |  |
| --- | --- |
| *4.1. Газеты* | 51 |

Нельсон изучает темы в газетах времен Гражданской войны, в том числе в официальной газете Конфедерации Richmond Daily Dispatch. 3 Подобно Блоку и Ньюману, цель Нельсона состоит в том, чтобы организовать коллекцию по темам и измерить изменение распространенности этих тем во времени. Веб-интерфейс выделяет временное представление коллекции в виде ряда тематических временных рядов. Режим анализа не является ни полностью автоматизированным, ни ручным, а скорее сочетает в себе два подхода. Нельсон вручную помечает темы и группирует их в более крупные категории, такие как «рабство», «национализм и патриотизм», «солдаты» и «экономика».

Он проверяет модель, сравнивая темы с известной и ранее аннотированной категорией «объявления о беглых рабах». Эти документы представляли собой дофотографические описания беглых рабов и имели специфический язык, состоящий из аспектов внешности и возможных мест, где могли прятаться порабощенные люди. Он находит почти идеальное соответствие между преобладанием с течением времени объявлений о беглых рабах, помеченных вручную, и документами, которые имеют высокую концентрацию конкретной темы, что делает высокую вероятность таких условий, как *вознаграждение, годы* и *цвет* (ручные метки не использовались в обучение модели).Нельсон отмечает, что лишь немногие из этих документов полностью относятся к этой теме: в качестве критерия он использует отсечку в 21,5%.

Крупномасштабные группы тем Нельсона выделяют нити дискурса, которые могут или не могут быть коррелированы во времени. Модель выделяет три темы, которые имеют одинаковое временное распределение, достигнув пика в начале войны в 1861 году и в значительной степени исчезнув впоследствии. Это связанные, но разные темы: антисеверные настроения, выраженные в поэтической форме, антисеверные настроения, выраженные в язвительной прозе, и обсуждение отделения. Все три аспекта составляют один и тот же процесс, риторический призыв к войне. Другие связанные темы имеют несколько иное временное распределение. Нельсон группирует шесть тем, связанных с солдатами, и отображает их в порядке их максимальной концентрации с течением времени. Они переходят от «призыва в армию» и «приказов докладывать» к более поздним темам, связанным с «дезертирами», «пострадавшими» и «военнопленными». Опять же, это связанные темы, но вместо того, чтобы включать в себя одно событие, они прослеживают развитие все более тяжелой военной ситуации в

* Добыча Dispatch, http://dsl.richmond.edu/dispatch/

52 *исторических документа*

Конфедерация.

Ян и др. [2011] моделирует коллекцию исторических газет из Техаса, охватывающую период с конца Гражданской войны до наших дней. Цель является как исследовательской, чтобы узнать об интересах техасцев на протяжении 19 и 20 веков, так и *полуисследовательской* , чтобы узнать больше об истории и контексте конкретных, заранее определенных тем, таких как производство хлопка. В условиях тематической модели полуисследовательский анализ начинается с определения одной или нескольких тем, которые кажутся соответствующими интересующей теме, а затем использования этих тем в качестве оси исследования корпуса. Например, историк рассмотрел документы с темами, связанными с хлопком, и темами, которые одновременно встречаются в этих документах. Исследование также привело к более полным исследовательским результатам. Тема битвы при Сан-Хасинто , заключительного конфликта Техасской революции, приведшего к отделению от Мексики, появилась раньше, чем ожидалось. Дальнейшее расследование показало, что значение ключевой битвы при Сан-Хасинто было установлено намного раньше, чем ранее предполагали историки.

Газетное исследование Техаса поднимает несколько интересных методологических вопросов, касающихся предварительной обработки и итеративного моделирования. Авторы серьезно поработали над качеством оцифровки. Многие факторы, влияющие на качество оцифрованных исторических газет, от качества исходной печати до сканирования, сегментации статей и оптического распознавания символов (OCR). Для этого исследования была проведена обширная работа по автоматизированному исправлению орфографии. Еще одним примечательным фактором в этом исследовании является широкое использование нескольких тематических моделей. Часто предполагается, что единый корпус должен привести к единой модели, но на практике моделирование часто является итеративным и тесно связано с разработкой систем предварительной обработки. Аль-Сумаит и др. [2008] итеративно уточняют свою модель с каждым сегментом коллекции новостей, в то время как Yang et al. [2011] обучают разные модели на разных временных срезах корпуса. Хотя есть некоторое преимущество в сохранении согласованного тематического пространства с течением времени, разделение корпуса на отдельные разделы имеет определенные преимущества. В этом случае историков интересовали конкретные исторические периоды, такие как полный тираж газеты в течение нескольких ключевых лет, которые меньше, чем полный корпус, но все же слишком велики, чтобы их было легко читать. Авторы также описывают итеративный рабочий процесс, который включает

|  |  |
| --- | --- |
| *4.2. Исторические записи* | 53 |

сравнение выходных данных тематической модели после каждого из нескольких шагов предварительной обработки. Тематические модели часто эффективны при выявлении последовательных ошибок подготовки данных, таких как переносы в конце строки и постоянные ошибки распознавания.

**4.2**  **Исторические записи**

Другие типы записей, кроме газет, представляют интерес и представляют свои проблемы. В этом разделе рассматриваются два тематических исследования, в которых простота модели документа с набором слов является преимуществом, поскольку она допускает существенную вариативность правописания и языка как на английском, так и на других языках.

Эрлин [2017] ищет работу, связанную с эпистемологией, в большом корпусе книг на английском и немецком языках. Они «засевают» модели для каждого языка несколькими ключевыми словами, которые, по мнению авторов, связаны с этим предметом. Этот подход ближе к стандартному поиску информации, чем многие другие приложения тематических моделей, поскольку модель используется и как способ организации корпуса, и как способ сосредоточить внимание на конкретных аспектах. Использование ими тематической модели отличается от стандартного ИО тем, что они более преднамеренно открыты для связанных терминов и понятий: ожидается, что эпистемология будет широкой и скорее будет представлена комбинацией слов, чем каким-либо одним запросом.

Миллер [2013] использует китайские записи для исследования значения слова *zei* , или «бандит», в Китае времен династии Цин (1644–1912). Слово само по себе может подразумевать несколько различных форм антиобщественного поведения, которые трудно отличить только по частоте слов. Тематическая модель использует контекстную информацию для разделения этих эффектов.

Применение тематических моделей на китайском языке подчеркивает важность токенизации. Обычно мы получаем документы в виде длинных строк, но нас интересует идентификация *токенов* , которые представляют собой короткие строки с определенным значением. Разбиение документа на отдельные токены — часто упускаемая из виду часть процесса анализа документа. В европейских языках мы можем добиться хороших результатов, отделив строки буквенных символов от последовательностей небуквенных символов, хотя есть много особых случаев [Boyd-Graber et al., 2014]. Токены могут содержать

54 *исторических документа*

небуквенные символы, такие как апострофы и дефисы, и могут состоять из нескольких слов ( *королева Виктория* , *черная дыра* ). Во многих восточноазиатских системах письма мы не можем полагаться на орфографические соглашения для идентификации токенов. Миллер утверждает, что в классическом китайском языке один символ можно рассматривать как токен без ущерба для моделирования, но для японского и современного китайского языка мы часто должны полагаться на инструменты предварительной обработки, которые сами по себе потенциально ненадежны.

Кэмерон Блевинс моделирует дневник Марты Баллард (1735–1812), революционной акушерки времен войны, которая делала записи более 27 лет. 4 Модель предоставляет полезный способ обнаружения связей между словами и повторяющимися дискурсами. Как и в случае с другими историческими корпусами, Блевинс фокусируется на связи между темами и временем. Конкретные события, такие как рождение, можно выделить, просматривая всплески в определенной теме в ежедневном временном ряду. Но очевидны и более крупные тенденции. В качестве калибровочного эксперимента Блевинс измеряет связь темы, которая, по-видимому, относится к холодной погоде ( *холод, ветер, холод, снег, воздух* ) с месяцами года. Как и ожидалось, концентрация по этой теме минимальна с мая по август, повышается с сентября по январь и падает с февраля по апрель.

Блевинс выделяет несколько других тем, концентрация которых со временем меняется. Две темы, связанные с работой по дому, с упором на уборку и приготовление пищи, по-видимому, связаны во времени и нарастают на протяжении десятилетий. Блевинс связывает это открытие с предположениями о том, что по мере того, как Баллард становилась старше, а ее дети уезжали, ей меньше помогали члены семьи. Более тонкая тема включает описания усталости и болезни. Эта тема также увеличивается со временем и, по-видимому, коррелирует с темами работы по дому, за исключением последнего года дневника, когда усталость и болезни достигают наибольшей концентрации, а работа по дому снижается.

Этот анализ иллюстрирует исследовательский характер тематического моделирования: сами по себе эти наблюдения не являются окончательными, но они наводят на размышления и указывают на области дальнейшего анализа. Ученый может взять дневниковые записи, получившие высокие оценки по отдельной теме, в качестве списка для чтения и определить, насколько хорошо конкретный автоматически обнаруженный дискурс

* http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/

|  |  |
| --- | --- |
| *4.3. Научная литература* | 55 |

сопоставляется с темами из личного опыта Балларда. Например, можно проверить, относятся ли ссылки Баллард к усталости и болезни к ней самой или к пациентам. Модель не рассказывает всей истории, но указывает на то, где истории могут лежать.

Блевинс утверждает, что характеристики формы дневника делают ее хорошо подходящей для тематического анализа: «Короткие, ориентированные на содержание записи, которые обычно затрагивают ограниченное количество тем, кажутся удивительно связными и точными темами». Кроме того, недостаточная лингвистическая сложность тематической модели является преимуществом. Дневник написан лаконично, со множеством сокращений и с неправильным написанием, характерным для 18 -го века: «Болезнь госпожи Пейдж началась в Евнге, и Ши была доставлена в 11 часов от сына, который весил 12 фунтов». Модели, обученные на современных текстовых корпусах, могут даже не распознать этот пример как английский, но алгоритм тематического моделирования по-прежнему способен находить семантически значимые группы слов.

**4.3**  **Научная литература**

Исторические записи науки являются ценным источником для интеллектуальной истории. Многие пользователи используют API jstor «Данные для исследований». 5 API-интерфейс dfr является важным примером, поскольку он обеспечивает доступ к статьям, которые были отсканированы jstor и могут быть защищены авторским правом. Доступ к базовым документам в их исходной форме в виде читаемых последовательностей слов может быть ограничен по юридическим или коммерческим причинам. dfr обеспечивает простое представление избранных статей, предоставляя только частоту униграмм слов. Хотя допущение набора слов, используемое в тематических моделях, является ограничительным, в данном случае оно может быть преимуществом, поскольку исходная последовательность слов в любом случае не используется для вывода.

Mimno [2012] изучает коллекцию классических журналов, оцифрованных jstor, чтобы обнаружить изменения в этой области за 20 -й век. Отличительной особенностью данного исследования является использование *полиязычной* тематической модели [Mimno et al., 2009]. Детали этой модели и ее отличие от других моделей более подробно описаны в главе 8.1; для этого обсуждения нам нужна некоторая тематическая модель, которая может обнаруживать темы, которые непротиворечивы.

* http://dfr.jstor.org/

56 *исторических документов*

через языки. Англоязычный журнал сравнивают с немецкоязычным, изучая общий набор тем, каждая из которых имеет словарный запас на обоих языках. Другими словами, у топика есть два «модуса», в одном из которых он выделяет слова, взятые из распределения английских терминов, а в другом — слова, взятые из распределения немецких терминов. Связь между английскими и немецкими словами создается с помощью статей Википедии. Статьи в Википедии существуют на многих разных языках, и статьи на одном языке часто ссылаются на сопоставимые статьи на другом языке. Сначала автор выбирает статьи из англоязычной Википедии, соответствующие ключевым терминам в англоязычных журналах, а затем собирает статьи из немецкой Википедии, которые указаны как сопоставимые с выбранными англоязычными статьями.

Путем совместного обучения тематической модели на объединенном корпусе оригинальных журнальных статей и сопоставимых статей из Википедии модель дает представление об относительной концентрации научных интересов в двух языковых сообществах. Журнальные статьи на немецком языке содержат больше работ по праву и ораторскому искусству, темы, которые присутствуют в англоязычных статьях, но менее распространены. Модель также показывает значительный рост интереса к поэзии в немецком журнале в период после Второй мировой войны. В английских журналах, начиная с 1980-х годов, наблюдается значительный рост культурных и экономических исследований наряду с критической теорией, чего нет в немецких журналах.

Riddell [2012] также обращается к немецкой научной литературе ХХ века. Он считает, что темы хорошо сочетаются с такими авторами, как Гёте, и такими темами, как фольклор. Очевидные всплески использования этих тем, по-видимому, совпадают с годовщинами авторов (Гёте, братья Гримм). Ридделл подчеркивает, что модели полезны для постановки вопросов, но не заменяют исследования. Он отмечает, что «становится важным, чтобы те, кто использует тематические модели, подтверждали описание, предоставленное тематической моделью, ссылкой на что-то другое, чем сама тематическая модель».

Голдстоун и Андервуд [2014] используют тематическую модель в качестве инструмента для структурирования исследования корпуса, охватывающего более века. Их интересуют как изменения на уровне темы, так и на уровне

|  |  |
| --- | --- |
| *4.4. Резюме* | 57 |

использование слов в темах. Для этих авторов привлекательность тематического моделирования заключается в том, что модели способны лучше представлять контекстное значение, чем простые списки ключевых слов. Они пишут, что «значения слов меняются и зависят от контекста. По этой причине рискованно создавать группы слов, которые, по нашему мнению, эквивалентны какому-то заранее определенному понятию».

Они анализируют материалы Ассоциации современного языка 6 , чтобы найти сдвиги в фокусе внимания в английской литературе. Модель, обученная 150 темам по 21 000 статей, определяет тему, связанную с описаниями насилия : *сила, насилие, страх, кровь, смерть, убийство, действие, вина* . Используя Темпоральный сюжет утверждают, что концентрация этой темы во второй половине 20 века больше, чем в первой половине. Они контекстуализируют этот вывод, сравнивая частоту этих слов в более общем корпусе из n-грамм Google; нет сопоставимого изменения. Этот подход удерживает фиксированную тему и ищет связанные слова. Затем они поворачиваются и фиксируют слово «власть» и ищут связанные темы. В этом случае тема насилия оказывается относительно стабильной в своей ассоциации с целевым словом. Наибольшее увеличение наблюдается в другой теме, характеризуемой *формой слова «собственная власть* », в контексте которой она появляется почти исключительно после 1980 года. Как и во многих других темах, содержание этой темы трудно оценить только по основным словам. Потребуется дальнейшее изучение отдельных документов (например, с помощью инструментов, обсуждаемых в главе 3).

**4.4**  **Резюме**

В этой главе основное внимание уделяется поиску тем в документе, отражающих временные тенденции. Когда мы рассматриваем газеты, исторические отчеты и исторические научные журналы, мы ищем не только тематические фокусы каждого периода времени, но и то, как эти темы меняются в концентрации под влиянием исторических событий. Моделирование больших коллекций документов позволяет нам выявить, как события отражаются в письме и как возникают идеи и эмоции в ответ на изменяющиеся события.

В следующей главе мы продолжим обсуждение научных журналов.

6 MLA — профессиональная организация литературоведов в США.

58 *исторических документов*

сосредоточиться на том, как появляются новые идеи. В отличие от газет и дневников, отражающих реальность мира, научные рукописи действительно могут изменить мир за счет внедрения инновационных технологий. В следующей главе спрашивается, можем ли мы обнаружить и описать эти инновации.

**5**

**Понимание научных публикаций**

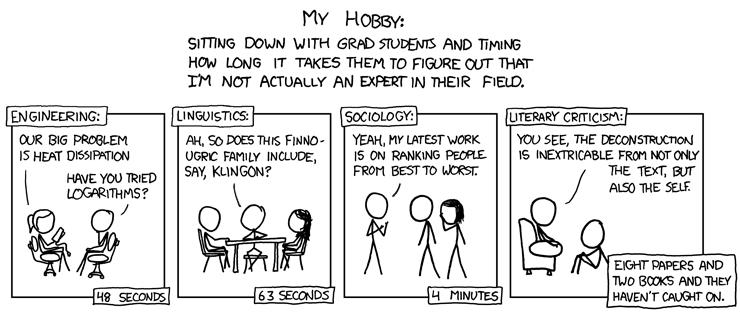
В главе 4 мы обсуждаем, как ученые используют тематические модели для понимания нехудожественных документов. Эта глава посвящена особому поджанру документальной литературы: научным документам. Научные документы заслуживают отдельной главы, потому что в этих документах используется очень специализированная лексика, они являются проводниками инноваций и формируют важные политические решения. Мы обсудим каждый из этих аспектов по очереди.

**Специализированные словари определяют области исследования** Во-первых, научные документы уникальны, потому что в отличие от общих документов их словарный запас точен и тщательно взвешен. «Сопротивление», «сращивание», «использование» и «требование» — общие слова, имеющие совершенно разные значения при использовании в специализированных технических контекстах. Их использование является маркером принадлежности к определенной дисциплине (рис. 5.1). Таким образом, способность тематических моделей фиксировать паттерны употребления слов также охватывает сообщество и принадлежность; это выходит далеко за рамки тематической организации тематических моделей, описанных в предыдущих главах.

**Научные документы создают инновации** Не каждая научная публикация является инновационной; на самом деле, большинство не таковы. Однако некоторые научные публикации

59

60 *Понимание научных публикаций*



**Рисунок 5.1:** Использование соответствующего языка является необходимым условием для участия в поле (но недостаточным). Тематические модели используют это для автоматического обнаружения областей исследования.

Сотрясающий землю. Такие разработки могут быть теоретическими, методологическими или эмпирическими. Теория относительности произвела революцию в физике. Генетика произвела революцию благодаря открытию методов полимеразной цепной реакции. Революцию в геологии произвело открытие свидетельств тектоники плит в виде магнитных следов на дне океана. В отличие от других областей, которые мы обсуждали, научные документы — это не просто *сообщения* о новостях или событиях; они *новости* .

Что делает анализ коллекций научных документов одновременно сложным и интересным, так это то, что инновации трудно обнаружить и трудно атрибутировать. Новаторские работы Эйнштейна 1905 года не были полностью признаны до тех пор, пока много лет спустя; важные идеи часто выдвигаются малоизвестным исследователем, но принимаются только после популяризации и поддержки другими исследованиями. Какой документ (или исследователь) в данном случае был истинным источником инновации? Как мы увидим в этой главе, тематические модели могут помочь ответить на этот вопрос.

**Наука и политика** Понимание научных публикаций важно для финансирующих организаций, законодателей и общественности. Государственное финансирование науки может создавать рабочие места, улучшать культуру и является важной формой международной «мягкой силы». Однако трудно понять, какие исследования финансировать, поскольку природа науки такова, что области постоянно меняются, что исключает жесткие классификации [Szostak, 2004]. Одна проблема

|  |  |
| --- | --- |
| *5.1. Понимание областей исследования* | 61 |

моделирование научных документов — это моделирование того, как меняются поля; статические модели, которые мы обсуждали до сих пор, не всегда подходят.

**5.1**  **Понимание областей исследования**

Одним из первых применений тематических моделей было понимание «областей науки». Griffths и Steyvers [2004] обнаружили, что они могут автоматически реконструировать официальные тематические коды Proceedings of the National Academy of Sciences (pnas) с использованием тематических моделей (рис. 5.2). Это полезная проверка здравомыслия: да, тематические модели коррелируют с тем, что мы часто называем научными дисциплинами. Они используют разные языки для методов, предметов изучения и имеют разных ключевых игроков.

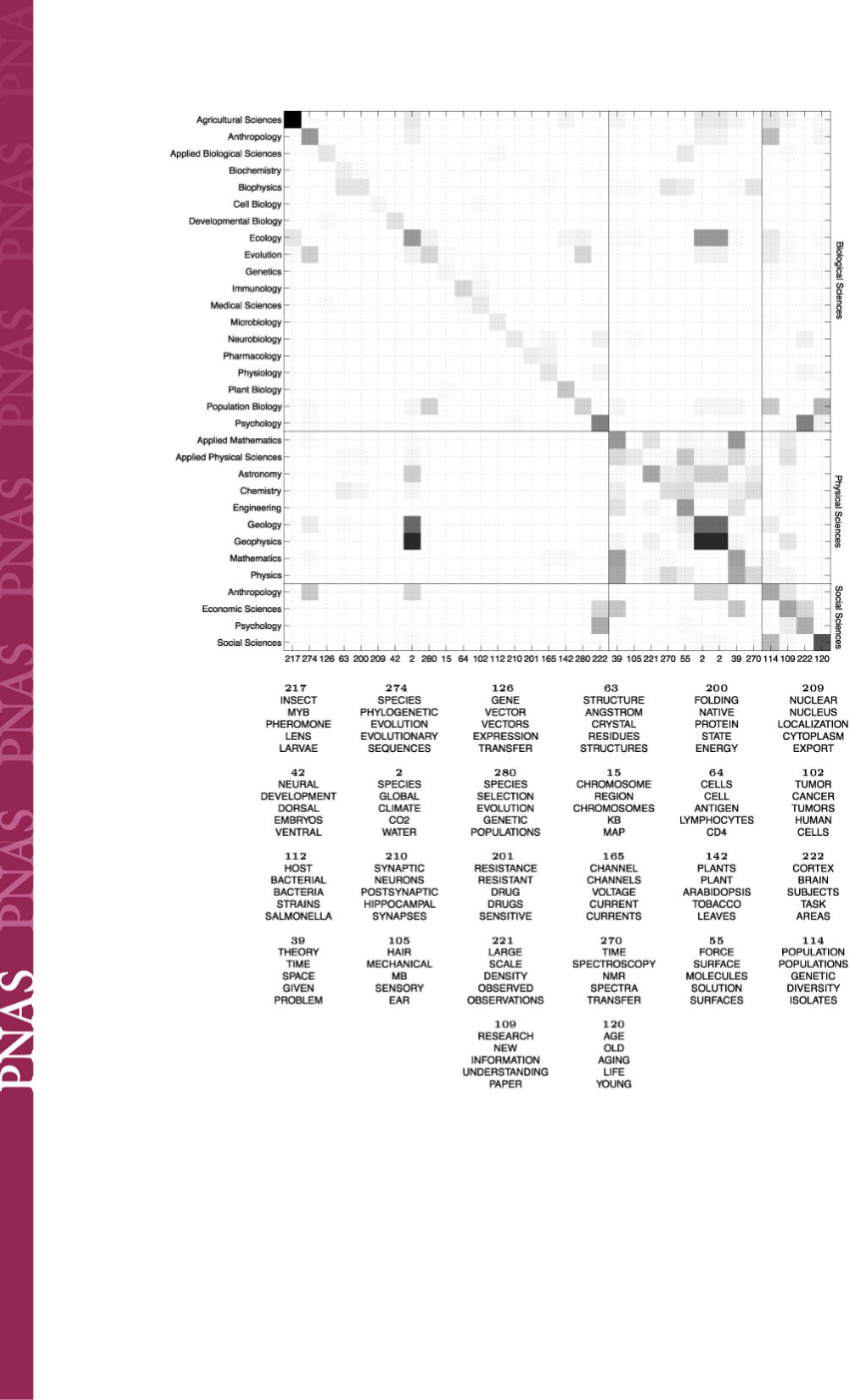
Эта работа стремилась найти различия между областями науки, но с ретроспективной точки зрения. Напротив, Talley et al. [2011] стремились составить карту приоритетов финансирования Американских национальных институтов здравоохранения (NIH) внутри организации.

Национальные институты здравоохранения являются ведущим американским агентством по финансированию биологических исследований и исследований в области здравоохранения. Них состоит из нескольких институтов, занимающихся конкретными заболеваниями, исследовательскими методами или системами организма; каждый из этих институтов управляет своим собственным независимым портфелем финансирования, что иногда затрудняет понимание «общей картины» финансирования.

Талли и др. [2011] используют тематические модели, чтобы помочь создать эту общую картину, в отличие от более трудоемких методов (например, ключевые слова из тщательно организованной онтологии). Их анализ обнаружил неожиданное совпадение приоритетов исследований в разных институтах. Например, многие институты изучают ангиогенез, образование новых кровеносных сосудов; в качестве лечения рака, при визуализации сердца, молекулярных основах ангиогенеза в глазах и того, как ангиогенез может сигнализировать об осложнениях при диабете.

На практике применить тематическое моделирование к коллекции рефератов грантов NIH было непросто. Создание моделей, приемлемых для пользователей, привыкших к ручному вводу ключевых слов, требовало тщательного изучения словарного запаса. Тематические модели пытаются найти темы, объясняющие все аспекты учебного корпуса. Сбор предложений по исследованиям будет

62 *Понимание научных публикаций*



**Рис. 4.** ( *Верхний* ) Средние значения *!* по каждой из диагностических тем для всех 33 второстепенных категорий PNAS, рассчитанных с использованием всех рефератов, опубликованных в 2001 г. Более высокие вероятности указаны более темными ячейками. ( *Нижний* ) Пять наиболее вероятных слов в самих темах перечислены в том же порядке, что и на горизонтальной оси в *Верхнем* .

**Рисунок 5.2:** После запуска тематической модели на pnas, Griffths and Steyvers [2004]

|  |  |
| --- | --- |
| это может быть менее очевидно при анализе, учитывающем только | темы регулярно растут и падают в популярности, 54 темы |
| найдены темы ( ось *x* ), которые могут воссоздать области исследований, определенные вручную. | |
| частоты отдельных слов. | показали статистически значимую линейную тенденцию к увеличению, и 50 |
| Чтобы найти темы, популярность которых постоянно росла или падала с | показал статистически значимый линейный тренд снижения, как в |
| по pnas (ось *y* ). |  |
| С 1991 по 2001 год мы провели линейный анализ тенденций *! Дж* по годам, | уровень *Р* 0,0001. Три самые горячие и самые холодные темы, |
| используя тот же единственный образец, что и в наших предыдущих анализах. Мы | оцениваются по размеру статистики теста линейного тренда, показаны |
| применил этот анализ к образцу, использованному для создания рис. 4. | на рис. 5. Самыми горячими темами, обнаруженными в ходе этого анализа, были |
| В соответствии с идеей о том, что наука демонстрирует сильные тенденции, с | темы 2, 134 и 179, соответствующие глобальному потеплению и |

Гриффитс и Стейверс PNAS **6 апреля 2004 г.,**  том. 101 доп. 1 **5233**

|  |  |
| --- | --- |
| *5.2. Как меняются поля* | 63 |

иметь значительный дискурс, связанный с темами, не связанными с исследованиями, представленными такими словами, как «предлагать», «поддерживать» и «финансировать». Для удаления было определено большое количество стоп-слов, характерных для корпуса, в основном связанных с темами, не относящимися к исследованиям. Кроме того, Талли и соавт. [2011] обнаружили, что предварительная обработка документов для объединения обширного списка многословных терминов в отдельные токены имеет существенное значение. Научная и другая техническая лексика часто образует некомпозиционные сложные термины из-за необходимости конкретики. Например, «аминокислота» содержит слово «кислота», но аминокислоты не имеют функционального сходства с «жирной кислотой» или «соляной кислотой». Объединение таких терминов в соединения с одним токеном привело к существенному повышению специфичности и понятности тем.

**5.2**  **Как меняются поля**

Одно из отличий науки от областей, обсуждавшихся в предыдущих главах, состоит в том, что ученые считают себя создателями единой связной структуры знаний. Каждая газета по-своему стоит на плечах гигантов. Таким образом, тематические модели для науки должны учитывать связи между документами во времени. Другое отличие науки состоит в том, что сами документы вносят новые идеи (об обнаружении этих инновационных идей мы поговорим в следующем разделе).

Один из первых методов *моделирования* смены тем рассматривал темы как слегка меняющиеся каждый год с помощью динамической тематической модели [Blei and Lafferty, 2006, dtm]. Каждая тема имеет четкое распределение по словам для каждого периода времени. Например, вероятность того, что тема физики выдаст слово «струна» в 1910 году, может быть низкой, но после 1970 года намного выше. Конечно, мы не хотим, чтобы темы каждый год были абсолютно разными — мы хотим, чтобы темы менялись, но не *слишком сильно* .

dtm рассматривает темы как изменяющиеся посредством *броуновского движения* : тема в год *t* берется из гауссовского распределения со средним значением в теме для года *t* - 1 (отдельный параметр дисперсии контролирует, насколько темы могут меняться каждый год). Здесь вы можете возразить, учитывая наше обсуждение распределений из главы 1.3.1: гауссианы производят непрерывные наблюдения, которые могут быть отрицательными или больше 1,0,

64 *Понимание научных публикаций*

в то время как темы представляют собой полиномиальные распределения по дискретным результатам.

Чтобы перейти от гауссовых розыгрышей от *~ x* ∈ R *d* к дискретным распределениям по *d* результатам, Блей и Лафферти [2006] используют логистическую нормальную форму для создания полиномиального распределения.

|  |  |  |
| --- | --- | --- |
| опыт *х к* |  |  |
| *р* ( *г* знак равно *к* | *~ х* ) = П *я* опыт *х я* | *,* | (5.1) |

вместо того, чтобы рисовать дискретное распределение из распределения Дирихле (см. уравнение 1.3). Это значительно усложняет вывод, но позволяет постепенно менять темы из года в год.

С помощью этой модели dtm обнаруживает, как поля меняются с течением времени. В начале двадцатого века язык физики сосредоточился на понимании того, как «эфир» распространяет волны и фундаментальные силы; к середине века понимание «квантовых» эффектов взяло верх; к концу века экспериментальная физика с большими ускорителями частиц возглавила поиск еще более экзотических членов субатомного зверинца. Хотя последняя тема почти неузнаваема по сравнению с первой, все они явно относятся к физике; допущения моделирования dtm фиксируют эти почти незаметные изменения каждый год.

Однако перелистывание страницы календаря не управляет наукой; изменения могут произойти в любое время. Ван и др. [2008] фиксирует изменения тем в непрерывном времени; каждый документ получает «собственное» представление темы, которое может незначительно отличаться от предыдущей версии темы. Это может помочь зафиксировать внезапные изменения в научных темах, например, в результате инновационного вклада.

Альтернативный взгляд на новаторство исходит от авторов. Стейверс и др. [2004], Розен-Цви и соавт. [2004] построили генеративную модель, включающую личность авторов. У автора есть набор тем, о которых он пишет, и каждый документ представляет собой комбинацию тем, которые волнуют его авторов. Чжоу и др. [2006] расширяют этот аргумент, моделируя использование темы на основе марковского перехода от других тем. Затем они используют эту модель, чтобы найти авторов, которые управляют этими переходами.

|  |  |
| --- | --- |
| *5.3. Инновации* | 65 |

**5.3**  **Инновации**

Изменения в полях происходят из-за *инноваций* . Ученые разрабатывают новые методы, новые термины и новое понимание мира. Эти понятия требуют новых слов, которые находят отражение в их научных публикациях. В отличие от других областей, где документы просто сообщают об изменяющемся мире, научные документы сами по себе являются той силой, которая может изменить мир: от « *Происхождения видов » Дарвина* до работ Эйнштейна по теории относительности.

Можем ли мы найти, где происходит это изменение? Таким образом, с точки зрения моделирования нам нужны модели, которые также могут изменяться. В отличие от подхода, описанного в главе 4.3, в котором основное внимание уделялось *статическим* темам, здесь мы сосредоточимся на *динамических* темах, которые могут меняться, и на том *, кто* несет ответственность за эти изменения.

С исторической точки зрения нам может быть интересно узнать, кто первым провел новаторские исследования. Измерение контекста инноваций также может быть полезным для политиков [Largent and Lane, 2012]. Вместо того, чтобы ждать «волшебных» или случайных открытий, мы могли бы захотеть измерить, как условия, команды и формы исследовательского сотрудничества приводят к прорывам. Тогда более совершенные прогностические модели могли бы помочь направить новые инициативы по воссозданию условий, ведущих к важным выводам.

С точки зрения темы оценка воздействия сводится к определению того *, кто* несет ответственность за изменение темы. Манн и др. [2006] найти оченьцитируемые статьи в контексте отдельных тем. Этот подход помогает контекстуализировать влияние по отношению к подобластям: очень влиятельная статья по математике может иметь такое же количество цитирований, что и умеренно успешная статья по молекулярной биологии, просто потому, что последняя область намного больше. Они также ищут статьи, имеющие тематически разнообразное влияние, путем измерения тематической энтропии статей, цитирующих данную статью. Эти широко значимые документы, как правило, являются методами и инструментами. С институциональной точки зрения Ramage et al. [2010b] рассматривают ретроспективную перспективу: после подгонки стандартной тематической модели *lda* найдите распределение по темам исследований во всем исследовательском сообществе в момент времени *t* и найдите места в прошлом, исследования которых больше всего напоминают настоящее. Они предполагают, что эти институты «приводят» другие институты к принятию их идей.

66 *Понимание научных публикаций*

Цитаты — полезное, но неполное руководство по решению этой проблемы. Дитц и др. [2007] разработали модель «копирования», которая использует сеть цитирования для моделирования использования языка в цитируемых документах: если вы цитируете статью, вы, вероятно, повторно используете часть языка оригинальной статьи. Он и др. [2009] распространили эту идею на более сложные графы. Однако это предполагает идеальный график цитирования, который не всегда доступен для некоторых полей.

Учет более тонких эффектов на индивидуальном или лабораторном уровне требует уточнения модели. Gerrish and Blei [2010] адаптировали модель случайного блуждания Wang et al. [2008] (Глава 5.2) для научных изменений. Герриш и Блей [Gerrish and Blei, 2010] предлагают вместо того, чтобы тема беспорядочно блуждала по новым понятиям, чтобы инновационные статьи «подталкивали» темы, чтобы они выглядели так, чтобы использовать словоупотребление этих инновационных документов. Эта модель называется «моделью динамического воздействия» (dim). Предполагается, что понятия и идеи представлены языком. Если мы можем определить изменения в использовании слов, это изменение означает, что в основе лежали изменения в понятиях.

Например, Penn Treebank [Marcus et al., 1993] произвел революцию в естественном языковом процессе и помог совершить статистическую революцию в вычислительной лингвистике. Одним из многих его последствий является то, что люди стали использовать слово «дерево» гораздо чаще, чем в прошлом. dim фиксирует это, явно моделируя влияние *l k,d* документа *d* в теме *k* .

Документы, которые не производят фурора, не имеют заметного влияния, в то время как влиятельные документы поглощаются другими учеными, которые перенимают влиятельные идеи и, что особенно важно, их язык. В большинстве документов темы вообще не перемещаются, поэтому разумно предположить, что *l d,k* для большинства документов равно нулю. Однако влиятельные документы изменят тему.

Тема изменяется влиятельным документом, делая распределение темы по словам более похожим на слова во влиятельном документе. Например, в статье, посвященной Penn Treebank, «treebank» используется гораздо чаще, чем «картофель», поэтому тема будет иметь более высокую вероятность для «treebank» после включения влияния документа.

Это включение происходит аналогично дрейфу динамической тематической модели (dtm, Глава 5.2). Вместо случайного дрейфа *направление*

|  |  |
| --- | --- |
| *5.4. Резюме* | 67 |

изменения темы основаны на словах, используемых во влиятельных документах, а *величина* отклонения зависит от того, насколько влиятельным является документ.

Каждое из этих условий является случайной величиной; вывод в модели обнаруживает настройки случайных величин, которые лучше всего объясняют данные. Оценки влияния, сделанные димом, хорошо коррелируют с количеством цитирований статьи (традиционная мера влияния). Однако, в отличие от цитат, затемнение можно использовать в более неформальной обстановке для обнаружения влиятельных документов: например, когда блог или письмо представляет влиятельные идеи.

**5.4**  **Резюме**

Понимание научной коммуникации позволяет нам увидеть, как с годами продвинулось наше понимание природы, технологий и техники. Якобы тематические модели могут фиксировать, как эти области изменились и получили дополнительные знания с каждым новым открытием. По мере того как научная деятельность становится все более распределенной и динамичной, эти инструменты становятся важными для ученых, стремящихся понять тенденции и развитие, а также для политиков, стремящихся направлять инновации.

В отличие от научных тенденций, в следующей главе рассматривается менее буквальное использование слов. В отличие от науки, художественная литература и художественная литература используют слова и фразы для выражения эмоций и настроения. Однако, как и в случае с наукой, исследователи могут использовать тематические модели для выявления закономерностей использования слов, отражающих художественные и литературные тенденции.

**6**

**Художественная литература и литература**

В этой главе рассматриваются документы, которые ценятся не только за их информационное содержание, но и за их художественное выражение. Есть много способов читать художественную литературу, поэзию и риторику. То, как мы читаем, влияет на выводы, которые мы можем сделать. Ученые традиционно сосредотачивались на подходе «внимательного чтения», при котором цель состоит в том, чтобы определить конкретные черты отрывка, которые передают более общий смысл, эмоции или атмосферу. Эти функции могут включать нюансы выбора слов, эхо звука через рифму или аллитерацию или просодические особенности, такие как ритм или каденция.

**6.1**  **Тематические модели в гуманитарных науках**

В то время как внимательное чтение является основополагающим инструментом в изучении литературы, оно обязательно ограничено его масштабом. Мы ценим литературу, потому что это один из лучших способов запечатлеть дух эпохи и опыт тех, кто ее пережил. Но стандартные методы внимательного чтения требуют узкого внимания и тщательной интерпретации. Тематические модели дополняют внимательное чтение двумя способами: в качестве метода опроса и в качестве средства отслеживания и сравнения крупномасштабных закономерностей.

68

|  |  |
| --- | --- |
| *6.1. Тематические модели в гуманитарных науках* | 69 |

Метод опроса относительно прост, он связывает отрывки, о которых читатель, возможно, не знал. Внимательное чтение — лучший способ проанализировать короткий отрывок текста, но какие короткие отрывки текста мы хотим проанализировать? Поскольку никто не может прочитать — а уж тем более внимательно прочитать — весь доступный материал о культуре или периоде времени, ученые часто вынуждены пытаться делать масштабные аргументы об истории литературы на основе небольших свидетельств. И это маленькое исследование выбрано не случайно: один и тот же маленький канон изучается подробно, в то время как огромная часть остается «большим непрочитанным» [Moretti, 2000]: произведения, которые *никогда не* изучаются. Выявление широких тем, а затем сопоставление этих тем с их реализацией в различных контекстах может выявить произведения или разделы произведений, которые «прячутся у всех на виду», неизвестные современной науке из-за безвестности.

Альтернативный и менее традиционный метод анализа часто называют «дистанционным чтением» [Moretti, 2013a]. Этот подход использует компьютерные методологии. Тематическое моделирование стало центральным инструментом дистанционного чтения, как способом организовать наше чтение крупномасштабных паттернов [Blei, 2012].

Тематический анализ, рассматриваемый как способ выявления повторений языка или дискурса в нескольких произведениях, перекликается со многими более привычными подходами к изучению литературы. В самом широком смысле определение жанра или литературного периода означает разделение корпуса на разделы на основе некоторого наблюдаемого критерия. Мы утверждаем, что это «готическая» литература, характеризующаяся атмосферными описаниями замков, или литература «киберпанк», характеризующаяся противоречивыми отношениями с информационными технологиями. В меньшем масштабе темы или тропы вновь появляются в разных контекстах. На самом детальном уровне ученые определяют повторяющиеся фразы, такие как описательные эпитеты, используемые в устной поэзии Гомера.

Статистический тематический анализ преследует ту же цель, но достигает ее другими средствами. Вместо жестких границ, определяемых датой публикации или национальностью, алгоритмы идентифицируют жанр по повторяющимся словам, образующим следы этих тем. Темы сами по себе не представляют темы, а скорее идентифицируют имплицитные статистические закономерности в использовании слов, вызванные наличием жанров, тем и дискурсов.

70 *художественная литература*

Однако применение тематических моделей к художественной литературе ставит новые задачи. Jockers [2013] обучает модель из 500 тем на корпусе из 4000 англоязычных романов. Из этого корпуса вытекает несколько проблем. Они присутствуют в других контекстах, но они гораздо более очевидны в художественной литературе.

**6.2**  **Что такое документ?**

В большей части литературы о тематических моделях термин «документ» используется исходя из неявного предположения, что у пользователей есть то, что называется документами. В канонической статье журнала lda [Blei et al., 2003] это слово используется 143 раза, но никогда не определяется. Значение «документа» часто достаточно ясно: новостная статья или научный реферат. Что не было ясно в этой более ранней работе, так это то, что это определение может быть проблематичным, особенно для документов, длиннее нескольких страниц текста.

Например, рассматривать романы как единый набор слов не получится. Темы, возникающие в результате обработки этого корпуса, чрезмерно расплывчаты и лишены тематической согласованности. Мы не должны удивляться этому открытию. Предположение о тематической модели состоит в том, что концентрация тем в документе фиксирована и неизменна от начала документа до конца. Естественное письмо редко соответствует предположению о тематической модели, и роман, не имеющий тематических вариаций на всем протяжении, вряд ли будет опубликован.

Нам нужно найти хорошую сегментацию на более короткие контексты (в отличие от социальных сетей, которые часто необходимо объединять в более длинные документы, см. главу 7.5.1). Мы предполагаем, что темы выражены в разных разделах длинного документа, такого как роман. Если сегментация хорошо справляется с определением границ между этими разделами, каждый результирующий сегмент должен иметь относительно небольшое количество тем. Если сегментация не справляется с определением границ, мы должны видеть сегменты, которые в среднем содержат больше тем, потому что наши сегменты объединяют фрагменты нескольких тематических сегментов.

Джокерс [2013] предпочитает не полагаться на структурные маркеры, такие как разделение на главы, и делит романы на куски по 1000 слов. Такая обработка приводит к связным, четко сфокусированным темам, которые могут быть обоснованно

|  |  |
| --- | --- |
| *6.3. Люди и места* | 71 |

используются в качестве прокси для узнаваемых тем.

Хотя сегментация фиксированной длины эффективна, она не обязательно идеальна. Элджи-Хьюитт и др. [2015] сравнивают различную сегментацию фиксированной длины с сегментацией на основе абзацев. Они оценивают разницу между обработками, измеряя концентрацию тем в каждом сегменте текста после моделирования. Индекс Герфиндаля — это мера концентрации в дискретных распределениях вероятностей, рассчитываемая как сумма квадратов вероятностей каждого возможного значения:

|  |  |
| --- | --- |
| Икс | (6.1) |
| Герфиндаль( *п* ) знак равно *п* ( *Икс* ) 2 *.* |

*Икс*

Например, рассмотрим два распределения *P* и *Q* по набору символов { *a, b, c, d, e* }. Если *P* имеет ненулевую вероятность только для одного символа *P* ( *a* ) = 1 *.* 0 и нулевая вероятность для всех остальных символов, Герфиндаляиндекс *P* будет равен 1,0. Если *Q* имеет одинаковую вероятность для всех пяти символов *Q* ( *a* ) = *Q* ( *b* ) = *...* = *Q* ( *e* ) = 0 *.* 2 индекс Герфиндаля будет равен 5 · 1 5 · 1 5 = 0 *.* 2.

Когда корпус романов XIX века делится на абзацы, индекс Герфиндаля по концентрации тем внутри каждого сегмента постоянно больше, чем тот же индекс, рассчитанный при разделении того же корпуса на фрагменты по 200 слов одинакового размера, что означает, что распределение -ция над темами для каждого сегмента в большей степени ориентирована на меньшее количество тем. Установка размера фрагмента на среднюю длину абзацев в корпусе, восемьдесят два слова, увеличивает показатель концентрации Герфиндаля, но результирующее значение все еще меньше, чем значение, основанное на абзацах. Этот результат обнадеживает, так как предполагает, что абзацы действительно имеют какое-то последовательное значение, по крайней мере, в этом сборнике романов XIX века.

**6.3**  **Люди и места**

Поскольку действие большинства художественных произведений происходит в воображаемых мирах, которые не существуют вне самого произведения, в них есть такие слова, как имена персонажей, которые очень часто встречаются в данном месте, но никогда не встречаются где-либо еще. Этот шаблон совпадения слов проблематичен для тематических моделей, потому что их можно рассматривать как машины для поиска групп слов, которые

72 *Художественная литература и художественная литература*

часто встречаются вместе, а не в других контекстах. Имена персонажей — по этому критерию — идеальная тема. Моделирование этих документов может привести к темам, которые по существу представляют собой списки имен персонажей.

В качестве примера рассмотрим модель с пятьюдесятью темами из четырнадцати романов Чарльза Диккенса и ее главными словами из набора тем (таблица 6.1). Заглавные буквы не преобразуются в строчные, чтобы подчеркнуть наличие имен собственных. В нескольких темах преобладают имена, написанные с заглавной буквы, с четко идентифицируемыми отдельными романами: Тема 4 — *Оливер Твист* , Тема 5 — *Николас Никльби* , Тема 6 — *«Записки Пиквикского клуба»* иТема 7 — *«Повесть о двух городах»* . На самом деле, ровно половина отдельных слов в топ-20 слов по всем темам пишется с заглавной буквы, и почти все они являются именами собственными.

Сосредоточение внимания на персонажах не всегда неинформативно, а в некоторых случаях может подчеркнуть структуру произведений. Все темы 1–3 относятся в первую очередь к « *Холодному дому»* (за исключением *Скруджа* ), но сосредоточены на разных взаимосвязанных сюжетных линиях. Первый посвящен леди Дедлок, второй — мистеру Джарндису и двум его подопечным, Ричарду и Аде, а третий — расследованиям детектива мистера Бакета. Сюжет сосредоточен вокруг раскрытия связей между этими явно не связанными между собой группами.

Джокерс [2013] подходит к этой проблеме, создавая список стоп-слов, который удаляет все имена персонажей перед моделированием. Существует много способов построения таких списков. Списки общих имен — хорошее начало, но они могут не соответствовать конкретному корпусу. Некоторые языки помечают имена собственные орфографическими соглашениями, такими как использование заглавных букв, но это, как правило, зашумлено. Полезной эвристикой в английском языке является определение терминов, которые пишутся с заглавной буквы более чем в 90% случаев. Даже в этом случае имена, которые также являются общеупотребительными словами, такими как *daisy* и вышеупомянутый Mr. Bucket, или слова, написанные с заглавной буквы по другим причинам, такие как *god* , могут привести к непредвиденным результатам. Более того, в некоторых языках регистр букв не различается (иврит, корейский), а в других он используется для других целей (все существительные в немецком). Инструменты распознавания именованных сущностей сканируют текст на наличие языковых шаблонов, указывающих на личные имена, и могут обеспечить большую точность, чем более простые методы. Тем не менее, не существует известного способа избежать внимательного рассмотрения значения слов.

|  |  |
| --- | --- |
| *6.3. Люди и места* | 73 |

**Таблица 6.1:** Примеры тем из романов Чарльза Диккенса без удаления имен персонажей (упорядочены вручную).

Условия темы

* Леди Лестер Скрудж Дедлок Раунсуэлл ледишип Уолд Чесни Призрак Волумния Рождество Семья Талкингхорнов Дух Баронет племянник Роза Скрудж экономка Леди
* Ричард Джарндайс опекун Ада Чарли Кэдди дорогая Ским-пол мисс Саммерсон Эстер Джеллиби мисс Фоулз Кендж Вудкорт вполне себе Гуппи Канцелярия
* говорит Джордж Бакет Снэгсби Гуппи возвращается Мелкая травка Сумка-нет приходит Талкингхорн смотрит берет солдат делает дружит

идет спрашивает плачет Чадбанд

* Ответил Оливер Бамбл Сайкс Еврей Феджин мальчик девочка Роуз Браун-лоу дорогой джентльмен Монахи Ной доктор Джайлс Доджер леди Нэнси Билл
* Николас Никльби Ральф Кейт Ньюман ответил Тим Малберри Манталини Брат Криви Н оггс Мадам Грайд Линкинвотер
* Пиквик Уинкль ответил джентльмену Тапмену Уордлу Сноду:

трава Пиквик Перкер толстяк Барделл дорогой Джингл осведомился Фогг Додсон друзья друг леди

* Грузовик Дефарж Доктор Манетт Просс Картон Дарней Мадам

Люси Монсеньор Кранчер Джерри Страйвер заключенный Маркиз Чарльз месье Тельсон отец Пэрис

* карета дядя джентльмен леди ящик кучер джентльмены домовладелец получить Лондон охрана внутри лошади официант мальчики почта пассажиры большой лучшая шляпа
* уличная дверь улицы окна дома окна комнаты несколько железных стен стены комнаты темные внутри двери магазина угол маленький стоял

большой

1. деньги письмо бумага бизнес читать фунты бумаги пятьсот офис тысяч клерк оплачивается лет ручка следующий юридический стол письма неделя

74 *Художественная литература*

в контексте.

Романы описывают людей и места, но они также создаются людьми (авторами), на которых влияет их культурная среда. Джокерс и Мимно [2013] проводят ретроспективный анализ более ранней модели Джокерса с 500 темами, чтобы определить, существует ли связь между использованием конкретных тем и переменными метаданных, такими как пол автора, национальность автора и год публикации. Они обнаружили, что концентрация многих тем сильно коррелирует с полом автора, и что эти корреляции статистически значимы. Такая проверка значимости может быть проведена с помощью тестов рандомизации и бутстреп-тестов. Оба метода создают «поддельные» корпуса, похожие на настоящий корпус, но отличающиеся определенным образом. Тесты рандомизации или перестановки случайным образом меняют присвоение меток (например, пол автора). Если наблюдаемая корреляция между темой и внешней переменной находится в пределах корреляций, полученных путем случайного присвоения документам меток, мало статистических доказательств того, что наблюдаемая корреляция имеет смысл. Тесты Bootstrap сохраняют взаимосвязь между документами и переменными метаданных, но передискретизируют документы с заменой. Этот тест показывает, зависит ли результат от наличия или отсутствия определенного документа. Если существуют большие различия между случайно сгенерированными корпусами, наблюдаемая корреляция может быть результатом необычных выбросов, а не постоянной закономерности.

В то время как использование методов проверки статистических гипотез ценно в контексте крупномасштабного дистанционного чтения, литературный анализ не похож и не должен быть похож на клиническое испытание, существуют различия между их использованием в научном контексте и их использованием в более типичные научные исследования. Во-первых, наличие необычных выбросов или единичных примеров на самом деле может быть положительным результатом. Предположение о том, что конкретная работа может радикально отличаться от предположительно подобных примеров, могло бы стать началом новой точки зрения. По крайней мере, он может выявить проблемы редактирования и курирования. Во-вторых, критической переменной в анализе статистической значимости является размер выборки. В отличие от запланированного эксперимента, этот размер выборки обычно не находится под нашим контролем: у нас есть литература, которая у нас есть. Наконец, жизненно важно избегать стремления трактовать показатель значимости как двоичный результат валидный/невалидный. Если числовое

|  |  |
| --- | --- |
| *6.4. Помимо буквального* | 75 |

баллы должны использоваться вообще, они должны быть представлены как «уровень поддержки» с учетом имеющихся документов. Гуманисты также могут фундаментально чувствовать себя более комфортно с сомнительными гипотезами: наблюдаемая связь с 10-процентной вероятностью быть чисто случайной все же может быть очень сильным результатом.

Например, Джокерс и Мимно [2013] оценивают интригующую гипотезу о том, что тема о религиозных фондах (монастырях и аббатствах) чаще используется неизвестными авторами 1 , чем (известными) авторами мужского или женского пола. Предполагается, что авторы предпочитали оставаться анонимными, чтобы писать о политически и религиозно щекотливых темах. Эта корреляция, однако, показала большую вариабельность при проверке начальной загрузки, и действительно, одно из предположительно анонимных произведений оказалось сокращением романа Энн Рэдклифф. Образец все еще присутствует без влияния этих работ, но нет четкой и неоспоримой связи между анонимностью и сомнением в религиозном авторитете.

Действие художественной литературы иногда происходит в контексте реальных мест. Тангерлини и Леонард [Tangherlini and Leonard, 2013] рассматривают вложенные модели подкорпусов в датской литературе таким образом, чтобы выявить связи между реальными событиями, культурными движениями и вымышленными отголосками. Их метод, который они описывают как «линию трала», использует указанный пользователем подкорпус в качестве запроса, а затем ищет в оставшейся части корпуса работы, соответствующие этому запросу. В качестве примеров они находят произведения, на которые повлиял перевод Чарльза Дарвина на датский язык, произведения, на которые повлиял «Прорыв в современность», и произведения, на которые повлиял фольклор и региональная литература.

**6.4**  **За пределами буквального**

Одной из отличительных черт художественной литературы и литературы является использование образного языка. Не очевидно, что неразумные машины, не понимающие культуру, способны обрабатывать такие метафоры. Тем не менее, Rhody [2012] демонстрирует на поэтическом корпусе, что, хотя

* Авторы, неизвестные современной науке, а не авторы, публикующиеся под известными псевдонимами.

76 *художественная литература*

темы не представляют собой символические значения, они являются хорошим способом обнаружения конкретного языка, связанного с повторяющимися метафорами.

В частности, Роди исследует корпус из 4500 стихов, описывающих произведения искусства (или *экфрастические* стихи). Она тренирует модель по шестидесяти темам и выделяет несколько особенно интересных тем. Одна из этих тем с высокой вероятностью указывает на *ночь, свет, луну, звезды, день, темноту, солнце, сон, небо, ветер, время, глаза, звезды, темноту, яркий свет* . Очевидное значениетема ясна и хорошо выражена одним главным словом: ночь. Но Роди обнаруживает, что когда она исследует *контекст* этой темы, во всех стихотворениях используется последовательная метафора, связывающая ночь и сон со смертью. Понятие смерти не фигурирует в первых словах — поэты не обращаются к этому вопросу напрямую. Тем не менее, модель определила пример небуквального, образного языка, хотя, поскольку он основан на реальных словах, он не может представить более глубокий смысл поэтов. Это потому, что поэты используют последовательный «поверхностный» язык для представления последовательной метафоры. Метафору невозможно обнаружить напрямую, но использование метафоры поэтом имеет характерную особенность, которую можно наблюдать.

Роди выделяет вторую тему, которая представляет собой пример другого типа небуквального значения. Эта тема придает высокую вероятность *смерти, жизни, сердцу, мертвому, длинному, миру, крови, земле, человеку, душе, людям, лицу, дню, боли, смерти* . В отличие от предыдущей темы, тема напрямую ссылаетсясмерть и жизнь, но ему также не хватает того, что Роди называет «недвусмысленной понятностью» *ночной* темы. Но изучение контекста стихов, содержащих эту тему, обнаруживает другую закономерность. Эти стихи имеют последовательную *форму* , которую Роди описывает как элегическую. Она пишет, что «в произведении Пола Лоуренса Данбара «Мы носим маску» ни разу не упоминается слово «смерть», дискурс, который Данбар использует для описания стирания идентичности и оков расовой несправедливости, определяется моделью как в значительной степени основанный на языке, связанном с со смертью, утратой и внутренней суматохой — язык, из которого, бесспорно, взята и «Звездная ночь».

|  |  |
| --- | --- |
| *6.5. Сравнение со стилометрическим анализом* | 77 |

**6.5**  **Сравнение со стилометрическим анализом**

В дополнение к обсуждению того, что исследователи сделали в литературном анализе с помощью тематических моделей, полезно рассмотреть, как другие технологии использовались в тех же условиях. Одним из наиболее устоявшихся применений вычислений в литературоведении является стилометрия, или, точнее, вопрос атрибуции авторства [Juola, 2006]. Показательно противопоставление целей и методов стилометрии целям и методам тематического моделирования.

Критическое понимание современной стилометрии заключается в том, что авторам легко сместить фокус своей работы, но гораздо труднее изменить полусознательный стиль своего языка [Mosteller and Wallace, 1964]. Подразумевается, что содержательные слова, такие как существительные и прилагательные, являются относительно плохим индикатором авторства или, по крайней мере, авторского стиля, в то время как функциональные слова, такие как определители, союзы и предлоги, несут больше информации об авторстве. Поэтому такие меры, как дельта Берроуза [Burrows, 2002], ограничивают внимание наиболее часто встречающимися словами в корпусе.

Отличие от тематического моделирования очевидно: стилометрический анализ фокусируется на частых, малоинформативных словах и игнорирует содержательные слова, в то время как тематическое моделирование обычно делает прямо противоположное. Обычно мы удаляем часто встречающиеся слова, используя стоп-лист, а в художественной литературе мы идем еще дальше, удаляя слова, которые слишком характерны для конкретного произведения. Таким образом, предположение о тематическом моделировании состоит в том, что цель состоит в том, чтобы найти тематические компоненты, которые *не являются* специфическими для одного автора, а скорее повторяются с большей или меньшей вариацией в нескольких работах. Там, где стилометрия стремится заглянуть за пределы того, что говорят авторы, и сосредоточиться на том, как они это говорят, использование тематического моделирования заключается в поиске случаев, когда разные авторы пишут об одном и том же.

**6.6**  **Внедрение «Темы»**

Использование тематического моделирования при изучении литературы оказалось полезным как для ученых-гуманитариев, так и для исследователей машинного обучения. Для ученых эти модели предлагают возможность более точного подхода к концепциям, которые традиционно были расплывчатыми и импрессионистскими, такими как

78 *Художественная литература*

тема, жанр и мотив. В то же время, что несколько парадоксально, литературные документы представляют собой настолько радикально отличающийся способ языка, чем новостные статьи или научные публикации, что они заставляют нас усомниться в кажущейся точности статистических подходов.

Тематические модели позволяют реализовать концепцию дистанционного чтения. Моретти [2013b] определяет этот термин как «взятие концепции и преобразование ее в серию операций». Он приписывает это определение Бриджмену [1927], который вводит этот термин в контексте измерения в физике: «Чтобы найти длину объекта, мы должны выполнить определенные физические операции. Таким образом, понятие длины является фиксированным, когда фиксированы операции, посредством которых измеряется длина: то есть понятие длины включает в себя столько же, сколько и ничего более, чем совокупность операций, посредством которых определяется длина». Хотя тематические модели являются несовершенным инструментом для измерения темы в литературе, они обеспечивают гораздо более мощное приближение темы, чем все, что у нас было раньше.

Но применение статистических моделей к литературе также ставит ряд проблем, которые подчеркивают объем человеческой интерпретационной работы, которая должна быть направлена на успешное тематическое моделирование. Литературные документы разной длины, описывают самодостаточные воображаемые миры и наполнены символическим языком. Мы можем решить эти проблемы с помощью курирования корпуса и интерпретативного прочтения моделей, но при этом мы обязательно должны признать тот факт, что мы не применяем фиксированный набор операций Бриджмена.

**6.7**  **Резюме**

Тематические модели сами по себе не могут изучать литературу, но они являются полезными *инструментами* для ученых, изучающих литературу. Модели обеспечивают четкую перспективу, которая может привлечь наше внимание к связям между различными частями корпуса, которые могут быть неочевидны при внимательном чтении. Литературные концепции сложны, но они часто имеют удивительно сильные статистические подписи. Модели по-прежнему могут быть полезны для выявления областей потенциального интереса, даже если они «не понимают» того, что обнаруживают. В то же время художественная литература бросает вызов практике моделирования, потому что

|  |  |
| --- | --- |
| *6.7. Резюме* | 79 |

каждое вымышленное произведение намеренно создает свой собственный закрытый мир — мир, персонажи и декорации которого настолько ярки, что могут затмить более тонкие связи между произведениями. Решение этих проблем может послужить приглашением к более глубокому размышлению о словах и их контекстуальном значении.

Точно так же, как тематические модели обеспечивают методологию для анализа творческого, разнообразного творчества авторов, а также эмоций и мыслей вымышленных персонажей, тематические модели также могут помочь нам понять реальных людей. Благодаря социальным сетям у нас есть огромное количество информации о людях, которые делятся своими мыслями и взглядами в Интернете. Тематические модели могут помочь нам использовать эти данные, чтобы лучше фиксировать эмоции, убеждения и отношения. В следующей главе обсуждается, как тематические модели могут понимать эти запутанные, интересные свойства текста.

**7**

**Вычислительная социальная наука**

В то время как предыдущие главы были в основном ретроспективными анализами, вычислительная социальная наука в основном находится «здесь и сейчас». Роль анализа текста состоит в том, чтобы предоставить данные о том, как люди относятся друг к другу и к своему окружению в определенных контекстах, например, социальных, политических или экономических взаимодействиях. Конкретное выражение того или иного документа обычно имеет меньшее значение. В результате социальная наука сосредотачивается на данных, полученных за самые последние часы, дни или недели, чтобы информировать аналитиков разведки, наблюдателей за брендами, журналистов или социологов. Однако основная проблема та же: эти заинтересованные стороны заинтересованы в том, что говорят люди, но не могут прочитать все имеющиеся в их распоряжении данные.

Исторически сложилось так, что социальные науки задают вопросы о мнениях. Какому кандидату отдают предпочтение в той или иной части страны? Нравится ли людям новый ресторан или продукт? На эти вопросы часто отвечают опросы: социологи отправляются в мир, собирают статистически значимую выборку опроса и экстраполируют результаты на более широкие слои населения.

Эти методы остаются основополагающими, но они требуют времени. Компания должна знать, есть ли у нее проблема с продуктом, немедленно,

80

81

особенно если его доброе имя поливают грязью в социальных сетях [Bowen, 2016]. Однако причина острой нехватки времени также может быть решением: если компания сможет быстро увидеть, что у нее есть проблема с социальными сетями, она сможет быстрее вмешаться и исправить проблему.

Традиционные методы социальных наук трудоемки, требуют много времени или невозможны для деликатных предметов. Например, исследования гриппа занимают слишком много времени, чтобы быть полезными по сравнению с жизненным циклом прогрессирования гриппа [Broniatowski et al., 2015], а рейтинги одобрения могут быть слишком низкими в преддверии выборов [O'Connor et al. ., 2010]. Использование поиска в Twitter и Google приводит к более быстрому получению более точной информации.

Прямое общение с некоторыми группами населения может быть затруднено. Использование социальных сетей представляет собой альтернативу [Wang et al., 2015], так как люди более свободно делятся информацией, чем официальные информационные агентства (которые, например, могут страдать от официальной цензуры в случае мнений о загрязнении в Китае) или в школах. проводимые опросы (которые могут страдать от самоцензуры в случае сексуальности или употребления наркотиков). Тематические модели и другие подходы к работе с большими данными, позволяющие анализировать огромное количество текста, помогают преодолеть некоторые препятствия на пути социальных наук с быстрым откликом.

Наблюдательный характер тематических моделей является одновременно и слабостью, и силой. Анализ документов через коллекцию социальных сетей может вызвать угрозу достоверности. Исследователи не обязательно могут контролировать группы, производящие документы, форум, на котором пишутся документы, или предмет документов. В то же время наблюдательные модели обладают преимуществом повышенного потенциала для открытий. В опросе исследователи должны указать каждый вопрос, который будет задан. Такой контроль хорош для обеспечения достоверности, но рискует упустить целые категории мнений, которые могут быть неочевидны для исследователей. Тематические модели могут дополнять этот тип тщательно разработанного опроса, выявляя проблемы или факторы, важные для участников, независимо от того, ожидались ли эти проблемы. Таким образом, неконтролируемые модели могут помочь в выявлении вопросов, которые исследователи «забыли задать».

82 *Вычислительные социальные науки*

**Прогнозирование и интерпретация** Общая тема при использовании тематических моделей заключается в том, должны ли модели отдавать приоритет *прогнозированию* или *интерпретации* . Различные тематические модели отдают предпочтение каждому из этих подходов. Различие между модельными приложениями в некоторой степени отражает различие между количественными и качественными социальными науками. Прогнозные приложения ближе к количественным методологиям, которые сосредоточены на методах, основанных на регрессии, которые имеют четкие входные и выходные переменные. Интерпретационные приложения ближе к качественным методологиям, использующим человеческую интуицию для объяснения сложных процессов. В то же время использование тематических моделей стирает границу между этими двумя методологиями. Даже когда они используются для поддержки качественной работы, тематические модели применяют вычисления и в определенной степени изолируют исследователей от предвзятых предубеждений (хотя они привносят свои собственные предположения при моделировании). Точно так же, даже когда они используются для поддержки количественной работы, тематические модели позволяют исследователям устанавливать количественные отношения между переменными метаданных документа и переменными, полученными из беспорядочного, неструктурированного текста, который в противном случае мог бы не поддерживать количественный анализ.

В предыдущих главах основное внимание уделялось интерпретации: может ли пользователь понять вывод модели? Но для контролируемых моделей возникает вопрос, насколько хорошо модель может предсказать некоторые интересующие параметры, такие как настроения или вовлеченность пользователей.

В какой-то степени они не всегда противоречат друг другу. Рэймидж и др. [2010a] показывают, что функции тематической модели могут улучшить категоризацию твитов, как это делают Blei and McAuliffe [2007] для контролируемого lda. Однако изменение целевой функции может еще больше улучшить прогнозы [Zhu et al., 2009].

Однако иногда улучшенная интерпретируемость (глава 3.4) препятствует способности модели предсказывать контент. Это относится как к словам внутри документа, так и к меткам документа. Чанг и др. [2009] показали, что сложные тематические модели лучше предсказывают протянутые документы, но имеют меньше смысла для пользователя. Нгуен и др. [2015a] показывают, что контролируемые модели предлагают лучшие прогнозы с дополнительными темами, но эти темы менее интерпретируемы.

|  |  |
| --- | --- |
| *7.1. Тематические модели для качественного анализа* | 83 |

**7.1**  **Тематические модели для качественного анализа**

Общей задачей качественной социальной науки является разработка теорий высокого уровня, объясняющих социальные процессы на основе наблюдений низкого уровня, таких как полевые отчеты или этнографические заметки. Одним из способов операционализации этого процесса является метод обоснованной теории [Glaser and Strauss, 1967]. Обоснованная теория описывает процесс итеративной разработки теорий посредством многократного чтения исходного материала. Баумер и др. [2017] сравнивают ручной анализ обоснованной теории и анализ на основе тематической модели текста ответа на опрос, описывающего опыт людей, пытающихся добровольно покинуть Facebook.

Они обнаруживают, что между этими двумя подходами существуют как теоретические, так и эмпирические связи. На теоретическом уровне как обоснованная теория, так и алгоритмы вероятностных тематических моделей являются итеративными, начиная с грубых, низкокачественных моделей/теорий и уточняя их путем многократного просмотра документов. Оба метода также стремятся поддерживать тесную связь между абстрактным представлением и исходным набором данных: в выборке Гиббса темы «основаны» на конкретных словарных лексемах. На эмпирическом уровне Baumer et al. [2017] находят тесные связи между темами, обнаруженными исследователями вручную, применяющими процедуры обоснованной теории, и темами, обнаруженными с помощью модели lda. Но этот результат не должен означать, что человеческий анализ не представляет никакой дополнительной ценности. Они сообщают, что тематическое значение тем лда не сразу становится очевидным из простых списков высокочастотных слов. Скорее, эта модель была наиболее полезна как способ предложить «список для чтения» по конкретной теме. «Смысл» тем был ясен только на теоретическом уровне при просмотре документов, которые имели необычно высокое представление этой темы. Таким образом, тематическая модель лучше всего может рассматриваться как инструмент для более эффективного применения обоснованной теории с большей изоляцией от человеческих предубеждений.

**7.2**  **Анализ настроений**

Полезно рассматривать применение тематических моделей в количественных социальных науках как средство получения числовой переменной из текста. Определенные темы в документе могут заменять темы, которые могут не

84 *Вычислительные социальные науки*

иначе легко измерить. Затем эти предполагаемые тематические переменные можно добавить к стандартным статистическим методам, чтобы найти связи между темами и нетекстовыми переменными. В качестве мотивирующего примера сначала рассмотрим анализ настроений [Pang and Lee, 2008]. Здесь цель состоит в том, чтобы определить «настроение» — например, положительное или отрицательное мнение — связанное с фрагментом текста. Например, «Чипотле великолепен!» будет ассоциироваться с положительным настроением, в то время как «от Chipotle меня тошнит» будет связано с отрицательным настроением.

В то время как коммерческие приложения анализа настроений в основном предназначены для определения того, нравится ли людям продукт или компания, существуют более широкие приложения в социальных науках для изучения больших корпусов для определения *внутреннего состояния авторов* . Например, политологи могут захотеть классифицировать пользователей социальных сетей как либеральных или консервативных на основе их онлайн-комментариев.

Тематические модели могут помочь в решении этих задач, разделив проблему на темы. Контекстное устранение неоднозначности, обеспечиваемое темами, может быть полезно для сужения диапазона применимых тем. Например, слово «Apple» может появляться как в технических новостях, так и в качестве пищевого ингредиента; кто-то, следящий за продавцом iPod и iPhone, не хотел бы, чтобы комментарии в социальных сетях жалуются на низкое качество Red Delicious. Контекстное устранение неоднозначности также может быть полезным, потому что подразумеваемая тональность слов может значительно различаться в разных доменах. Термин «удивительный» может быть положительным в обзоре книги, но резко отрицательным в обзоре автомобиля.

Тем не менее, анализ настроений на основе темы может быть проблематичным. Цель тематической модели состоит в том, чтобы идентифицировать и отделить скрытые факторы, которые лучше всего объясняют то, как сочетания слов появляются вместе в документах. Настроение может быть едва уловимым и не представлять для алгоритма сильного сигнала. Например, несмотря на то, что обзоры ресторанов специально предназначены для передачи настроений, тематические модели, обученные на обзорах ресторанов, в основном совпадают с *типами* ресторанов, создавая темы, представляющие кухни, такие как китайская, мексиканская и тайская. Тематические модели также теряют свою ценность, если вы хотите *противопоставить* настроения в теме. Хотя тематическая модель может найти людей, обсуждающих буррито Chipotle в Интернете, она не может отделить любителей от ненавистников. Таким образом, *различая*

|  |  |
| --- | --- |
| *7.3. Восходящие и нисходящие модели* | 85 |

темы, основанные на их настроении, могут помочь пользователю лучше понять, как темы и настроения взаимодействуют в наборе данных. Для этого требуется изменить модель темы, чтобы она знала об основных настроениях.

**7.3**  **Восходящие и нисходящие модели**

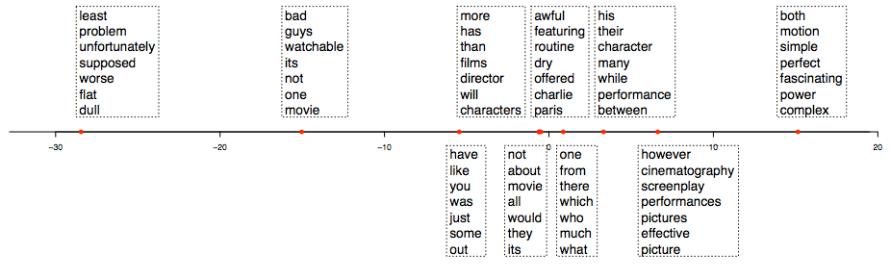
Чтобы различать темы на основе их настроений, модель должна учитывать нетекстовую переменную, которая представляет настроение. На языке вероятностных моделей настроение и тема моделируются *совместно* . То есть для каждого документа существует распределение вероятностей как по переменной тональности *y , так* и по темам *z* .

Существует два основных типа совместных моделей, которые включают метаданные, такие как настроения: восходящие и нисходящие модели. Различие основано на генеративной истории тематических моделей (глава 1): в генеративной истории настроения предшествуют (вверх по течению) или после (вниз по течению) темам?

Восходящие модели предполагают, что внешние переменные, такие как настроения, стоят на первом месте в порождающей истории. То есть будут разные темы, учитывая основное настроение. Это может происходить в виде жестко запрограммированного распределения [Mei et al., 2007a], априорного определения настроений [Mimno and McCallum, 2008] или скрытой переменной, которая может служить показателем настроений [Lin and Он, 2009]. Восходящие модели часто легче реализовать и они более гибкие [Roberts et al., 2014], поскольку им не нужно задавать генеративное распределение, соответствующее форме переменной.

Напротив, нижестоящие модели явно предсказывают переменные, такие как настроение, *заданное* текстом. Если цель состоит в том, чтобы позже предсказать настроение по необработанному тексту с помощью тематических моделей, нижестоящие модели могут работать лучше, чем восходящие модели. Эти модели часто называют «контролируемыми» тематическими моделями после контролируемого lda [Blei and McAuliffe, 2007, slda], которые используют темы документа для прогнозирования нисходящей тональности с помощью регрессии: предполагается, что тональность документа *y d* исходит из распределения Гаусса с означает *η* > *z* ¯, где *z* ¯ — нормализованный вектор всех тем, используемых в документе, а *η* — параметр регрессии, описывающий настроение каждой темы.

86 *Вычислительные социальные науки*



**Рисунок 7.1:** Примеры тем, изученных лда под наблюдением из Blei and McAuliff [2007]. Каждая тема представляет собой не просто набор слов, но также имеет показатель регрессии *η* , который объясняет, связана ли она с положительным настроением (справа) или отрицательным настроением (слева).

Во время вывода слова и настроения работают вместе, чтобы найти комбинации темы и настроения, которые имеют смысл. В то время как «ванильные» тематические модели стремятся найти кластеры слов, которые имеют смысл вместе, если тема связана с документами, которые имеют много разных значений тональности, ей придется изучить менее сфокусированное распределение по баллам тональности, что приводит к более низкой вероятности.

Рассмотрим рисунок 7.1. Если тема имеет непоследовательное значение тональности (например, документ отрицательной тональности в теме положительной тональности), вывод попытается переместить документы негативной тональности в темы с согласованной тональностью *η* **и** последовательные слова.

Эти модели составляют основу моделей и проблем, которые мы обсуждаем в оставшейся части этого раздела.

**7.4**  **Понимание стойки и поляризации**

Другой формой внутреннего состояния является *позиция* : какую сторону занимает человек в вопросе. Это может принимать разные формы: вы за или против предложения, вы демократ или республиканец, или вы поклонник оригинального «Звездного пути» или новой версии?

Восходящие модели могут обнаружить эти стороны, включив позицию в генеративную модель. Например, несколько авторов — Zhai et al. [2004], Лу и Чжай [2008] и Пол и Гирью [2010] — развитие темы

|  |  |
| --- | --- |
| *7.5. Социальные сети и СМИ* | 87 |

модели, которые позволяют читателям сравнивать аспекты темы. Они утверждают, что каждая сравнительная «сторона» имеет распределение слов, которые она обычно использует , *и* что у каждой стороны есть свое мнение о том, как она обсуждает тему. В документе каждое слово выбирается либо из фонового распределения стороны, версии темы стороны, либо из «нейтральных» слов темы. Например, и израильтяне, и палестинцы используют термины «нападения», «гражданские лица» и «военные» при обсуждении беспорядков в оккупированной Израилем Палестине, но израильская сторона использует термины «террорист» и «подстрекательство», в то время как палестинская сторона сосредотачивается на «сопротивлении». » и «оккупация».

Взаимодействие между чувством и аспектом неясно. Некоторые аспекты не зависят от настроения, а другие аспекты особенно заряжены. Джо и О [2011] разрабатывают модель, которая сначала рисует распределения настроений, которые ограничивают аспекты, обсуждаемые в документе.

Модели нисходящего потока также могут охватывать эти подразделения. Нгуен и др. [2013] предсказывают, является ли спикер республиканцем или демократом 1 , основываясь на версиях тем, которые он обсуждает, расширяя непрогностическую модель Гриммера [2010]. Например, республиканцы чаще обсуждают налоги, чем демократы, но демократы сосредотачиваются на пользе, получаемой от налогов (рис. 7.2).

Однако не всегда есть две стороны вопроса. Вероятностным решением этой модели является вложенный процесс Дирихле [Blei et al., 2010]. Эти иерархии создают непараметрическую иерархию по неограниченному числу тем. Это соответствует повестке дня из политической науки [Nguyen et al., 2015b].

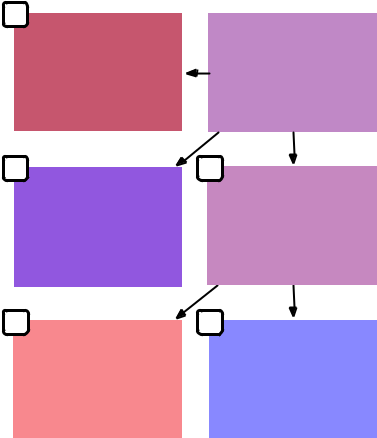
**7.5**  **Социальные сети и СМИ**

Мы говорили о метаданных, независимых для каждого пользователя. Однако иногда нас интересуют метаданные, описывающие отношения *между* документами: какие пользователи следят друг за другом

* В последующих моделях вопрос о том, какие переменные использовать в прогнозе, часто является предметом споров. Использование лексических терминов [Titov and McDonald, 2008; Zhao et al., 2010] в логарифмической модели обычно работает лучше: она способна уловить характерные для слова нюансы тональности и смоделировать ситуационную тональность (например, «непредсказуемый» — это хорошо). для книги, но плохо для руля автомобиля). Однако это приводит к более сложной и менее интерпретируемой модели.

88 *Вычислительные социальные науки*

Е



конституция флага

свобода Supreme\_court

избирательные права

преемственность

конституционная\_поправка

энт

Р: 1,1

А

минимальная\_заработная плата

комиссия

Independent\_commissio

1. расследовать ураган\_катрина

усилить расследование

Д:1.7

С

вакансий

бизнес бизнес

семейный\_бизнес

производительность оборудования

repeal\_permanency

работники капитальных ферм

Р: 4,3

выставление счетов за время выступающего

поправка

председатель люди



джентльмен

законодательство

поддержка конгресса

Р:0

Б

процентный налог экономика недвижимость\_налог капитал\_прибыль деньги налоги предприятия семьи налог\_сокращение платить налог\_льготная\_помощь социальное\_обеспечение

Р:0,4

1. миллиард детей бюджет сокращает долг tax\_cuts

дефицит поддержки детей образование студенты здравоохранение республиканцы национальный\_долг

Д:4,5

gses кредит\_рейтинг



регулятор fannie\_mae

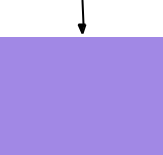
рынок freddie\_mac

финансовые\_услуги

конкурс агентств

инвесторы фанни

Р: 1,0



доступное жилье

фонд управляющего жильем

фонды деятельности

организации

избиратель\_регистрация

религиозные некоммерческие организации

Д:2.2

РЕСПУБЛИКАНСКИЙ



ДЕМОКРАТ

**Рисунок 7.2:** Темы, выявленные в ходе дебатов в зале Конгресса с использованием модели нисходящего потока для захвата идеологии оратора. Многие темы первого уровня выделены двухпартийным цветом (фиолетовый), в то время как темы более низкого уровня связаны с определенными идеологиями (синий цвет демократов, красный цвет республиканцев). Например, тема «налоги» (B) является двухпартийной, но ее дочерняя часть с демократическим уклоном (D) фокусируется на социальных целях, поддерживаемых налогами («дети», «образование», «здравоохранение»), а ее республиканская Начинающий ребенок (С) фокусируется на последствиях для бизнеса («налог на смерть», «рабочие места», «бизнес»). Число под каждой темой обозначает величину изученного параметра регрессии, связанного с этой темой. Цвета и числа под каждой темой показывают параметр регрессии *η ,* связанный с этой темой. Из Нгуен и др. [2013].

Twitter, какие научные статьи цитируют друг друга или какие веб-страницы ссылаются друг на друга. Это усложняет моделирование, но мы по-прежнему видим то же разделение между восходящими и нисходящими моделями: восходящие модели предполагают, что сообщества формируются до того, как мы видим слова, тогда как нисходящие модели используют слова для объяснения того, какие связи мы видим.

Стохастическая блочная модель [Holland et al., 1983] и ее потомок со смешанным составом [Airoldi et al., 2008] являются прототипами восходящих моделей. Они утверждают, что существуют внутренние группы документов, и ссылки более вероятны внутри группы, чем вне группы. Эти группы аналогичны темам в тематических моделях, за исключением того, что ссылки «общие» между документами.

|  |  |
| --- | --- |
| *7.5. Социальные сети и СМИ* | 89 |

Однако первые вероятностные модели сетевой структуры игнорировали слова в документах. Поскольку сетевая структура связана с личностью автора, естественно объединить личность автора с моделью восходящего потока McCallum et al. [2007], Лю и соавт. [2009], обусловливая темы авторами и сообществами, к которым принадлежат авторы.

Link lda является образцом для последующих моделей [Nallapati and Cohen, 2008] и включает текст в документы. Он использует регрессию распределения тем ( *θ* ), а не назначений тем ( *z* ), в отличие от контролируемого lda выше. Точно так же Ча и Чо [2012] используют отслеживаемых пользователей для моделирования последующих документов.

Тем не менее, обусловленность тематических назначений может улучшить способность алгоритма предсказывать ссылки в протянутых документах [Chang and Blei, 2009]. Это связано с тем, что регрессия, основанная только на распределении, может использовать темы для объяснения ссылок, которых нет в документе. Например, если модель считает, что между документами существует связь, потому что они оба используют тему 14, но в документе нет назначенных слов ( *z n* = 14), то модель не может воссоздать этот прогноз в отложенном документе.

Не все тематические модели, применяемые к социальным сетям, пытаются предсказать ссылки. Венг и др. [2010] используют сетевую структуру Twitter, чтобы определить, кто имеет влияние в рамках темы, а некоторые модели используют ссылки, чтобы ограничить документы, связанные друг с другом, чтобы они были похожими [Mei et al., 2008, Sun et al., 2009, Daumé III , 2009].

В дополнение к явным связям в социальных сетях социальные сети также формируются неявными связями между людьми в сходных контекстах — событиями, культурными или региональными моделями, которые влияют на то, как люди говорят и о чем они говорят. Мэй и др. [2006] фиксируют, как темы меняются в зависимости от региона и времени (например, когда ураган обрушивается на регион, те, кто находится ближе всего к эпицентру шторма, будут говорить об этом больше и конкретнее). Более поздняя работа строит модели местоположения и темы совместно [Yin et al., 2011].

Напротив, Эйзенштейн [2017] фокусируется на лексической изменчивости, фиксируя, как неологизмы социальных сетей, такие как «af» (усилитель постпозиции, особенно прилагательных: например, «описание было уклончивым af»), распространялись из двойных эпицентров в южных странах. Калифорния и Атланта на всю территорию США.

90 *Вычислительные социальные науки*

**7.5.1**  **Особенности социальных сетей**

Хонг и Дэвисон [2010] обсуждают, как короткие документы платформ социальных сетей, таких как Twitter, могут запутать алгоритмы тематического моделирования, а Zhao et al. [2011] расширяют анализ, показывая тематические различия (например, Твиттер часто кратко следит за мимолетными темами страстно). Поскольку документ ограничен 140 символами, допущения о примесях тематических моделей ограничены. Чтобы отслеживать тенденции во времени или между пользователями, алгоритмы также должны знать связи между пользователями или группировать сообщения во времени [Mehrotra et al., 2013]. Другие исследователи разрабатывают модели с особыми свойствами разреженности [Lin et al., 2014], чтобы учесть особенности Twitter.

**7.6**  **Резюме**

Вычислительная социальная наука может раскрыть эмоции и скрытые фракции, часто присутствующие в онлайн-дискуссиях. Это полезно для компаний, пытающихся понять своих клиентов, для политиков, пытающихся ориентироваться на избирателей, для служб экстренного реагирования, реагирующих на стихийные бедствия [Киреев и др., 2009], и для ученых, пытающихся понять, как онлайн-коммуникация трансформирует социальные нормы.

Однако, поскольку социальные сети все больше охватывают весь земной шар, предположение, что тематическая модель существует только на одном языке, часто является неверным предположением. Действительно, даже в пределах одной страны тематические модели могут обнаруживать региональные различия [Eisenstein et al., 2010]. В следующей главе мы обсудим, как справляться с многоязычными наборами данных и при этом находить связные, независимые от языка темы.

**8**

**Многоязычные данные и машинный перевод**

До сих пор мы фокусировались на одноязычных тематических моделях и их приложениях. Но многие коллекции содержат документы более чем на одном языке. На практике мы часто обнаруживаем это явление неожиданно после запуска начальной одноязычной тематической модели: тематические модели оказываются очень хорошими при идентификации языка. Такое поведение имеет смысл, поскольку модель ищет группы слов, которые часто встречаются вместе, но не в других контекстах, а отдельные языки обладают этим свойством. В некоторых случаях мы можем отфильтровать небольшое количество документов на других языках, но мы хотели бы воспользоваться преимуществами связей между многими языками.

Многоязычные тематические модели были разработаны для анализа и понимания корпуса на нескольких языках. Вулич и др. [2015] дают хороший обзор. Приложения в многоязычных корпусах можно разделить на две категории. Первая и более простая категория — это те, которые выравнивают языки на тематическом уровне, а не на уровне отдельных типов слов. Эти модели полезны для организации корпусов, но не пытаются поддерживать анализ для пользователей, незнакомых с каким-либо конкретным языком. Вторая категория — это те, которые явно моделируют выравнивание на уровне слов в разных языках. Эти модели поддерживают приложения в

91

92 *Многоязычные данные и машинный перевод*

статистический машинный перевод (СМТ).

Как одно из наиболее частых приложений для многоязычных тематических моделей, smt пытается найти последовательность слов на одном языке, которая соответствует значению введенного текста на другом языке. В то время как обучающие данные smt требуют явно выровненных предложений на разных языках, многоязычные тематические модели ослабляют это ограничение данных и позволяют исследовать только слабо выровненные данные. В этой главе мы сначала обсудим, как тематические модели адаптируются для использования нескольких языков, а затем покажем, как эти многоязычные тематические модели могут помочь smt.

Прежде чем обсуждать конкретные методы, полезно определить термины, относящиеся к источникам данных. Наиболее заметной характеристикой многоязычных корпусов является степень их выравнивания. Параллельные корпуса наиболее тесно связаны друг с другом. Эти коллекции состоят из подмножеств документов, каждое из которых содержит документы на разных языках с одинаковым семантическим содержанием (вплоть до пределов перевода). Общие примеры включают переводы литературных произведений или переведенные правительственные документы, где стенограмма речи на французском языке сопровождается стенограммой той же речи на немецком языке с минимально возможной семантической разницей. Сопоставимые корпуса менее тесно связаны. Эти коллекции также содержат подмножества документов, но каждый набор ограничен только *тематическим* сходством, а не обязательно прямым переводом. Типичным примером являются статьи в Википедии. Статьи о французском городе Лилль в английской и французской Википедии относятся к одному и тому же месту и содержат большую часть той же информации, но французская версия значительно длиннее. Смешанные корпуса наименее согласованы. Эти коллекции просто содержат документы более чем на одном языке, но не обязательно существует какая-либо связь между любым документом на одном языке и документом на другом языке. Примером может служить журнал, издаваемый на английском, французском, немецком и итальянском языках. Ни одна статья не является переводом какой-либо другой статьи. Тематические совпадения между статьями могут быть, но не обязательно какие-либо структурные признаки таких отношений. Последняя категория полезных данных, не обязательно в форме документов, — это двуязычный словарь, который сопоставляет слова одного языка со словами другого. Лексиконы этой формы можно рассматривать как примеры параллельных корпусов с одноточечными документами, но часто

|  |  |
| --- | --- |
| *8.1. Выравнивание на уровне документа из многоязычных корпусов* | 93 |

полезно относиться к ним специально.

**8.1**  **Выравнивание на уровне документа из многоязычных корпусов**

В случае, когда пользователь просматривает многоязычную коллекцию, обладая только одноязычным знанием, чтобы найти соответствующие документы, могут помочь многоязычные тематические модели. Такие коллекции содержат несколько языков, но не обязательно имеют точное соответствие или перевод слов и предложений. Необходимо только грубое выравнивание документов, если в них обсуждаются одни и те же темы, например, статьи Википедии на разных языках. Такая связь между языками также полезна для вывода более надежных тем, поскольку разные языки могут дополнять друг друга, чтобы уменьшить двусмысленность.

Этот подход предшествует вероятностным тематическим моделям. Ландауэр и Литтман [1990] соединяют выровненные документы на разных языках, проецируя оба документа в общее пространство скрытой семантической индексации.

Точно так же двуязычные тематические модели [Zhao and Xing, 2006, De Smet and Moens, 2009] и, более конкретно, многоязычное латентное распределение Дирихле [Mimno et al., 2009, plda] предполагают, что выровненные документы на разных языках имеют общие одинаковое распределение тем, и каждый язык имеет уникальное распределение тем по типам слов. Таким образом, порождающий процесс полиязычной тематической модели выглядит следующим образом: дана пара документов ( *d l* 1 *, d l* 2 ), мы сначала выбираем распределение по теме документа *θ d* ; для документа *d l i* на языке *l i* мы затем выбираем тему *z dn* из *θ d* и генерируем слово из темы *φ z dn ,l i* на языке *l i* .

Тематические модели, обученные на основе выравнивания на уровне документов, находят применение в исследовательском анализе данных и поиске информации [Vulić et al., 2013]. Мимно и др. [2009] используют модель, обученную на нескольких языках в Википедии, для сравнения относительного интереса к различным темам в разных языковых областях. Например, в Википедии на персидском языке больше, чем в среднем, статей о науке, а в Википедии на финском языке больше, чем в среднем, статей о лыжах. Эти методы требуют параллельных или сопоставимых корпусов, но для смешанных корпусов обучающие данные могут быть дополнены дополнительным корпусом сопоставимых документов при условии, что сопоставимые

94 *Многоязычные данные и машинный перевод*

документы охватывают достаточно схожие темы [Mimno, 2012].

Не обязательно понимать «язык» в его строгом значении. Слабо выровненные модели применялись при поиске информации для расширения запроса [Gao et al., 2011, 2012]. В то время как применение тематических моделей в поиске информации (глава 2) сосредоточено на одном языке, Gao et al. [2011] предполагают, что запросы и веб-документы написаны на разных «языках». Язык запросов от пользователей обычно представляет собой неформальный устный язык, который менее форматирован и может также включать аббревиатуру. Однако язык документов более формальный и хорошо организованный письменный язык. Например, при запросе «dtd amc» в соответствующем веб-документе может содержаться «downtown disney amc» [Jiang et al., 2016].

Этот семантический разрыв [Müller and Gurevych, 2009] между запросами и документами дает возможность рассматривать запросы и документы как разные языки, а релевантность между запросами и документами делает их слабо выровненными. Основываясь на этом предположении, они также предполагают, что запросы и документы имеют одно и то же распределение тем документа *θ Q* , но имеют разные распределения слов темы *φ Q z* и *φ D z* соответственно. Таким образом, документы и запросы связаны через скрытые темы, несмотря на то, что их словари (распределение слов темы) различны. Суммируя все возможные темы, отношение между термином документа *e* и запросом *q* равно

|  |  |
| --- | --- |
| Икс |  |
| *п* ( *е* | *q* ) знак равно *п* ( *е* | *φ k D* ) *п* ( *k* | *θ q* ) *.* | (8.1) |

*к*

Некоторые формы расширения запросов на несколько языков не требуют явного моделирования связей между темами [Vulić et al., 2011a, 2015]. Как отмечалось в главе 2, Эрлин [Erlin, 2017] использует две независимые исходные модели для английских и немецких книг для поиска работ по эпистемологии. После ручной идентификации одной эпистемологической темы из каждой языковой модели эти две темы используются в качестве формы расширения запроса для идентификации документов, связанных с целевым предметом.

Хотя сравнительно легко получить сопоставимые темы из сопоставимых корпусов, определить конкретные слова, которые являются переводами друг друга на разных языках, сложнее. Учитывая слово *we* , которое имеет высокую вероятность в теме *k* языка *e* , вполне вероятно, что хороший перевод *слова we в* языке *f также будет* иметь высокую вероятность в теме *k* . Вулич и др.

|  |  |
| --- | --- |
| *8.2. Выравнивание на уровне слов из лексических данных* | 95 |

[2011b] оценивают несколько методов определения таких пар перевода с учетом двуязычной или многоязычной тематической модели. Они нашли два хорошо работающих метода, оба из которых учитывают частоту данного целевого слова во многих темах. Интуиция подсказывает, что слова с высокой вероятностью в данной теме, потому что они специфичны для этой темы, с большей вероятностью будут хорошими переводами, чем слова с высокой вероятностью в теме, потому что они часто встречаются в корпусе в целом и, следовательно, представлены во многих темах. Затем авторы смогли вывести алгоритм для поиска высококачественных выровненных пар перевода [Vulić and Moens, 2012]. Этот метод позволяет использовать образцы слов в качестве подсказок для этимологически связанных слов, когда они доступны, но он также эффективен даже для неродственных языков. Однако наша способность находить пары прямого перевода с выравниванием только на уровне документа ограничена как из-за недостаточности данных, так и из-за того, что языки могут выравниваться больше на уровне *понятий* , чем на уровне конкретных лексических единиц [Vulić and Moens, 2014] . ].

**8.2**  **Выравнивание на уровне слов из лексических данных**

Выровненные документы полезны, когда коллекция спроектирована так, чтобы быть слегка многоязычной: например, когда создатели создают версию Википедии на родном языке. Ссылки на документы дешевы и просты. Однако они требуют активной поддержки создателей коллекции, что не всегда возможно. Многие сборники пишутся изолированно.

Однако одним из самых распространенных многоязычных инструментов является словарь. В этом разделе обсуждается, как мы можем использовать *лексическую информацию* , такую как многоязычные словари [Zhang et al., 2010], и орфографические отношения между словами [Boyd-Graber and Blei, 2009], чтобы помочь пользователям, которые хотят понять коллекцию.

Например, древовидные тематические модели, такие как древовидное латентное распределение Дирихле [Boyd-Graber et al., 2007, Andrzejewski et al., 2009, tlda], включают положительные корреляции между словами в одном и том же или разных языках, поощряя слова, которые появляются вместе в **понятии** , чтобы иметь схожие вероятности для данной темы. 1 Эти понятия

1 Чжан и др. [2010] используют мягкие ограничения на уровне темы для достижения аналогичного эффекта.

96

Словарь:



компьютер,

рынок,市

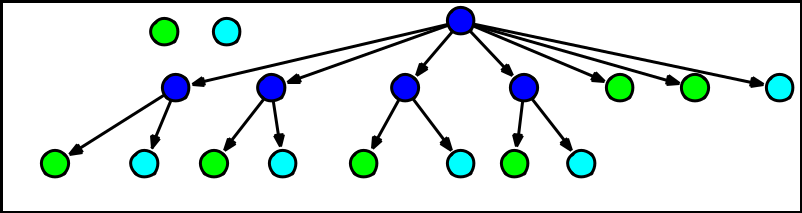
правительство,政府

наука,科学

*Многоязычные данные и машинный перевод*

Словарный запас: английский (0), китайский (1)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0 | компьютер | 0 | научный 1 | | 政府 |
| 0 | рынок | 0 | политика 1 | | 科学 |
| 0 | правительство 1 | | 市 | 1 | 天气 |
| 0 | наука | 1 |
|  |  |
|  |  |  |  |  |  |



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Предыдущее дерево: | 0 | 1 |  |  |  |
|  |  |  |  | научная политика | 天气 |
| компьютер | рынок | 市 | правительство | 政府наука科学 |  |

**Рисунок 8.1:** Пример построения априорного дерева из двуязычного словаря: пары слов с одинаковым значением, но в разных языках являются концептами; общий родительский узел создается для группировки слов в понятие, а затем соединяется с корнем; несоответствующие слова связаны с корнем напрямую.

могут поступать из WordNet [Boyd-Graber and Resnik, 2010], экспертов по предметной области [Andrzejewski et al., 2009] или пользовательских ограничений [Hu et al., 2014a]. Если эти понятия находятся на одном языке, внутренняя модель будет такой же, как одноязычное интерактивное тематическое моделирование, представленное в главе 3. Однако когда мы собираем понятия из двуязычных ресурсов, эти понятия могут связывать разные языки. Например, если двуязычный словарь определяет «5» как «компьютер», мы объединяем эти слова в понятие.

Эти концепции (положительные корреляции) организованы в **априорную древовидную** структуру. Слова в одном и том же понятии имеют общего предкаузел (рис. 8.1). Затем это понятие становится одним из многих дочерних элементов корневого узла. Слова, не входящие ни в одно понятие, — **несоотнесенные слова** — напрямую связаны с корневым узлом. Таким образом, тема становитсяраспределение по всем путям в этом предшествующем дереве, и каждый путь связан со словом.

Вероятность пути в теме зависит от вероятностей перехода в теме. Каждое понятие *i* в теме *k* имеет распределение по своим дочерним узлам, которое регулируется априорным правилом Дирихле: *π k,i* ∼ Dir( *β i* ). Каждый путь заканчивается словом (т. е. конечным узлом), а вероятность пути является произведением всех переходов между темами, которые он пересекает.

|  |  |
| --- | --- |
| *8.3. Выравнивание из параллельных корпусов и лексической информации* | 97 |

Темы имеют корреляции со словами, потому что параметры Дирихле могут кодировать положительные или отрицательные корреляции [Andrzejewski et al., 2009].

В результате для выборки слова *w dn* по заданной теме *z dn выбирается* путь *y dn* из дерева тем темы *z dn* : мы начинаем с корня *n* 0 и сначала выбираем дочерний узел *n* 1 корня; если узел *n* 1 является узлом понятия, мы продолжаем выборку узла слова *n* 2 и генерируем слово, связанное с *n* 2 ; если узел *n* 1 уже является узлом слова, мы генерируем слово напрямую.

Когда это дерево служит априором для тематических моделей, слова в одном концепте имеют положительную корреляцию в темах. Например, если «5» имеет высокую вероятность в теме, то же самое будет и «компьютер», поскольку они имеют один и тот же родительский узел. С древовидными априорами каждая тема больше не является распределением по типам слов; вместо этого это распределение по путям, и каждый путь связан с типом слова. Одно и то же слово может встречаться в нескольких путях, и каждый путь представляет уникальное значение этого слова.

**8.3**  **Выравнивание из параллельных корпусов и лексической информации**

Двуязычные словари и другие источники информации на уровне слов полезны при обучении многоязычных моделей, потому что они могут легко определять простые лексические отношения, которые может быть трудно извлечь из параллельных корпусов. Но такие сгенерированные вручную данные могут быть ненадежными, некачественными или в них могут отсутствовать контекстуальные различия при фактическом использовании. Однако эти два подхода не исключают друг друга; они обнаруживают различные связи между языками. Ху и др. [2014b] объединяют существующее древовидное скрытое распределение Дирихле (tlda) и многоязычное скрытое распределение Дирихле (plda) и создают многоязычное древовидное латентное распределение Дирихле (ptlda), которое включает корреляции на уровне слов и выравнивание на уровне документа. Информация.

Чтобы построить предыдущую древовидную структуру, Hu et al. [2014b] рассматривают два ресурса, которые соотносят слова в разных языках. Первый — это многоязычные словари, которые сопоставляют слова с одинаковым значением на разных языках вместе. Другой - это выравнивание слов, извлеченное из выровненных предложений в параллельном корпусе. Эти отношения между словами

98 *Многоязычные данные и машинный перевод*

используемые в качестве понятий [Bhattacharya, 2006] в предшествующем дереве (рис. 8.1). Учитывая предшествующую древовидную структуру, генерация документов представляет собой комбинацию tlda и plda. Для каждой выровненной пары документов ( *d l* 1 *, d l* 2 ), мы сначала выбираем распределение по темам *θ d* из априорного Dir( *α* ) Дирихле. Для каждого маркера в выровненном документе *d l i* мы сначала выбираем тему *z dn* из полиномиального распределения *θ d* , а затем выбираем путь *y dn* по дереву темы *z dn* . Поскольку каждый путь *y dn* ведет к слову *w dn* в языке *l dn* , мы добавляем выбранное слово *w dn* к документу *d l dn* в

язык *л дн* .

Если мы используем плоский симметричный априор Дирихле вместо априорного дерева, модель эквивалентна plda. Точно так же, если все документы одноязычны (т. е. с различным распределением по темам *θ* ), модель эквивалентна tlda. ptlda связывает разные языки как на уровне слов (используя сопоставление слов), так и на уровне документа (используя выравнивание документов), таким образом, он лучше изучает темы, рассматривая больше информации из обоих языков.

**8.4**  **Тематические модели и машинный перевод**

Наиболее часто многоязычные тематические модели применяются в машинном переводе. При вводе текста на одном языке (исходном языке) статистический машинный перевод пытается найти аналогичную последовательность слов на другом языке (целевом языке). Современные системы машинного перевода [Koehn, 2009] используют миллионы обучающих примеров для изучения правил перевода и применения этих правил к тестовым данным. Тематические модели полезны в этом приложении, когда они могут помочь сообщить значение слова и выбор слова в определенных контекстах. Хотя правила перевода изучаются в локальном контексте, эти системы работают лучше всего, когда учебный корпус имеет согласованную *область* , такую как жанр (например, спорт, бизнес) или стиль (например, лента новостей, сообщения в блогах).

Переводы в пределах одного домена лучше, чем переводы между доменами, поскольку они сильно различаются по выбору слов и стилю. Правильный перевод в одном домене может быть неуместным в другом домене. Например, «\ **4** » в спортивном домене обычно означает «подводное плавание», но в домене социальных сетей это означает не-

|  |  |
| --- | --- |
| *8.5. Компоненты статистического машинного перевода* | 99 |

содействующий «скрытень». Чтобы избежать таких ошибок перевода, вызванных сдвигом домена, перевод поезда должен быть устойчивым к таким систематическим изменениям в обучающей выборке. Это называется *адаптацией домена* .

Для обучения таких smt-систем ранние усилия были сосредоточены на построении отдельных моделей с учетом помеченных вручную доменов [Foster and Kuhn, 2007, Matsoukas et al., 2009, Chiang et al., 2011]. Однако эта установка в лучшем случае дорогая, а в худшем — невозможна для больших данных. Тематические модели представляют собой многообещающее решение, позволяющее автоматически обнаруживать домены. Каждая извлеченная тема рассматривается как программный домен. 2 Таким образом, обычные одноязычные тематические модели, обученные только на исходных документах, применялись для извлечения знаний предметной области для машинного перевода [Eidelman et al., 2012].

Однако исходный язык и целевой язык могут дополнять друг друга для создания более точных тематических моделей. Например, если мы знаем только китайскую фразу «\ **4** », трудно решить, является ли это доменом спорта или доменом социальных сетей . Однако с помощью выровненного английского перевода «lurker» легко определить домен «социальные сети». Так, многоязычные тематические модели [Ni et al., 2009; De Smet and Moens, 2009] применялись для извлечения знаний предметной области для машинного перевода [Hu et al., 2014b].

**8.5**  **Компоненты статистического машинного перевода**

Статистический машинный перевод представляет перевод как комбинацию вероятностных процессов, модели перевода на уровне фразы и языковой модели на уровне предложения [Koehn et al., 2003, Koehn, 2009]. Тематические модели применялись к обоим аспектам этого процесса.

Фактический процесс перевода также известен как *декодирование* , которое заключается в поиске наилучшего целевого предложения перевода **e** по заданному исходному предложению **f** . Формально, учитывая исходное предложение **f** , лучший перевод в целевом **лучший** язык \_ \_

**e** лучший = **argmax e** *p* ( **e** | **f** ) = **argmax e** *p* ( **f** | **e** ) *p* ( **e** ) *,*

(8.2)

который разделен на *модель перевода p* ( **f** | **e** ) и *языковую модель*

*р* ( **е** ).

* В дальнейшем мы будем использовать термины «тема» и «домен» взаимозаменяемо: «тема» для обозначения распределения слов в тематических моделях и «домен» для обозначения smt-корпусов.

100 *многоязычных данных и машинный перевод*

Интуитивно, хороший перевод должен быть как хорошим соответствием исходному предложению (высокий балл в модели перевода), так и хорошим предложением сам по себе (высокий балл в языковой модели).

На этапе *декодирования* исходное предложение **f** сегментируется на несколько исходных фраз *f* ¯ *n* , которые переводятся в набор целевых фраз *e* ¯ *n* . Таким образом, вероятность перевода *p* ( **f** | **e** ) может быть дополнительно разложенавероятности перевода фразы *p* ( *f* ¯ *n* | *e* ¯ *n* ). На этапе *переупорядочивания* может потребоваться изменить положение целевых фраз, чтобы получить наилучший результат перевода. Изменение порядка фиксируется относительным распределением вероятности искажения *d* ( *a n* − *b n* −1 ), где *a i* обозначает начальную позицию исходной фразы, которая была переведена в *n* -ю целевую фразу, а *b n* −1 обозначает конечная позиция исходной фразы, переведенная в ( *n* − 1) -ю целевую фразу. В результате модель перевода распадается на

*p* ( **f** | **e** ) = *p* ( *f* ¯ *n* | *e* ¯ *n* ) *d* ( *a n* − *b n* −1 ) (8.3)

Д

*н*

В smt, основанном на фразе, вероятность фразы *p* ( *f* ¯ *n* | *e* ¯ *n* ) может быть дополнительно оценена путем комбинирования вероятностей лексического перевода слов, содержащихся в этой фразе [Koehn et al., 2003], которая обычно пересчитывается. называется *лексическим взвешиванием* . Лексические условные вероятности *p w* ( *f* | *e* ) являются оценками максимального правдоподобия по относительным лексическим частотам,

.

*p ш* ( *f* | *e* ) = *c* ( *f, e* ) P *f c* ( *f, e* ) (8.4)

где *c* ( *f, e* ) — количество наблюдаемых лексических пар ( *f, e* ) в наборе обучающих данных. Учитывая выравнивание слов *a* , лексический вес для этой пары фраз *p w* ( *f* ¯ | *e* ¯; *a* ) представляет собой нормализованное произведение лексических вероятностей выровненных пар слов в этой паре фраз:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *п ш* ( *ж* ¯ | *е* ¯; *а* ) знак равно |  |  |  | 1 |  |  |  | Х ) ∈ | *п ш* ( *ж я* | *е j* ) | (8.5) |
|  |  |  |  |  |  |  |
| {| | *Дж* | | | ( *я, дж* ) | *а* | | ∀ ( |
| Д |  |  | ∈ }| | | *а* |  |
| *я* |  |  |  |  |  |  |  | *я, дж* |  |

где *i* и *j* - позиции слов в целевой фразе *e* ¯ и исходной фразе *f* ¯ .

Далее мы расскажем, как применять тематические модели для улучшения моделей перевода, языковых моделей и моделей переупорядочения соответственно.

|  |  |
| --- | --- |
| *8.6. Тематические модели для перевода на уровне фраз* | 101 |

**8.6**  **Тематические модели для перевода на уровне фраз**

Модели перевода сопоставляют слова и фразы с одного языка на другой. Как одноязычные тематические модели, так и двуязычные тематические модели полезны для улучшения моделей перевода. Как мы упоминали в разделе 8.4, наиболее заметное применение тематических моделей — *адаптация предметной области* .

Ранние работы по извлечению знаний о предметной области были сосредоточены на доменах, помеченных вручную [Foster and Kuhn, 2007, Matsoukas et al., 2009, Chiang et al., 2011]. Эти этикетки не только дороги и требуют много времени для получения, но также не сглажены и чувствительны к ошибкам и несоответствиям маркировки. Кроме того, такие жесткие метки доменов трудно применять, и они могут снизить надежность переводов: домены фундаментально неопределенны, и если вы ошибетесь в домене, вы можете отсечь полезную информацию.

Тематические модели обеспечивают способ **автоматического** обнаружения назначений **программных** доменов. Если мы приравняем *K* распределений тем по словарю в тематической модели к *K* доменам smt, распределение тем каждого документа можно рассматривать как присвоение мягкого домена для этого документа. Если есть две темы: спорт и социальные сети , а тестовый пример, скорее всего, посвящен спорту, он может иметь мягкое распределение домена: 85% для спортивного домена и 15% для домена социальных сетей . Эти автоматически получаемые программные метки доменов хорошо сглажены, и их не только дешево получить, но и они гораздо более устойчивы к ошибкам темы. Далее мы опишем применение одноязычных и многоязычных тематических моделей для улучшения моделей перевода [Eidelman et al., 2012, Hu et al., 2014b].

**Переводы из одноязычных тематических моделей** Мы можем обучить модель перевода, подсчитав частоту пар из данных выравнивания на уровне слов. Эйдельман и др. [2012] строит тематические модели перевода, переоценивая частоту пар слов на основе назначений мягкой темы/домена для документов. Поскольку предполагается, что переведенный документ имеет одни и те же темы на обоих языках, нам нужна только одноязычная тематическая модель, обученная на одном или другом языке. Распределение темы документа *p* ( *k* | *d* ) используется для сглаживания ожидаемого количества *c* ˆ k ( *f, e* ) числа

102

*Многоязычные данные и машинный перевод*

пара перевода слов по теме *k* ,

*c* ˆ k ( *f, e* ) =

P *d p* ( *k* | *d* ) *c d* ( *f, e* ) *,*

(8.6)

где *c d* (•) — количество вхождений пары слов в документе *d* . Лексическая вероятность, обусловленная темой *k* , является несглаженнойоценка вероятности этих ожидаемых значений

|  |  |
| --- | --- |
| *p w* ( *f* | *e* ; *k* ) = *c* ˆ k ( *f, e* ) */* P *f c* ˆ k ( *f, e* ) *,* | (8.7) |

из которого мы можем вычислить лексический вес этой пары фраз *p w* ( *f* ¯ | *e* ¯; *a, k* ) с учетом выравнивания слов *a* [Koehn et al., 2003]:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *н* |  |  |  |  | 1 |  |  |  |  |  | Х ) ∈ |  |  |
| *p w* ( *f* ¯ | *e* ¯; *a, k* ) = |  |  |  |  |  |  |  |  |  | *п ш* ( *ж я* | *е j* ; *к* ) | (8.8) |
| {| | | *Дж* | | | ( *я, дж* ) | е | *а* | }| | | ∀ ( |
| Д |  |  |  | *а* |  |
| *я* =1 |  |  |  |  |  |  |  |  |  |  | *я, дж* |  |

где *i* и *j* — позиции слов в целевой фразе *e* ¯ и исходной фразе *f* ¯ соответственно. Уравнения 8.7 и 8.8 эквивалентны уравнениям 8.4–8.5,но с добавлением присвоения мягких тем/доменов. Эйдельман и др. [2012] объединяют стандартные *f* ( *f* ¯ | *e* ¯) и *f* (¯ *e* | *f* ¯ ) с двумя направлениями адаптированных к теме вероятностей *p w* ( *f* ¯ | *e* ¯; *a, k* ) и *p w* (¯ *e* | *f* ¯ ; *a, k* ), что эквивалентно введению 2 *тыс* . новых таблиц перевода слов. Веса объектов оптимизируются с помощью алгоритма расслабления с добавлением маржи [Crammer et al., 2006, mira].

Для тестового документа *d* распределение тем документа *p* ( *k* | *d* ) выводится на основе тем, полученных из обучающих данных. Признак лексического веса пары фраз ( *f* ¯ *, e* ¯ ) равен

|  |  |  |
| --- | --- | --- |
| н | о |  |
| *f k* ( *f* ¯ | *e* ¯) = − log | *p w* ( *f* ¯ | *e* ¯; *k* ) · *p* ( *k* | *d* ) *,* | (8.9) |

сочетание зависимого от темы лексического веса и тематического распределения документа, из которого мы извлекаем фразу.

Эти адаптированные функции позволяют нам смещать переводы в соответствии с темами. Например, если в тестовом документе преобладает тема *k , признак f k* ( *f* ¯ | *e* ¯) будет большим, что может сместить декодер к переводу с малым значением стандартного признака *f* ( *f* ¯ | *e* ¯). Кроме того, сочетание адаптированных функций со стандартными функциями делает эту модель более гибкой. Для тестового документа с менее четкими темами распределение тем будет довольно равномерным. В

|  |  |
| --- | --- |
| *8.6. Тематические модели для перевода на уровне фраз* | 103 |

В этом случае тематические функции будут меньше влиять на результаты перевода, а стандартные функции будут доминировать в результатах перевода.

Хаслер и др. [2012] также применяют одноязычные тематические модели для адаптации домена к smt в той же структуре, что и Eidelman et al. [2012], за исключением того, что они применяют *модели Маркова со скрытой темой* [Gruber et al., 2007, htmm] вместо lda для изучения тем и извлечения различных признаков. В то время как lda предполагает, что каждое слово в документе генерируется независимо, htmm моделирует тему слова в документе как цепь Маркова, где всем словам в предложении назначается одна и та же тема. В результате тема для каждой пары фраз в выровненном предложении является последовательной и может использоваться непосредственно для тематических функций.

Су и др. [2012] используют htmm для включения тематической информации в вероятность фразы напрямую, а не через вероятность перевода слова. Учитывая данные для обучения двуязычному переводу без какой-либо конкретной предметной информации (называемые двуязычными данными вне предметной области), они включают тематическую информацию с исходного языка в перевод.

оценку вероятности, и разложить фразу вероятность *p* (¯ *e* | *f* ¯ ) как

|  |  |  |  |
| --- | --- | --- | --- |
| *p* (¯ *e* | *f* ¯ ) = | *p* (¯ *e, k вых* | *f* ¯ ) = | *p* (¯ *e* | *f* ¯ *, k out* ) · *p* ( *k out* | *f* ¯ ) | (8.10) |
| Икс |  | Икс |  |
| *k из* |  | *k из* |  |

где *p* (¯ *e* | *f* ¯ *, k out* ) — вероятность перевода с учетом исходной темы *k out* , а *p* ( *k out* | *f* ¯ ) обозначает вероятность фразы в теме *k out* .

Кроме того, Су и соавт. [2012] предполагают одноязычный корпус в том же домене, что и тестовое предложение (так называемые «одноязычные данные в домене»). Таким образом, они также применяют htmm для оценки внутренней темы *k in* и *p* ( *k in* | *f* ¯ ). Тем не менее, темы внутри домена *k in* и вне доменатемы *k out* могут находиться в разных местах, поэтому Su et al. [2012] вводят вероятность отображения темы *p* ( *k out* | *k in* ), чтобы сопоставить тему в предметной области с темой вне предметной области:

|  |  |  |
| --- | --- | --- |
| *p* ( *k вых* | *f* ¯ ) = | *p* ( *k вне* | *k в* ) · *p* ( *k в* | *f* ¯ ) | (8.11) |
| *к в* |  |  |
| Икс |  |  |
| В результате вероятность конечной фразы становится | |  |
| *p* (¯ *e* | *f* ¯ ) = ХХ *п* ( ¯ *е* | *ж* ¯ *, k вне* ) · *п* ( *k вне* | *k в* ) · *п* ( *k в* | *ж* ¯ ) *.* | | (8.12) |

*к выходу к входу*

104 *Многоязычные данные и машинный перевод*

Отношения темы и отображения темы между обучающими данными и тестовыми данными могут быть построены в автономном режиме, поэтому весь процесс не создает дополнительной нагрузки на систему перевода.

Одноязычные тематические модели могут добавлять контекстную информацию о выборе слова в модели перевода, но сами по себе не используют преимущества многоязычной информации. Затем мы обратимся к тематическим моделям, которые явно изучают многоязычные связи между словами.

**Многоязычная информация для адаптации предметной области** Использование двуязычных данных усложняет моделирование, но также может повысить качество тематической модели. Можно рассматривать тематические модели как инструменты для устранения неоднозначности значения слов на основе их контекста. Выравнивание по нескольким языкам является распространенным способом разрешения таких неоднозначностей. Например, «(l» в китайском документе может быть либо «коньком» в детской теме, либо «троянским вирусом» в теме технологии. короткий китайский документ, но эти термины однозначны на английском языке, более точно указывая на соответствующую тему.

Хотя многие подходы, описанные в этой главе, пытаются одновременно моделировать исходный и целевой языки для извлечения тем, некоторые преимущества многоязычных моделей могут быть достигнуты за счет согласования одноязычных моделей. Сяо и др. [2012] применяют тематические модели к исходным документам и целевым документам отдельно, чтобы изучить распределения тем документов *p* ( *k f* | *d f* ) и *p* ( *k e* | *d e* ), а затем оценить вероятности фраз-темы *p* (¯ *e, k f* | *f* ¯ ) и *p* (¯ *e, k e* | *f* ¯ ) из каждой модели. Кроме того, они вычисляют оценки сходства темы между распределением темы фразы и распределением темы документа как функции для декодирования для улучшения результатов smt.

Чтобы перевести новый документ *d f* , они сначала оценивают распределение тем документа *p* ( *k f* | *d f* ). Затем для заданной фразы *f* ¯ в исходном документе они ищут целевую фразу *e* ¯, которая максимизирует сходство между распределением темы исходного документа *p* ( *k f* | *d f* ) и распределением фразы-темы *p* (¯ *e, k f* | *f* ¯ ) согласно квадрату расстояния Хеллингера *H* 2 ( *p, q* ) = П *к* √ *п к* - √ *д к* 2 . Во-вторых, они вычисляют проекцию между двумя одноязычными тематическими моделями *p* ( *k f* | *k e* ) путем нормализации

|  |  |
| --- | --- |
| *8.7. Тематические модели для языкового моделирования на уровне предложений* | 105 |

количество совпадений в выровненных обучающих предложениях и использовать это отношение для расчета условного распределения целевых фраз и целевых тем *p* (¯ *e, k e* | *f* ¯ ).

Эта тематическая проекция напоминает тематическое картирование Su et al. [2012], но применяется между исходным и целевым языками. По сравнению с лексическими особенностями в Eidelman et al. [2012] и Hu et al. [2014b], Сяо и др. [2012] представили новую структуру для непосредственного применения тематической информации для измерения взаимосвязи между фразами и представили две функции сходства тем для декодирования. Эти два подхода можно комбинировать для дальнейшего улучшения smt.

**8.7**  **Тематические модели для языкового моделирования на уровне предложений**

Важнейшим компонентом систем машинного перевода является языковая модель, которая обеспечивает локальные ограничения и предпочтения, чтобы сделать перевод более связным. Языковая модель описывает вероятность появления слова *w при наличии* предыдущих слов контекста, которая также упоминается как история *h* (в главе 2.1 обсуждаются языковые модели для поиска информации). Они также помогают выбрать правильное или более подходящее слово во время статистического машинного перевода. Например, английские слова «дом» и «дом» часто являются синонимами, но перевод «я иду домой» лучше, чем «я иду домой».

Адаптация предметной области для языковых моделей [Bellegarda, 2004, Wood and Teh, 2009] использует дополнительные знания, чтобы скорректировать эту вероятность *p* ( *w* | *h* ), чтобы отразить изменение контекста, что является важным направлением для улучшения машинного перевода. Как отмечает Bellegarda [2004], «адаптивная языковая модель стремится поддерживать адекватное представление текущей предметной области в меняющихся условиях, включая возможные изменения в словарном запасе, синтаксисе, содержании и стиле».

Темы из тематических моделей могут быть одним из ресурсов для предоставления таких знаний для адаптации языковой модели. Например, китайская фраза « » переводится как «много вермишели» в домене еды , но означает «много фанатов» в домене развлечений . Такую неоднозначность можно уменьшить, используя знание темы/предметной области. Если тема развлечения извлекается на основе предыдущего контекста, эта китайская фраза

106 *Многоязычные данные и машинный перевод*

будет переведено как «много поклонников» без какой-либо двусмысленности. Далее мы представляем подробности о применении тематических моделей для адаптации языковой модели.

**Адаптация языковой модели из одноязычных тематических моделей на**  ранней стадии

Работа [Clarkson and Robinson, 1997, Seymore and Rosenfeld, 1997, Kneser and Peters, 1997, Iyer and Ostendorf, 1999] фокусируется на разделении обучающих данных на несколько тематических подмножеств и построении языковых моделей для каждого подмножества. Затем тематические языковые модели *p k* ( *w* | *h* ) линейно комбинируются с общей языковой моделью *p g* ( *w* | *h* ), построенной из всех обучающих данных в соответствии с уравнением 8.13. Веса *λ k* можно настроить в зависимости от тем тестовых документов.

|  |  |
| --- | --- |
| Икс | (8.13) |
| *п* адаптированный ( *ш* | *час* ) знак равно *λ k п k* ( *ш* | *час* ) + *λ г п г* ( *ш* | *час* ) |

*к*

Сеймор и др. [1998] дополнительно определяют наиболее подходящую тему для каждого слова в словаре и выбирают либо языковую модель, специфичную для темы, либо общую языковую модель. Интуиция подсказывает, что общая языковая модель обеспечивает наиболее надежную оценку для общих слов, а тематическая языковая модель более точно оценивает вероятность для более конкретных слов. В результате они разделили словарные слова на три группы: общее подмножество, подмножество по теме и подмножества не по теме. Они используют общую языковую модель для общего подмножества и не относящегося к теме подмножества и языковую модель, специфичную для темы, для тематического подмножества.

Все эти методы используют традиционную модель n-грамм, которая обуславливает конечную, ограниченную историю. Эти модели также предполагают, что каждый документ или история принадлежат только к одному тематическому кластеру. Чтобы решить эти проблемы, модели со смесями тем, такие как *латентный семантический анализ* [Deerwester et al., 1990, lsa] и его вероятностная интерпретациявероятностное латентное семантическое индексирование (plsi) [Hofmann, 1999a, plsi], изучение языковых моделей с большим охватом [Bellegarda, 1997, Coccaro and Ju-rafsky, 1998, Gildea and Hofmann, 1999]. Гильдеа и Хофманн [1999] разлагают языковую модель на темы,

Икс

*p* ( *w* | *h* ) = *p* ( *w* | *k* ) *p* ( *k* | *h* ) (8.14)

*к*

|  |  |
| --- | --- |
| *8.7. Тематические модели для языкового моделирования на уровне предложений* | 107 |

где темы изучаются из учебного корпуса путем оптимизации логарифмической вероятности,

|  |  |  |  |
| --- | --- | --- | --- |
| ХХ | | Икс | (8.15) |
| *л* ( *θ* ; *N* ) знак равно |  | *n* ( *w, d* ) log *p* ( *w* | *k* ) *p* ( *k* | *d* ) |
| *ж* | *г* | *к* |  |
|  |  |

где *d* — учебные документы, а *n* ( *w, d* ) — частота слова *w* в документе *d* . *p* ( *w* | *k* ) и *p* ( *k* | *d* ) изучаются с помощью алгоритма EM. Для тестовых документов они фиксируют *p* ( *w* | *k* ) для оценки *p* ( *k* | *h* ), а затем вычисляют *p* ( *w* | *h* ), используя уравнение 8.14.

Этот способ применения тематических моделей к языковым моделям аналогичен моделированию языка документа для поиска информации, как описано в главе 2.1. Однако, в отличие от информационного поиска, в процессе smt задействованы два разных языка, и они могут дополнять друг друга для изучения более точных тем. Далее мы обсудим многоязычные тематические модели для адаптации языковой модели.

**Адаптация языковой модели из многоязычных тематических моделей** Как мы объясняем в разделе 8.6, информация из разных языков может дополнять друг друга для извлечения лучших тем. Латентные семантические модели, такие как lsa, использовались в поиске многоязычной информации в течение многих лет [Carbonell et al., 1997]. Теперь мы опишем подходы к добавлению *многоязычной* информации в вероятностные тематические модели для адаптации языковой модели.

Там и др. [2007] вводят двуязычный латентный семантический анализ (blsa) для изучения тем как для исходного, так и для целевого языка и применяют изученные темы для адаптации языковой модели для smt. Подобно полиязычным тематическим моделям [Mimno et al., 2009], blsa переносит предполагаемые темы с исходного языка на параллельный целевой язык.

Более конкретно, Tam et al. [2007] предполагают, что выровненный исходный документ и целевой документ имеют одно и то же распределение тем документа. Сначала они изучают модель lsa на исходном языке, затем используют вектор темы документа из исходного документа в качестве вектора темы документа для выровненного целевого документа, а затем выводят вектор темы-слова на целевой стороне. Темы для целевого языка не изучаются итеративно, поэтому темы в параллельном корпусе могут быть изучены очень эффективно.

108 *Многоязычные данные и машинный перевод*

Чтобы применить темы для языковой адаптации, вычисляется маргинальное распределение *слов plsa (* w *)* для документа *d* ,

|  |  |
| --- | --- |
| *К* |  |
| Икс | (8.16) |
| *п лса* ( *ш* ) знак равно *п* ( *ш* | *k* ) *п* ( *k* | *d* ) |

*к* =1

Затем это предельное распределение слов интегрируется в модель целевого фонового языка путем минимизации расхождения KL между адаптированной языковой моделью и моделью фонового языка [Kneser et al., 1997]:

|  |  |  |
| --- | --- | --- |
| *п а* ( *ш* | *час* ) ∝ | *п лса* ( *ш* ) | |
|  | *п бг* ( *ш* ) |

! *β*

|  |  |
| --- | --- |
| · *р бг* ( *ш* | *ч* ) | (8.17) |

Ruiz и Federico [2011] применяют аналогичную идею для адаптации языковой модели. Вместо использования blsa Руис и Федерико [2011] объединяют выровненный исходный и целевой документы в один документ и обучают plsi. Обе идеи основаны на предположении, что выровненный исходный документ и целевой документ имеют одинаковое распределение тем документа. Окончательная адаптированная языковая модель сочетает в себе языковую модель на основе темы с моделью общего фонового языка, поэтому она более надежна в улучшении результатов smt.

Ю и др. [2013] представляют скрытую тему Марковская модель (htmm) для улучшения языковой модели в smt. Они строят модель темы на стороне источника и на стороне цели соответственно и изучают модель языка для конкретной темы на основе цели, оценивая максимальное правдоподобие. Чтобы сгладить резко распределенные вероятности, они возвращаются к другим распределениям:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *п* | *ж* | *я* | | *ш я* -1 |  | *, к* | | *λ* | *я* 1 | *п* | *MLE* ( | | *ж* | *я* | | | | *ш я* -1 | |  |  | *, к* | | (8.18) |
| ( |  | *я* - *п* +1 | |  | *д* ) = | | *ш я* - - *п* +1 |  |  |  | *я* - *п* +1 | | |  | *д* ) |
|  |  |  |  |  |  |  | +(1- | | *λ* | | *я* 1 |  |  | *п* | *MLE* ( | | | *ж* | *я* | | | *ш я* -1 | *, к* |
|  |  |  |  |  |  |  |  |  | *ш я* - - *п* +1 ) | | |  |  | *я* - *п* +2 | *д* ) (8.19) |
| где *λ* — параметр нормализации | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  | *N* 1+ ( *w i* −1 | | | | | |  |  | *, к э* ) | |  |  |  |  |
|  | *λ ш i* −1 | | | *,к е* | | знак равно |  | *я* 1 |  |  |  | *я* - *п* -1 | | | | |  |  | *я* | |  | (8.20) |
|  |  |  | *я* - *п* +1 |  |  |  | *N* 1+ ( *ш я* - - *п* -1 *, к е* ) + П *ш я c* ( *ш я* - *п* +1 *, к е* ) | | | | | | | | | | | | | | | |

где *N* 1+ ( *w i i* - - *n* 1 -1 *, k e* ) - количество слов, следующих за *w i i* - - *n* 1 -1 в теме *k e* , а *c* ( *w i i* - *n* +1 *, k e* ) — количество n-грамм *w i i* − *n* +1 в *k e* .

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *8.8. Изменение порядка с помощью тематических моделей* | | |  | 109 | |
|  |  |  |  |  |  |
|  | Тема | Тип |  | Пример |
|  |  |  |  | |  |
|  | Эконом | Источник · · · | Ôý **M** 3 *.* 8% · · · | |
|  |  | Цель · · · вниз 3 *.* 8% с мая · · · | | | |
|  | Спортивный | Источник | · · · | Ô 3 *.* 8% ··· |  |
|  |  | Цель | · · · | пять к одному · · · |  |

**Таблица 8.1:** Темы влияют на порядок слов: китайские слова, выделенные жирным шрифтом, расположены в разном порядке в разных темах. (Пример из Wang et al. [2014])

Во время декодирования, поскольку целевое предложение недоступно, они извлекают темы на исходной стороне и проецируют исходную тему на целевую сторону. Целевая вероятность:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *р* ( *е* ) = | Х *е* | Х *е* | Х *ф* | (8.21) |
| *п* ( *е* | *к е* ) *п* ( *к е* ) знак равно | *р* ( *е* | *к е* ) · | *п* ( *к е* | *к ж* ) *п* ( *к ж* ) |
|  | *к* | *к* | *к* |  |

где *p* ( *k e* | *k f* ) — вероятность прогнозирования темы, оцениваемая по совпадению назначения тем на стороне источника и на стороне цели.

**8.8**  **Изменение порядка с помощью тематических моделей**

В дополнение к моделям перевода и языковым моделям, третьим важным компонентом системы smt на основе фраз являются модели переупорядочивания, которые изучают, как порядок слов в исходных предложениях влияет на порядок слов в целевых предложениях и как сделать так, чтобы переводы в правильном порядке. Полезность тематических моделей при переупорядочении менее очевидна, чем их полезность для предметной адаптации моделей перевода и языковых моделей, но, тем не менее, она важна. Основное преимущество заключается в том, что порядок слов в разных доменах одного и того же языка может быть разным: Chen et al. [2013] обнаружили, что обучающие корпуса в разных областях значительно различаются по своим характеристикам переупорядочения для определенных пар фраз. Как показано в таблице 8.1 [Wang et al., 2014], в теме экономики китайское слово Ô находится слева от ; но в спортивной теме Ô справа от . В результате необходимо ввести предметные знания для моделирования такого отклонения порядка, и тематические модели обеспечивают хороший способ сделать это на основе данных.

110 *Многоязычные данные и машинный перевод*

Сюн и др. [2006] рассматривают проблему переупорядочения как классификацию с двумя метками: прямой и перевернутой между двумя последовательными блоками, и создают модель классификации с максимальной энтропией в качестве модели переупорядочения. Чен и др. [2013] вручную разделили обучающие данные на несколько доменов вместо использования автоматических методов, таких как тематические модели. Ван и др. [2014] интегрируют в модель переупорядочивания еще два типа тематических признаков в дополнение к признакам граничных слов [Xiong et al., 2006]. Во-первых, они выбирают тему с максимальной вероятностью в документе, чтобы быть *особенностью темы документа* для этого документа. Кроме того, они также используют темы слов содержания, которые находятся в крайнем левом и правом положениях исходных фраз, в качестве *признаков темы слова* для захвата шаблонов переупорядочения, зависящих от темы.

В процессе декодирования Xiong et al. [2006] сначала делают вывод о распределении тем тестовых документов, а затем применяют эту предложенную модель переупорядочения на основе тем в качестве одной подмодели к логлинейной модели максимальной энтропии, чтобы получить наилучший перевод:

|  |  |  |
| --- | --- | --- |
| Икс | *λ м ч м* ( *е , ж* ) ) |  |
| *e* лучший = argmax *e* ( *М* | (8.22) |

*м* =1

где *h m* ( *e, f* ) — подмодели или признаки всей лог-линейной модели, *λ m* — соответственно их веса, которые настраиваются на наборе для разработки.

Этот фреймворк очень гибкий и может кодировать любые тематические функции. Любые многоязычные тематические модели, которые мы обсуждали до сих пор, могут быть применены для извлечения лучших тем.

**8.9**  **Адаптация вне домена**

В дополнение к моделям перевода, языковым моделям и моделям переупорядочивания существуют и другие модули smt, такие как выравнивание слов, где также применяются тематические модели. Двуязычная тематическая смешанная модель Чжао и Син [2006, BiTAM] предполагает, что каждая пара документов представляет собой смесь тем, а темы для каждой пары предложений в этой паре документов выбираются из одного и того же распределения тем документа. Каждая тема также имеет таблицу перевода для конкретной темы. Следовательно

|  |  |
| --- | --- |
| *8.10. Резюме* | 111 |

Выравнивание слов на уровне предложения и переводы связаны скрытыми темами. BiTAM улавливает скрытую тематическую структуру и обобщает выравнивание слов и переводы с помощью тем, общих для пар предложений, таким образом улучшается качество выравнивания.

Кроме того, когерентность, которая связывает предложения текста в полноценную по смыслу структуру [Xiong and Zhang, 2013], является еще одним важным компонентом smt. Xiong and Zhang [2013] представили модель согласованности на основе тем для повышения качества перевода документов. Они узнают тему предложения для исходных документов, на основе которой они предсказывают целевую цепочку темы; затем они включают прогнозируемую целевую цепочку когерентности в процесс декодирования перевода документа.

**8.10**  **Резюме**

Тематические модели не ограничиваются одним языком, и разные языки могут быть связаны как на уровне документа, так и на уровне слов. Многоязычные тематические модели позволяют получить темы высокого качества, поскольку разные языки могут дополнять друг друга, чтобы уменьшить неоднозначность темы. Многие различные подходы применяют многоязычные тематические модели для улучшения различных частей конвейера статистического машинного перевода. Обладая таким знанием темы, можно лучше уловить вариации разных языков, чтобы сделать переводы более естественными и связными.

**9**

**Построение тематической модели**

В предыдущих главах основное внимание уделялось *существующим* моделям. До сих пор мы описывали модели, созданные исследователями для фиксации конкретных нюансов документов или процессов создания документов, существующих в мире. В этой главе основное внимание уделяется тому, как исследователь может создать, внедрить и проверить новую модель.

Прежде чем углубляться в детали построения новых моделей, мы призываем пользователей подумать, можно ли ответить на вопросы с помощью постфактум анализа простой тематической модели с дополнительной информацией. Например, простая динамическая тематическая модель может быть построена из стандартной модели lda путем разделения корпуса на разделы и оценки вероятностного распределения слов по каждой теме в каждом разделе. Построение и проверка пользовательских тематических моделей — это мощный инструмент, но он требует значительных инвестиций в кодирование и отладку, а также может быть не в состоянии воспользоваться вычислительными оптимизациями, доступными для более простых моделей. Апостериорные прогностические проверки [Mimno and Blei, 2011] предоставляют хорошие средства для определения того, демонстрируют ли темы lda или слова внутри тем паттерны, которые присутствуют, но явно не моделируются.

Невозможно охватить все детали исследования в тематических моделях.

112

|  |  |
| --- | --- |
| *9.1. Разработка модели* | 113 |

или машинного обучения в целом, но в этой главе представлены некоторые распространенные методы создания новых моделей. Мы сосредоточимся на рабочем примере, создав новую модель для предсказания идеологии политического оратора.

**9.1**  **Разработка модели**

Первым шагом в создании новой модели является определение того, что является важным. Например, в предыдущих главах основное внимание уделялось измерению инноваций, межъязыковых связей и настроений. Это концепции высокого уровня, которые мы хотим обнаружить из текста. Мы считаем, что эти свойства существуют в мире, но мы хотим, чтобы переменные наших моделей представляли эти понятия.

В тематической модели включение новой концепции в модель обычно включает добавление в модель новой случайной переменной. Здесь на первый план выходят интуиция и знание предметной области. Поскольку генеративный процесс пытается смоделировать реальный мир, новая модель должна уравновешивать несколько компонентов, которые часто находятся в противоречии друг с другом: точность, производительность, управляемость и интерпретируемость.

**Верность** Хорошая модель должна отражать мир. Один из способов, с помощью которого мы можем смоделировать отношения между словами и несловесной информацией, состоит в том, чтобы определить паттерны обусловливания. Если политологи считают, что идеология политика — это свойство, которое меняет то, как он говорит, то модель должна обусловливать выбор темы идеологией говорящего. Мимно и МакКаллум [2008] описывают эту формулировку как «восходящую» модель (подробнее см. главу 7). Если они считают, что успех на выборах является результатом политических выступлений, то модель должна обусловливать успех темами (модель «нижнего течения»).

Моделирование реальности — хорошая идея. Но точно так же, как для построения масштабной модели здания требуются компромиссные материалы и уровень детализации, для построения статистической модели иногда требуются нереальные предположения. Если генеративная модель точно соответствует процессу, производящему данные, мы можем доказать, что она сходится к правильному ответу [Neal, 1993]. Но люди и текст — это не Дирихле и не дискретные распределения; все модели будут

114 *Построение тематической модели*

приближение.

В дополнение к определению того, *где* моделировать функцию в генеративной модели, вы должны решить *, как* ее моделировать. Является ли это непрерывным значением, двоичным значением, членом дискретного набора или чем-то еще? Часто у нас есть несколько предопределенных представлений о том, как представлять интересующую величину. Настроение иногда представляется как постоянное положительное или отрицательное значение, в то время как корпуса обзоров включают дискретные положительные звездные рейтинги. Политическая идеология может быть представлена членством в одной из фиксированного числа партий, но политологи чаще придают идеологии политика непрерывное значение.

**Производительность** Более точное представление о том, как устроен мир, не всегда означает, что модели будут более полезными. Иногда нам приходится обменивать точность на производительность. Недавняя работа, по-видимому, согласуется с тем, что модели *ниже по течению* работают лучше, хотя они менее реалистичны [Nguyen et al., 2013]. 1

Почти невозможно *априори знать, будет* ли модель хорошо работать для конкретной задачи, учитывая только эту модель. Знание того, что будет работать лучше всего, часто является процессом проб и ошибок. Однако часто можно провести параллели с аналогичными моделями. Восходящие модели позволяют переменным метаданных лучше предсказывать, какие темы будут встречаться в контексте, но они не обязательно побуждают модель изучать другие темы, чем модель без метаданных. Нисходящие модели должны согласовывать темы, чтобы наилучшим образом предсказывать переменную метаданных, и поэтому имеют тенденцию находить разные (но не обязательно лучшие) темы. Таким образом, контролируемые тематические модели хорошо работают для прогнозирования настроений как нисходящей переменной, поэтому было бы разумно предположить, что нисходящая модель хорошо работает и для идеологии.

**Управляемость** Теперь, когда мы знаем, какую переменную мы хотим смоделировать, некоторые подходы к моделированию этой переменной могут быть проще, чем другие. Политическая идеология часто рассматривается как *спектр* , а не ярлык: некоторые

1 Можно возразить, что для *слушателя* , который должен интерпретировать речь, более реалистично присвоить идеологию после того, как она была услышана, но это больше не согласуется с тем, как был *создан текст* .

|  |  |
| --- | --- |
| *9.1. Разработка модели* | 115 |

политики более центристские, чем другие, даже если они могут носить один и тот же партийный ярлык.

Однако дискретные метки привлекательны с точки зрения моделирования. Дирихле-дискретные распределения (глава 1) легко комбинируются. Учитывая, что базовые тематические модели строятся из Дирихле и дискретных распределений, проще добавить дополнительную дискретность Дирихле, чем добавлять в модель непрерывную случайную величину. Например, модель «тема-аспект» [Paul and Girju, 2010] гораздо проще реализовать, чем тематические модели с учителем [Blei and McAuliffe, 2007].

Распределения Дирихле и дискретные распределения хорошо сочетаются друг с другом, потому что они сопряжены, а распределения Гаусса добавляют дополнительные трудности. Однако распределения Гаусса более удобны, чем другие распределения. Например, считается, что сферические распределения [Batmanghelich et al., 2016] лучше моделируют непрерывные вложения слов, чем распределения Гаусса, но за счет менее удобной модели.

Еще хуже комбинаторные распределения вероятностей. Например, в главе 8 мы обсуждали, как изучать сопоставления между языками. Boyd-Graber and Blei [2009] используют комбинаторное распределение для изучения отображения одного языка на другой. Это очень сложно, но не моделирует языки так же хорошо, как гораздо более простые подходы [Mimno et al., 2009].

Наш список все более сложных дистрибутивов не следует воспринимать как предостережение против их использования: иногда необходимы сложные модели. Однако не следует выбирать сложную модель только потому, что она сложна (заманчивое искушение для молодых аспирантов, желающих продемонстрировать свои навыки машинного обучения). Лучше всего попробовать самую простую модель, которая может работать; даже в случае неудачи эта модель может служить полезной основой.

В отличие от многих других измерений при построении модели сложность часто ищется и фетишизируется без улучшения других измерений (в этом часто играют роль порочные стимулы из публикаций). Таким образом, остерегайтесь излишнего усложнения модели!

**Интерпретируемость** Как мы описали в главе 3.4, интерпретируемость — это мера того, насколько легко человеку понять результаты

116 *Построение тематической модели*

модель. Часто об интерпретируемости модели забывают. То есть, если об этом вообще думают; многие статьи пренебрегают проверкой изученных параметров модели, вместо этого сосредотачиваясь на количественной оценке производительности.

Несмотря на то, что это всегда проблема, с ростом популярности глубокого обучения (более подробно это обсуждается в главе 10) это вызывает особую тревогу. Глубокое обучение имеет репутацию непостижимых параметров, но современной производительности. Одной из сильных сторон вероятностных моделей является их интерпретируемость и обоснованность порождающих процессов. Таким образом, исследователи, которые выбирают вероятностные модели, такие как тематические модели, не должны игнорировать интерпретируемость своих моделей.

Хотя в главе 3.4 обсуждаются строгие оценки интерпретируемости модели, даже простой просмотр параметров модели исследователем может показать, «имеет ли смысл» изученная модель или нет. Осмотр модели может помочь подтвердить, смог ли дизайн модели («точность» выше) отразить интуицию моделиста.

Однако часто приходится идти на компромиссы с интерпретируемостью (как при выборе модели глубокого обучения). Чанг и Блей [Chang and Blei, 2009] обнаружили, что коррелированная тематическая модель [Blei and Lafferty, 2007] имеет значительно более высокую скрытую вероятность за счет интерпретируемости.

**9.2**  **Реализация модели**

Итак, у вас в руках новая модель. Обычно это описывается как генеративный процесс (глава 1.4): последовательность вероятностных шагов, которая рассказывает историю о том, как появились ваши данные.

Реализация модели требует записи этой модели в форме, которую может понять компьютер, а затем с помощью *вероятностного вывода* работает в обратном направлении от данных, чтобы обнаружить конфигурацию скрытых переменных (темы, другие свойства данных, которые вы добавили как часть модели). процесс построения), которые лучше всего описывают ваши данные.

**9.2.1**  **Автоматические подходы**

Автоматические подходы к выводу привлекательны: запишите свою модель и закройте ее. Тем не менее, автоматические подходы имеют несколько

|  |  |
| --- | --- |
| *9.2. Реализация модели* | 117 |

недостатки. Они часто ограничены определенными платформами, медленнее, чем вывод «вручную», и ограничивают типы моделей, которые вы можете исследовать.

Использование инструмента (каким бы прекрасным он ни был), разработанного кем-то другим, означает, что вы должны принять его предположения. Это может заставить вас использовать язык программирования, с которым вы не знакомы, это может заставить вас форматировать ваши данные странным образом, или это может ограничить вас операционными системами, которые вы не используете (или не можете себе позволить).

**Stan** Stan [Stan Development Team, 2014] — это среда логического вывода, которая лучше всего работает с языком программирования R.

**Infer.Net** Infer.net [Minka et al., 2014] — созданная Microsoft среда с закрытым исходным кодом, предназначенная для сопряженных моделей, но может использоваться только в операционной системе Windows.

**Автоматическая дифференциация** Появление глубокого обучения создало множество платформ автоматической дифференциации. Их часто можно использовать для произвольных целевых функций, включая вариационный вывод. Если вы уже знакомы с Torch [Collobert et al., 2011] или Theano [Theano Development Team, 2016], эти инструменты относительно легко использовать для определения вариационных целей для произвольных вероятностных моделей.

**9.2.2**  **Вариационный вывод**

Хотя в главе 1.4 мы сосредоточились на выборке Гиббса, другим основным классом алгоритмов вывода является вариационный вывод. По сравнению с выборкой Гиббса вариационный вывод часто считается немного более сложным как для получения, так и для реализации.

Вариационный вывод напоминает алгоритмы максимизации ожидания [Liang and Klein, 2007]. Алгоритмы максимизации ожидания находят настройку локальных скрытых переменных *z* (которые представляют конкретные наблюдения) и глобальных параметров *θ* (которые представляют свойства модели), которые максимизируют вероятность данных *p* ( *x* | *z, θ* ). Во-первых, мы начинаем с некоторого предположения о том, какими могут быть скрытые переменные *z* 0 . Затем мы обновляем

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 118 |  |  |  | *Построение тематической модели* |
| параметры, которые должны быть |  |  |  |  |
| аргумент | макс *р* ( *х* | | | *z, θ* ) *,* | (9.1) |
| *θ* |  |  |

параметры, максимизирующие вероятность. Затем, учитывая эти параметры, вычислите ожидаемое значение скрытых переменных *z* , чтобы вычислить следующую итерацию скрытых переменных *z* .

Максимизация ожидания является полезным инструментом для вывода, когда распределение модели *p* достаточно просто для непосредственного решения уравнения 9.1. Однако для многих моделей (в том числе тематических) это неосуществимо. Вариационный вывод решает эту трудность путем поиска оптимальных *распределений q* ( *z* ) и *q* ( *θ* ), а не оптимальных значений *z* и *θ* . этоэтот поиск оптимальных функций (т. е. функций распределения вероятностей), а не оптимальных значений, дал название этому методу: «вариационное исчисление» — это ветвь исчисления, касающаяся оптимальных функций.

Вариационное распределение — это функция, которая присваивает оценку вероятности каждой настройке скрытых переменных модели. Это те же самые переменные, которые появятся в апостериорном распределении исходной модели по скрытым переменным с учетом наблюдаемых переменных и определяемых пользователем гиперпараметров. Хотя это распределение по тем же переменным, обычно оно проще. В исходном апостериорном распределении существует множество зависимостей между переменными: темы для одного документа ограничивают темы для других документов. В обычных подходах к тематическим моделям [Blei et al., 2003] распределение полностью факторизуется, чтобы обойти эти ограничения. Например, истинное распределение по скрытым переменным равно

|  |  |  |
| --- | --- | --- |
| Д | Д | Д |
| *p* ( *z, θ* ) = *p* ( *φk* | *β )* | *п* ( *θ d* | *α* ***ты*** ) | *р* ( *г д, п* | *θ d* ) *п* ( *ш d, n* | *β z d,n* ) (9.2) |
| *к* | *г* | *н* |
|  |

где ***u*** — *K -* мерное равномерное распределение, а вариационное распределение

Г Д

*q* ( *z, θ, β* ) =  *q* ( *β k* | *λ k* )  *q* ( *θ d* | *γ d* ) *q* ( *z d,n* | *φ d, n* ) *,*  (9.3)

* *г*

где *λ* и *γ* — параметры Дирихле, описывающие скрытые переменные модели. Хотя функционально то же самое, что и истинное распределение *p* , поскольку параметры больше не связаны, они называются вариационными параметрами.

|  |  |
| --- | --- |
| *9.2. Реализация модели* | 119 |

Вместо максимизации правдоподобия данных вариационный вывод минимизирует расстояние между вариационным распределением *q* (произведение независимых распределений по каждой из скрытых переменных) и апостериорным распределением исходной модели *p* . Здесь расстояние — это расхождение Кульбака-Лейблера *между* распределениями *p* и *q* . Минимизация этого расхожденияэквивалентно оптимизации нижней границы вероятности данных,

|  |  |
| --- | --- |
| *`* ≡ E *q* [log ( *p* ( *w* | *z, θ* ) *p* ( *z, θ* | *α, β* ))] − E *q* [log *q* ( *z, θ* )] | (9.4) |

Существует несколько вариантов такой оптимизации: прямая оптимизация, стохастический градиент [Hoffman et al., 2010] или координатный подъем [Blei et al., 2003].

**Получение целевой функции** После того, как вы выбрали вариационное распределение, вам необходимо вывести полную форму цели и вычислить обновления для каждого вариационного параметра. Это включает в себя получение производной уравнения 9.4 по этому вариационному параметру, установку ее равной нулю, а затем решение для вариационного параметра.

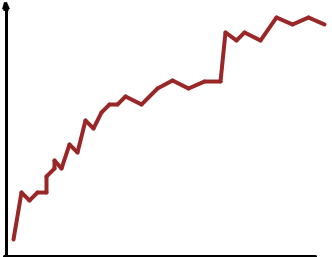
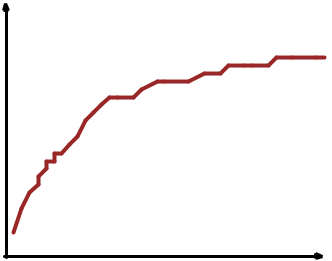
**Выбор вариационного распределения** С одной стороны, вы хотите выбрать вариационное распределение, близкое к истинному распределению скрытых переменных. На самом деле, если вы выберете *q* так, чтобы оно было равно *p ,* вариационный вывод сводится к максимизации ожидания.

Однако за выбор более точного вариационного распределения часто приходится платить: более сложные вычисления или более сложный (или даже невозможный) вывод. Возможно, вам не удастся решить уравнения для обновления отдельных вариационных параметров, или расчеты могут оказаться более сложными. Чем больше зависимостей в вариативном распределении, тем больше терминов необходимо учитывать. Когда во многих документах есть скрытые переменные для каждого токена, количество зависимостей может резко увеличиться.

Часто хорошим выбором является полностью факторизованное вариационное распределение. После реализации полностью факторизованного вариационного распределения исследователь должен тщательно отслеживать целевую функцию (уравнение 9.4). В идеале она будет быстро увеличиваться и достигать устойчивого (локального) оптимума (рис. 9.1,

120 *Построение тематической модели*

|  |  |  |
| --- | --- | --- |
| ` | Эффективный | ` |
|  |
| вариационный |  |  |
|  |  | Итерация |
|  | p(w | z, φ, θ) |  |

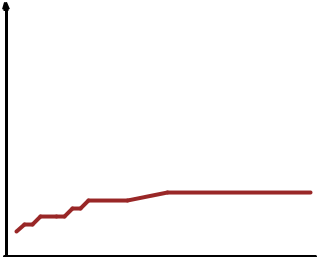


Гиббс

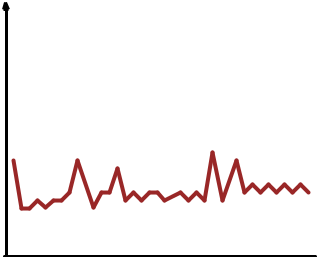
Итерация



Неэффективный



Итерация



p(w | z, φ, θ)

Итерация



**Рисунок 9.1:** Мониторинг целевой функции является важным компонентом диагностики правильности работы логического вывода. Правильный вариационный вывод должен монотонно увеличиваться, в то время как выборка Гиббса может немного уменьшаться. Однако в обоих случаях целевая функция должна резко возрастать со временем. Если цель плоская, это может означать, что вам может потребоваться пересмотреть выбор дизайна в вашем алгоритме вывода. Шкалы для вариационного вывода и выборки Гиббса несопоставимы.

|  |  |
| --- | --- |
| *9.2. Реализация модели* | 121 |

оставил). Однако, если это не так (рис. 9.1, справа), возможно, существуют связанные переменные, которые плохо обслуживаются вариационным распределением.

Связанные переменные должны изменяться вместе, а полностью факторизованное распределение заставляет их изменяться последовательно. Например, предположим, что *x* = 0 и *y* = 0 и что они сильно коррелированы, но более высокая вероятность установлена для *x* = 1 и *y* = 1. Оптимизация подъема координат с полностью факторизованным вариационным распределением потребует изменить *x* или *индивидуально* , когда на самом деле они должны двигаться вместе, чтобы поддерживать высокую производительность.

Чтобы полностью смоделировать взаимодействие между двумя переменными, мы можем совместно смоделировать *x* и *y* в вариационном распределении. Вместо независимых распределений *q* ( *x* ) *q* ( *y* ) вариационное распределение становится *q* ( *x , y* ) (в соответствии с формулировкой этих переменных).

**9.2.3**  **Выборка Гиббса**

Выборка Гиббса (как обсуждалось на высоком уровне в главе 1) находит скрытые переменные путем случайной выборки присваивания каждой случайной переменной, обусловленной всеми другими случайными переменными.

Таким образом, после создания вашей модели вам необходимо вычислить условное распределение каждой случайной величины, зависящее от всех остальных. Например, условное распределение переменной назначения темы слова *n z d,n* (выделено жирным шрифтом) равно

*p* ( ***z d,n*** | *z d,* 1 *. . . zd ,n* −1 *, zd ,n* +1 *, . . . z d,N d , θ, φ* ) =

*р* ( *г д,* 1 *. . . zd ,n* −1 *,* ***z d,n*** *, z d,n* +1 *, . . . г д, Н д , θ, ф* ) *,* (9.5)

*р* ( *г д,* 1 *. . . zd ,n* −1 *, zd ,n* +1 *, . . . zd , Nd , θ, φ* )

которое упрощается до уравнения 1.4.

Получение условных распределений часто проще, чем расширение вариационных ожиданий. Хардисти и др. [2010] предоставляют пошаговое руководство по получению выборки Гиббса для вероятностных моделей.

**Маргинализация и совместные переменные** Точно так же, как есть некоторое искусство в решении, какие переменные объединять в совместное распределение для вариационного

122 *Построение тематической модели*

логического вывода, в выборке Гиббса вам нужно решить, какие переменные выбрать для совместной выборки, а какие исключить.

Совместная выборка переменных решает ту же проблему, что и объединение переменных в одно и то же вариационное распределение, описанное выше. Вместо выборки *x* , обусловленной всеми другими переменными, а затем выборки *y* , *x* и *y* взятые *одновременно* из распределения, которое обусловливает *x* и *y* всеми остальными переменными.

**9.3**  **Отладка и проверка**

Теперь, когда вы реализовали вывод для своей тематической модели, как узнать, работает ли она так, как задумано?

**9.3.1**  **Синтетические данные**

Поскольку большинство тематических моделей имеют генеративную историю, мы можем *генерировать* данные. Это означает запуск вероятностной истории модели для генерации данных. Например, для lda выберите распределение тем для каждого документа и распределение типов для каждой темы из распределения Дирихле. Такой набор данных часто называют *синтетическим* набором данных.

Если вы создадите набор данных таким образом, скрытые переменные перестанут быть скрытыми. Вы *точно знаете,* что они собой представляют. Если вы выполняете вывод по этим данным, вы сможете воссоздать сгенерированные вами незамаскированные скрытые переменные. Если ваш алгоритм вывода не подходит близко, вероятно, проблема в вашей процедуре вывода.

Однако тематические модели могут быть сложными. Тема 7 в ваших синтетических данных может соответствовать теме 4, обнаруженной вами в результате логического вывода. Это не проблема, поскольку идентификаторы тем произвольны; таким образом, вам придется жадно сопоставлять темы или измерять какую-либо другую статистику, чтобы сравнить ваши предполагаемые переменные с истинными синтетическими данными.

Синтетические данные наиболее ценны, когда они близки к распределению реальных данных. Задачи-игрушки могут быть хорошей проверкой работоспособности, но в лучшем случае дают нечеткую верхнюю границу нашей способности моделировать данные, которые нам интересны. Таким образом, использование *полусинтетических* данных может быть хорошим компромиссом. В этом параметре вы обучаете модель на реальных данных, а затем создаете новые документы на основе этой модели. Получившийся полусинтетический корпус имеет

|  |  |
| --- | --- |
| *9.3. Отладка и проверка* | 123 |

многие из правильных свойств естественных документов, таких как размер словарного запаса и разреженность, но гарантированно соответствуют предлагаемой модели.

**9.3.2**  **Обновления**

Как для вариационного вывода, так и для выборки Гиббса наиболее важным шагом является обновление распределения для отдельных случайных величин. Таким образом, вам нужно будет предпринять как можно больше шагов, чтобы обеспечить их точность.

Для вариационного вывода каждое обновление вариационных параметров увеличивает целевую функцию (уравнение 9.4). Таким образом, вы можете проверять после каждого обновления вариационного распределения переменной, увеличилась ли эта цель или нет. Хотя не каждое обновление точно увеличивает цель (из-за ошибок числовой точности или инициализации), ни одно обновление не должно резко *снижать* цель. Если это так, у вас, вероятно, есть ошибка.

Выборка Гиббса является стохастической, поэтому ее сложнее отлаживать. Но как для вариационного вывода, так и для выборки Гиббса вы должны написать модульные тесты (разработанные ручкой и бумагой), чтобы убедиться, что ваши обновления дают правильный ответ при одних и тех же входных данных. Случайность выборки Гиббса может быть обработана в среде модульного тестирования с помощью заглушек, которые заменяют генератор случайных чисел (эти заглушки также могут быть полезны, когда вы позже захотите сгенерировать результаты с другими случайными начальными значениями, как обсуждается ниже).

**9.3.3**  **Базовые показатели и показатели**

Часто вы захотите сравнить свою модель с другой моделью, которая имеет аналогичную структуру или приложение. Это сравнение может помочь вам во время разработки.

Например, вы можете каким-то образом расширить контролируемый lda (глава 7). К счастью, slda — это четко определенная модель, которая доказала свою эффективность на широкодоступных наборах данных.

Таким образом, разумной стратегией разработки было бы создание ряда моделей, которые ведут вас от slda к вашей окончательной модели (например, добавляя одну скрытую переменную за раз), каждый раз сравнивая, насколько хорошо ваша модель работает с slda. Эти сравнения не только помогают развитию;

124 *Построение тематической модели*

они также помогут задокументировать, насколько важны каждое из изменений в модели.

Сравнение тематических моделей может быть затруднено, поскольку реализации и алгоритмы могут сильно различаться. Часто неясно, например, связано ли наблюдаемое различие с тем, что одна модель лучше другой, или со сравнением выборки Гиббса с вариационным выводом. Отличный способ протестировать новый алгоритм для сложной модели — найти способы реализации более простых моделей в том же коде. Например, вы можете эмулировать модель lda, используя модель Author-Topic [Rosen-Zvi et al., 2004], назначив каждому документу своего собственного «автора».

Стандартной метрикой для тематических моделей является их недоумение или ожидаемая вероятность. Это значение представляет собой вероятность задержек данных с учетом настроек модели. Уоллах и др. [2009b] подробно описывает метод оценки недоумения в тематических моделях с использованием алгоритма выборки Гиббса (который выходит за рамки данного обзора).

Однако недоумение часто не то, что вас волнует при применении тематических моделей. Например, если вы разрабатываете вариант slda, вы, вероятно, заботитесь о точности предсказания. Таким образом, по мере того, как вы медленно создаете свой вариант slda, вы также должны сообщать о производительности прогнозирования.

**9.4**  **Общение с вашей моделью**

После построения модели, реализации логического вывода и применения его к данным следующий захватывающий шаг — рассказать миру. Хотя очевидно, что необходимо мотивировать новую модель и описывать технические детали, одним из компонентов сообщения модели, который часто упускается из виду, является взаимодействие между этими двумя аспектами описания модели.

Приложение не должно просто стоять рядом с целью моделирования; он должен быть тесно интегрирован в вероятностную историю. Вместо того, чтобы перечислять шаги генеративного процесса, генеративный процесс должен быть описан вызывающими воспоминания именами переменных. Например, если ваша модель пытается отразить политическую поляризацию, вы можете назвать дискретную переменную «поляризация *π* », чтобы было ясно, где в модели будет моделироваться этот аспект.

|  |  |
| --- | --- |
| *9.5. Резюме* | 125 |

Однако имя переменной не делает ее таковой. Вам также нужно будет предоставить качественные доказательства, которые убедят скептически настроенного читателя в том, что ваша модель делает то, что вы обещали. Это возможно, предоставляя обзоры ваших данных на микро- или макроуровне.

На макроуровне вы можете показать, что ваша модель делает разумные вещи, показав распределение слов по заданным темам (или любой другой аналогичный компонент). Тем не менее, вам следует избегать искушения выбирать темы вишневого цвета. Лучше выбирать темы случайным образом или — если вы выбираете вишневые — также выбирать «плохие» темы, которые показывают режимы отказа вашей модели.

Хотя сигналы макроуровня могут показать, что модель находит хорошие сводки, вы также должны показать отдельные документы и то, как они взаимодействуют с данными. Например, если ваша модель пытается показать политическую поляризацию, вы можете показать поляризованные (или нет) документы и показать, как скрытые переменные в ваших моделях правильно отражают эти аспекты документов (или нет; как и в случае с темами, также отображайте виды ошибок). ).

Кроме того, у вас также будут количественные показатели для вашей задачи: точность предсказания, точность на ранге *K* или качество перевода (глава 8). При сообщении количественных результатов помните о вероятностных основах тематических моделей: вы не изучаете один ответ. Независимо от вывода, вы изучаете *распределение* по скрытым переменным с учетом набора данных.

Таким образом, передать присущую неопределенность в умозаключении. Вывод следует запускать несколько раз с разными случайными начальными значениями (выборка Гиббса) или случайными инициализациями (вариационный вывод). Количественные результаты с планками погрешностей позволяют проводить достоверные сравнения с другими моделями: вы не только получаете более высокий балл, но и показываете, что ваш более высокий балл не является результатом случайности.

**9.5**  **Резюме**

В этой главе обсуждается процесс создания новой тематической модели с нуля и то, как сообщить об этом процессе миру. Несмотря на то, что они внешне отличаются от наших глав, посвященных приложениям, большинство статей были созданы в процессе построения моделей, выводов,

126 *Построение тематической модели*

и оценка. Таким образом, эта глава помогает понять процесс построения моделей, обсуждаемых в этих главах.

В следующей главе мы рассмотрим будущее тематических моделей и то, как тематические модели могут вписаться в более широкую повестку дня исследований в области компьютерных наук и информатики.

**10**

**Вывод**

Хотя мы пытались охватить различные применения тематических моделей, чтобы помочь людям ориентироваться в больших наборах текстовых данных, ни одно конечное исследование не могло бы перечислить все приложения тематических моделей в тексте, которые были применены к маркировке частей речи [Toutanova and Johnson]. , 2008], индукция смысла слов [Brody and Lapata, 2009] и устранение неоднозначности сущностей [Kataria et al., 2011]. Само собой разумеется, что мы также пропустили многие другие приложения вне текста, такие как биология [Pritchard et al., 2000], понимание исходного кода [Maskeri et al., 2008], анализ музыки [Hu and Saul, 2009] и многое другое.

**10.1**  **Как справиться с информационной перегрузкой**

Проблема в тематическом моделировании заключается в том, как сделать логические выводы достаточно эффективными, чтобы масштабироваться до больших наборов данных и предоставлять интерактивные возможности с малой задержкой, чтобы помочь пользователю в цикле. Есть три общие стратегии для более быстрой обработки документов.

Во-первых, за счет уменьшения среднего количества просмотров документа компьютером для изучения тематической модели; т. е. для повышения *пропускной способности* . Только онлайн-алгоритмы [Hoffman et al., 2010]

127

128 *Заключение*

просмотрите документ один раз, обновите темы, а затем перейдите к следующему документу. Часто это намного быстрее, чем пакетные подходы, требующие многократных проходов по одному и тому же набору документов. Другой вариант — *распределить* вычисления по множеству машин [Zhai et al., 2012].

Дополнительный подход заключается в сокращении времени, затрачиваемого компьютером на любой конкретный документ: повышение *эффективности* . Это возможно за счет сокращения времени, необходимого для выборки назначений документов [Yao et al., 2009, Li et al., 2014] или вычисления вариационных параметров [Mimno et al., 2012].

Последним подходом к повышению эффективности вероятностных алгоритмов для тематических моделей является полное переосмысление процесса логического вывода. Новые подходы рассматривают вывод тематической модели как факторизацию матрицы совпадений [Arora et al., 2013] или как спектральную декомпозицию [Anandkumar et al., 2012]. Эти подходы часто намного быстрее, чем традиционные подходы, поскольку они используют типы слов, а не документы, в качестве центральной единицы вычислений.

**10.2**  **Более глубокие представления**

Частично преимущество тематических моделей заключается в том, что распределение тем документа ( *θ* ) служит низкоразмерным представлением того, что означает документ. Этот числовой вектор полезен для поиска похожих документов (глава 2), отображения документов пользователю (глава 3) или соединения документов на разных языках (глава 8).

Все чаще векторные распределенные представления становятся полезными «на всем пути». Представление слов и фраз на основе векторов может улучшить предсказание следующего слова [Bengio et al., 2003], анализ настроений [Socher et al., 2012] и перевод [Devlin et al., 2014]. И это касается не только текста — репрезентативное обучение охватило речь, зрение и машинное обучение в целом.

Влияние репрезентативного обучения на тематическое моделирование остается неясным, поскольку мы готовимся к печати в 2017 году. Мы видим несколько способов, которыми репрезентативное обучение и тематическое моделирование могут принести пользу друг другу в будущем.

|  |  |
| --- | --- |
| *10.3. Автоматический анализ текста для людей* | 129 |

**Оценка** Методы оценки из тематических моделей (глава 3.4) нашли свое применение в репрезентативном обучении [Schnabel et al., 2015, Iyyer et al., 2016], что позволяет предположить, что некоторые уроки, полученные при обеспечении интерпретируемости тематических моделей, также могут быть применены. в репрезентативном обучении. Это успокаивает некоторых критиков репрезентативного обучения, которые утверждают, что результаты часто не поддаются интерпретации или вводят в заблуждение [Szegedy et al., 2014].

**синтеза** также смешивается с более выразительными моделями скрытого представления [Ranganath et al., 2015]. Тематические модели могут помочь репрезентативному обучению решить некоторые трудности, связанные с обобщением больших сегментов текста. Абзацы и предложения сложно моделировать как единый вектор, а методы, более сложные, чем простое усреднение, не стоят хлопот [Iyyer et al., 2015].

**Параллельная эволюция** Другой возможный путь менее переплетен: тематические модели и репрезентативное обучение с глубоким обучением решают разные проблемы и не являются прямыми конкурентами. Тематические модели предлагают преимущества скорости и интерпретируемости, в то время как обучение на основе представлений может лучше справляться с задачами, основанными на прогнозировании. Тематические модели никогда не были идеальными, например, для определения признаков классификации текста: почти всегда лучше использовать признаки подсчета слов. Если интерпретируемость и узнаваемость не являются фундаментальной целью вашего анализа, возможно, вам лучше использовать что-то другое. Однако в тех случаях, когда интерпретируемость и распознавание являются основными и конечными целями анализа, методы глубокого обучения мало что дают, потому что само их преимущество — большая репрезентативная сложность — также является их слабостью. Оба подхода должны быть инструментами, которые есть у многих текстовых майнеров, с особыми условиями для использования каждого из них.

**10.3**  **Автоматический анализ текста для людей**

Однако, на наш взгляд, главная исследовательская задача тематических моделей состоит не в том, чтобы усложнить эти модели и их выводы, а в том, чтобы сделать их более доступными. Как мы уже говорили, тематические модели

130 *Заключение*

может помочь ученым и обычным людям ориентироваться в больших коллекциях текстов. Однако использование тематических моделей по-прежнему требует обширных данных и знаний.

пьютерные навыки. Наша работа как специалистов по информации не завершена, пока эти инструменты (или подходящие альтернативы) не станут доступны всем, кто в них нуждается.

Эта цель требует сделать инструменты более удобными в использовании. Предварительная обработка корпуса и курирование словарного запаса, необходимые для тематических моделей, непросты: должны ли мы удалить документы не на английском языке, что мы должны считать документом, как мы должны использовать метаданные? Выбор модели также не является тривиальным, чтобы понять смысл данных: сколько тем мы должны использовать, какую из многих возможных моделей мы должны использовать и какой метод вывода дает нам наилучший компромисс между скоростью и точностью? Существующие тематические модели плохо сообщают, какие варианты доступны пользователю и какие последствия имеют эти варианты.

Однако, даже если процесс создания тематической модели становится интуитивно понятным, выходные данные также должны быть интерпретируемыми. Распределения по словам — это язык, который эти модели используют для создания представлений коллекций документов, но это не то, как пользователи думают о темах: они скорее предпочли бы фразы [Mei et al., 2007b], предложения [Smith et al., 2016]. ] или изображения [Lau et al., 2014]. Однако предоставление этих представлений нетривиально и требует более глубокого понимания корпуса, чем могут обеспечить современные тематические модели.

Наконец, тематические модели нуждаются в более систематическом исследовании того, как они могут помочь пользователям в выполнении типичных задач поиска информации, организации и управления. В то время как приложения, рассматриваемые в этом обзоре, демонстрируют примеры того, как люди могут использовать тематические модели, от приложений от истории до политологии, тем, как тематические модели могут дополнять или заменять существующие рабочие процессы, не уделяется такого же внимания, как, например, поисковым системам.

**10.4**  **Кода**

Мы надеемся, что вам понравился наш обзор приложений тематических моделей.

Для получения дополнительной информации мы рекомендуем читателю исследовать

|  |  |
| --- | --- |
| *10.4. Кода* | 131 |

библиографию по тематическому моделированию, 1 присоединитесь к списку рассылки по тематическому моделированию, 2 или к связанной с книгой веб-странице.

* <https://mimno.infosci.cornell.edu/topics.html>
* [http s://lists.cs.princeton.edu/mailman/listinfo/topic-models](https://lists.cs.princeton.edu/mailman/listinfo/topic-models)

**использованная литература**

Эдоардо М. Айролди, Дэвид М. Блей, Стивен Э. Финберг и Эрик П. Син. Стохастические блок-модели смешанного членства. *Журнал исследований машинного обучения* , 9:1981–2014, 2008 г.

Николаос Алетрас, Тимоти Болдуин, Джей Хан Лау и Марк Стивенсон. Отображение меток тем для изучения цифровых библиотек. В *материалах совместной конференции IEEE/ACM по цифровым библиотекам* , стр. 239–248, 2014 г.

Марк Элджи-Хьюитт, Райан Хьюзер и Франко Моретти. По пунктам. масштаб, тематика и форма повествования. *Брошюры Стэнфордской литературной лаборатории* , 1 (10), октябрь 2015 г.

Джеймс Аллан, редактор. *Обнаружение темы и отслеживание: организация информации на основе событий* . Спрингер, 2002.

Лулва Аль-Сумаит, Даниэль Барбара и Карлотта Доменикони. Онлайн LDA: Адаптивные тематические модели для интеллектуального анализа текстовых потоков с приложениями для обнаружения и отслеживания тем. На *Международной конференции по интеллектуальному анализу данных* , 2008 г.

Анима Анандкумар, Дин П. Фостер, Дэниел Дж. Су, Шам М. Какаде и Йи Кай Лю. Спектральный алгоритм скрытого распределения Дирихле. В *материалах достижений в области нейронных систем обработки информации* , 2012 г.

Дэвид Анджеевски и Дэвид Баттлер. Скрытая обратная связь по теме для поиска информации. Обнаружение *знаний и интеллектуальный анализ данных* , 2011.

Дэвид Анджеевски, Сяоцзинь Чжу и Марк Крэйвен. Включение знаний предметной области в тематическое моделирование с помощью априорных данных леса Дирихле. В *материалах Международной конференции по машинному обучению* , 2009 г.

132

|  |  |
| --- | --- |
| *использованная литература* | 133 |

Санджив Арора, Ронг Ге, Йони Халперн, Дэвид М. Мимно, Анкур Мойтра, Дэвид Зонтаг, Ичен Ву и Майкл Чжу. Практический алгоритм тематического моделирования с доказуемыми гарантиями. В *материалах Международной конференции по машинному обучению* , 2013 г.

Антон Бакалов, Эндрю МакКаллум, Ханна Уоллах и Дэвид Мимно. Тематические модели для таксономий. На *совместной конференции по электронным библиотекам* , 2012 г.

Кайхан Батмангелич, Ардаван Саиди, Картик Нарасимхан и Самуэль Гершман. Непараметрическое сферическое тематическое моделирование с встраиванием слов. В *Трудах Ассоциации компьютерной лингвистики* , 2016.

Эрик П.С. Баумер, Дэвид Мимно, Шион Гуха, Эмили Куан и Джери К. Гей. Сравнение обоснованной теории и тематического моделирования: крайняя дивергенция или маловероятная конвергенция? *Журнал Ассоциации информационных наук и технологий* , 2017 г.

Джером Р. Беллегарда. Среда скрытого семантического анализа для языкового моделирования с большим размахом. На *Европейской конференции по речевым коммуникациям и технологиям* , 1997 г.

Джером Р. Беллегарда. Адаптация статистической языковой модели: обзор и перспективы. том 42, 2004.

Йошуа Бенжио, Режан Дюшарм, Паскаль Винсент и Кристиан Жанвен. Нейронно-вероятностная языковая модель. *Дж. Мах. Учиться. Рез.* , 3:1137–1155, март 2003 г.

Адам Бергер и Джон Лафферти. Информационный поиск как статистическая трансляция. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 1999 г.

Индраджит Бхаттачарья. Разрешение коллективных сущностей в реляционных данных. *Докторская диссертация, Мэрилендский университет, Колледж-Парк* , 2006 г.

Дэвид М. Блей. Тематическое моделирование и цифровые гуманитарные науки. *Журнал цифровых гуманитарных наук* , 2(1), 2012 г.

Дэвид М. Блей и Джон Лафферти. Коррелированная тематическая модель науки. *Анналы прикладной статистики* , страницы 17–35, 2007 г.

Дэвид М. Блей и Джон Д. Лафферти. Динамические тематические модели. В *материалах Международной конференции по машинному обучению* , 2006 г.

Дэвид М. Блей и Джон Д. Маколифф. Курируемые тематические модели. В *материалах достижений в области нейронных систем обработки информации* , 2007 г.

Дэвид М. Блей, Эндрю Нг и Майкл Джордан. Скрытое распределение Дирихле.

*Журнал исследований машинного обучения* , 3, 2003.

134 *ссылки*

Дэвид М. Блей, Томас Л. Гриффитс и Майкл И. Джордан. Вложенный процесс китайского ресторана и байесовский непараметрический вывод иерархий тем. *Журнал ACM* , 57(2):7:1–7:30, февраль 2010 г.

Шеннон Боуэн. Псевдособытия приносят дивиденды от Клеопатры до Chipotle.

*Неделя по связям с общественностью* , 2016.

Джордж Э. П. Бокс и Норман Р. Дрейпер. *Построение эмпирической модели и поверхности отклика* . Ряды Вили в вероятности и математической статистике: прикладная вероятность и статистика. Уайли, 1987 год.

Джордан Бойд-Грабер и Дэвид М. Блей. Многоязычные тематические модели для невыровненного текста. В *материалах о неопределенности в искусственном интеллекте* , 2009 г.

Джордан Бойд-Грабер и Филип Резник. Целостный анализ настроений на разных языках: многоязычное контролируемое скрытое распределение Дирихле. В *Трудах эмпирических методов обработки естественного языка* , 2010.

Джордан Бойд-Грабер, Дэвид М. Блей и Сяоджин Чжу. Тематическая модель для устранения неоднозначности смысла слова. В *Трудах эмпирических методов обработки естественного языка* , 2007.

Джордан Бойд-Грабер, Дэвид Мимно и Дэвид Ньюман. *Уход и кормление тематических моделей: проблемы, диагностика, доработки* . Справочники по CRCсовременных статистических методов. CRC Press, Бока-Ратон, Флорида, 2014 г.

Перси Уильямс Бриджмен. *Логика современной физики* . Макмиллан, Нью-Йорк, 1927 год.

Сэмюэл Броди и Мирелла Лапата. Байесовская индукция смысла слов. В *материалах Европейского отделения Ассоциации компьютерной лингвистики* , 2009 г.

Андре Дэвид Бронятовски, Марк Дредзе, Дж. Майкл Пол и Андреа Дуга. Использование социальных сетей для местного эпиднадзора за гриппом в городской больнице: ретроспективное обсервационное исследование. *JMIR Public Health and Surveillance* , 1(1):e5, май 2015 г.

Джон Берроуз. Дельта: мера стилистического различия и указатель вероятного авторства. *Lit Linguist Computing* , 17(3):267–287, 2002.

Хайме Дж. Карбонелл, Имин Ян, Роберт Э. Фредеркинг, Ральф Д. Браун, Ибинг Генг и Дэнни Ли. Транслингвальный поиск информации: сравнительная оценка. На *Международной объединенной конференции по искусственному интеллекту* , 1997 г.

Марк Джеймс Карман, Фабио Крестани, Морган Харви и Марк Бэйли. На пути к персонализации на основе журнала запросов с использованием тематических моделей. В *материалах Международной конференции ACM по управлению информацией и знаниями* , 2010 г.

|  |  |
| --- | --- |
| *использованная литература* | 135 |

Ёнчул Ча и Чонху Чо. Анализ социальных сетей с использованием тематических моделей. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 2012 г.

Эллисон Чейни и Дэвид М. Блей. Визуализация тематических моделей. На *Международной конференции AAAI по блогам и социальным сетям* , 2012 г.

Джонатан Чанг и Дэвид М. Блей. Реляционные тематические модели для сетей документов. В *Трудах искусственного интеллекта и статистики* , 2009 г.

Джонатан Чанг, Джордан Бойд-Грабер, Чонг Ван, Шон Герриш и Дэвид М. Блей. Чтение чайных листьев: как люди интерпретируют тематические модели. В *материалах достижений в области систем обработки нейронной информации* , 2009 г.

Бокс Чен, Джордж Фостер и Роланд Кун. Адаптация моделей переупорядочивания для статистического машинного перевода. В *материалах Ассоциации компьютерной лингвистики* , 2013 г.

Стэнли Чен, Дуглас Биферман и Рональд Розенфельд. Метрики оценки языковых моделей. Технический отчет, Школа компьютерных наук Университета Карнеги-Меллона, 1998 г.

Дэвид Чанг, Стив ДеНиф и Майкл Пуст. Два простых улучшения лексического взвешивания. В *материалах конференции по технологиям человеческого языка* , 2011 г.

Джэгул Чу, Чанхён Ли, Чандан К. Редди и Хэсун Пак. UTOPIAN: Тематическое моделирование, управляемое пользователем, основанное на интерактивной неотрицательной матричной факторизации. *IEEE Transactions on Visualization and Computer Graphics* , 19(12):1992–2001, 2013.

Джейсон Чуанг, Кристофер Д. Мэннинг и Джеффри Хир. Термит: методы визуализации для оценки текстовых тематических моделей. В *расширенных визуальных интерфейсах* , 2012.

Джейсон Чуанг, Маргарет Э. Робертс, Брэндон М. Стюарт, Ребекка Вайс, Дастин Тингли, Джастин Гриммер и Джеффри Хир. TopicCheck: интерактивное выравнивание для оценки стабильности тематической модели. На *конференции Североамериканского отделения Ассоциации компьютерной лингвистики* , 2015 г.

Филип Р. Кларксон и Энтони Дж. Робинсон. Адаптация языковой модели с использованием смесей и экспоненциально распадающегося кеша. На *Международной конференции по акустике, речи и обработке сигналов* , 1997 г.

Ной Коккаро и Даниэль Джурафски. К лучшей интеграции семантических предикторов в статистическом моделировании языка. На *Международной конференции по акустике, речи и обработке сигналов* , 1998 г.

136 *ссылок*

Ронан Коллобер, Корай Кавукчуоглу и Клеман Фарабет. Torch7: подобная Matlab среда для машинного обучения. На *семинаре NIPS по большому обучению (Biglearn)* , 2011 г.

Коби Краммер, Офер Декель, Джозеф Кешет, Шай Шалев-Шварц и Йорам Сингер. Онлайн пассивно-агрессивные алгоритмы. *Журнал исследований машинного обучения* , 7:551–585, 2006 г.

У. Брюс Крофт и Джон Лафферти. Языковое моделирование для поиска информации.

В *международной серии Kluwer по информационному поиску* , 2003 г.

Ван Данг и В. Брюс Крофт. Диверсификация результатов поиска на уровне терминов. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 2013 г.

Хэл Доме III. Марковские случайные тематические поля. В *Трудах искусственного интеллекта и статистики* , 2009 г.

Вим Де Смет и Мари-Франсин Моэнс. Межъязыковое связывание новостей в Интернете с использованием межъязыкового тематического моделирования. На *семинаре по поиску и добыче полезных ископаемых в социальных сетях* , 2009 г.

Скотт Дирвестер, Сьюзан Дюме, Томас Ландауэр, Джордж Фернас и Ричард Харшман. Индексирование с помощью скрытого семантического анализа. *Журнал Американского общества информационных наук* , 41(6):391–407, 1990.

Джейкоб Девлин, Раби Збиб, Чжунцян Хуан, Томас Ламар, Ричард Шварц и Джон Махоул. Быстрые и надежные совместные модели нейронных сетей для статистического машинного перевода. В *Трудах Ассоциации компьютерной лингвистики* , 2014.

Лаура Дитц, Штеффен Бикель и Тобиас Шеффер. Неконтролируемое прогнозирование влияния цитирования. В *материалах Международной конференции по машинному обучению* , 2007 г.

Чжичэн Доу, Жуйхуа Сун и Цзи-Ронг Вэнь. Масштабная оценка и анализ стратегий персонализированного поиска. В *материалах конференции World Wide Web* , 2007 г.

Владимир Эйдельман, Джордан Бойд-Грабер и Филип Резник. Тематические модели для адаптации модели динамического перевода. В *материалах Ассоциации компьютерной лингвистики* , 2012 г.

Джейкоб Эйзенштейн. *Письменные диалектные вариации в социальных сетях* . Уайли, 2017.

Джейкоб Эйзенштейн, Брендан О'Коннор, Ной А. Смит и Эрик П. Син. Модель скрытой переменной для географической лексической вариации. В *Трудах эмпирических методов обработки естественного языка* , 2010.

|  |  |
| --- | --- |
| *использованная литература* | 137 |

Джейкоб Эйзенштейн, Дуен Хорнг Чау, Аникет Киттур и Эрик Син. TopicViz: интерактивное изучение тем в коллекциях документов. В *расширенных тезисах конференции ACM по человеческому фактору в вычислительных системах* , 2012 г.

Джейкоб Эйзенштейн, Айрис Сан и Лорен Ф. Кляйн. Исследовательский анализ текста для больших архивов документов. В *цифровых гуманитарных науках* , 2014.

Мэтт Эрлин. Тематическое моделирование, эпистемология и английский и немецкий роман.

*Культурная аналитика* , май 2017 г.

Джордж Фостер и Роланд Кун. Адаптация смешанной модели для SMT. В *материалах второго семинара по статистическому машинному переводу* ,2007.

Цзяньфэн Гао, Кристина Тутанова и Вэнь Тау И. Скрытые семантические модели на основе кликов для веб-поиска. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 2011 г.

Цзяньфэн Гао, Шаша Се, Сяодун Хэ и Альнур Али. Изучение моделей лексики из журналов поиска для расширения запроса. В *Трудах эмпирических методов обработки естественного языка* , 2012.

Мэтью Гарднер, Джошуа Лютес, Джефф Лунд, Джош Хансен, Дэн Уокер, Эрик Ринггер и Кевин Сеппи. Тематический браузер: интерактивный инструмент для просмотра тематических моделей. На *семинаре NIPS по проблемам визуализации данных* , 2010 г.

Шон Герриш и Дэвид М. Блей. Языковой подход к измерению научного воздействия. В *материалах Международной конференции по машинному обучению* , 2010 г.

Дэниел Гилдеа и Томас Хофманн. Тематические языковые модели с использованием EM.

На *Европейской конференции по речевым коммуникациям и технологиям* , 1999 г.

Барни Г. Глейзер и Анслем Штраус. *Открытие теории заземления:*

*Стратегии качественных исследований* . Альдин, 1967 год.

Эндрю Голдстоун и Тед Андервуд. Тихие трансформации литературоведения: что могли бы нам сказать тринадцать тысяч ученых. *Новая литературная история* , 45 (3), лето 2014 г.

Томас Л. Гриффитс и Марк Стейверс. Поиск научных тем. *Труды Национальной академии наук* , 101 (Приложение 1): 5228–5235, 2004 г.

Джастин Гриммер. Байесовская иерархическая тематическая модель для политических текстов: измерение повесток дня, выраженных в пресс-релизах Сената. *Политический анализ* , 18(1):1–35, 2010.

Амит Грубер, Майкл Розен-Цви и Яир Вайс. Скрытая тема Марковские модели.

В *искусственном интеллекте и статистике* , 2007.

138 *ссылок*

Эрик Хардисти, Джордан Бойд-Грабер и Филип Резник. Моделирование перспективы с использованием грамматик адаптеров. В *Трудах эмпирических методов обработки естественного языка* , 2010.

Джейкоб Харрис. Облака слов считаются вредными. [http://www.niemanlab.org/](http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/) [2011/10/слово-облака-считаются-вредными/](http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/) , 2011.

Морган Харви, Фабио Крестани и Марк Джеймс Карман. Создание профилей пользователей из тематических моделей для персонализированного поиска. В *материалах Международной конференции ACM по управлению информацией и знаниями* ,2013.

Ева Хаслер, Барри Хэддоу и Филипп Коэн. Редкие лексикализованные функции и тематическая адаптация для SMT. В *материалах Международного семинара по устному переводу* , 2012 г.

Ци Хэ, Би Чен, Цзянь Пей, Баоцзюнь Цю, Прасенджит Митра и Ли Джайлз. Выявление эволюции темы в научной литературе: чем могут помочь цитаты? В *материалах Международной конференции ACM по управлению информацией и знаниями* , 2009 г.

Линетт Хиршман и Роб Гайзаускас. Ответ на вопрос на естественном языке: Вид отсюда. *Natural Language Engineering* , 7(4):275–300, декабрь 2001 г.

Мэтью Хоффман, Дэвид М. Блей и Фрэнсис Бах. Онлайн-обучение скрытому распределению Дирихле. В *материалах достижений в области систем обработки нейронной информации* , 2010 г.

Томас Хофманн. Вероятностный латентный семантический анализ. В *материалах о неопределенности в искусственном интеллекте* , 1999a.

Томас Хофманн. Вероятностное скрытое семантическое индексирование. *Материалы конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 1999b.

Пол В. Холланд, Кэтрин Блэкмонд Ласки и Сэмюэл Лейнхардт. Стохастические блок-модели: первые шаги. *Социальные сети* , 5(2):109–137, 1983.

Лянцзе Хонг и Брайан Д. Дэвисон. Эмпирическое исследование тематического моделирования в Twitter. В *материалах первого семинара по аналитике социальных сетей* , 2010 г.

Дайан Ху и Лоуренс К. Сол. Вероятностная модель неконтролируемого обучения для профилей музыкальных тональностей. На *конференции Международного общества музыкального информационного поиска* , 2009 г.

Юнин Ху, Джордан Бойд-Грабер, Брианна Сатинофф и Элисон Смит. Интерактивное тематическое моделирование. *Журнал машинного обучения* , 95(3):423–469, июнь 2014 г.

|  |  |
| --- | --- |
| *использованная литература* | 139 |

Юнин Ху, Ке Чжай, Владимир Эдельман и Джордан Бойд-Грабер. Многоязычные древовидные тематические модели для адаптации домена перевода. В *Ассоциации компьютерной лингвистики* , 2014b.

Рукмини Айер и Мари Остендорф. Моделирование зависимостей на большом расстоянии в языке: смешение тем и модели динамического кэша. *IEEE Transactions on Speech Audio Process* , 7:236–239, 1999.

Мохит Ийер, Варун Манджунатха, Джордан Бойд-Грабер и Хэл Дауме III.

Глубокая неупорядоченная композиция конкурирует с синтаксическими методами классификации текста.

В *Ассоциации компьютерной лингвистики* , 2015.

Мохит Ийер, Анупам Гуха, Снигдха Чатурведи, Джордан Бойд-Грабер и Хэл Дауме III. Враждующие семьи и бывшие друзья: неконтролируемое обучение динамичным вымышленным отношениям. В *Североамериканской ассоциации компьютерной лингвистики* , 2016 г.

Бернард Дж. Янсен, Аманда Спинк и Тефко Сарацевич. Реальная жизнь, реальные пользователи и реальные потребности: изучение и анализ пользовательских запросов в сети. *Обработка информации и управление* , 36(2):207–227, 2000.

Фред Елинек и Роберт Л. Мерсер. Интерполированная оценка параметров марковского источника по разреженным данным. В *материалах семинара по распознаванию образов на практике* , 1980 г.

Шань Цзян, Юнин Ху, Чансун Кан, Тим Дейли, Давэй Инь и И Чанг. Изучение релевантности запроса и документа на основе графика кликов в масштабе Интернета. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 2016 г.

Йохан Джо и Элис Х. О. Модель объединения аспектов и настроений для анализа онлайн-обзоров. В *материалах Международной конференции ACM по поиску в Интернете и интеллектуальному анализу данных* , 2011 г.

Мэтт Л. Джокерс. *Макроанализ: цифровые методы и история литературы* . Темы цифровых гуманитарных наук. Университет Иллинойса, 2013.

Мэтью Джокерс и Дэвид Мимно. Значимые темы в литературе 19 века. *Поэтика* , 41(6):750–769, декабрь 2013 г.

Патрик Джуола. Атрибуция авторства. *Основы и тенденции информационного поиска* , 1(3):233–334, 2006.

Саураб С. Катариа, Кришнан С. Кумар, Раджив Р. Растоги, Притхвирадж Сен и Шринивасан Х. Сенгамеду. Устранение неоднозначности сущностей с помощью иерархических тематических моделей. Обнаружение *знаний и интеллектуальный анализ данных* , 2011.

СМ Кац. Оценка вероятностей по разреженным данным для компонента языковой модели распознавателя речи. В *IEEE Transaction on Acoustics, Speech and Signal Processing* , 1987.

140 *ссылок*

Кирилл Киреев, Лейсия Пален и Кеннет Андерсон. Применение тематических моделей к анализу данных Twitter, связанных со стихийными бедствиями. декабрь 2009 г.

Рейнхард Кнезер и Йохен Петерс. Семантическая кластеризация для адаптивного языкового моделирования. На *Международной конференции по акустике, речи и обработке сигналов* , 1997 г.

Райнхард Кнезер, Йохен Петерс и Дитрих Клаков. Адаптация языковой модели с использованием динамических маргиналов. На *Европейской конференции по речевым коммуникациям и технологиям* , 1997 г.

Филипп Коэн. *Статистический машинный перевод* . Издательство Кембриджского университета,

2009 г. URL -адрес <http://books.google.com/books?id=D21UAAAACAAJ>.

Филипп Кён, Франц Йозеф Ох и Даниэль Марку. Статистический фразовый перевод. На *конференции Североамериканского отделения Ассоциации компьютерной лингвистики* , 2003 г.

Томас К. Ландауэр и Майкл Л. Литтман. Полностью автоматический поиск документов на разных языках с использованием скрытой семантической индексации. В *материалах Центра Нового Оксфордского словаря английского языка UW* , 1990.

Джон Лэнгфорд, Лихонг Ли и Алекс Стрель. Вовпал Ваббит, 2007.

Марк А. Ларгент и Джулия И. Лейн. ЗВЕЗДНЫЕ ПОКАЗАТЕЛИ и наука о научной политике. *Обзор политических исследований* , 29(3):431–438, 2012 г.

Джей Хан Лау, Дэвид Ньюман, Сарвназ Карими и Тимоти Болдуин. Лучший выбор слова темы для маркировки темы. В *материалах Международной конференции по компьютерной лингвистике* , 2010 г.

Джей Хан Лау, Карл Гризер, Дэвид Ньюман и Тимоти Болдуин. Автоматическая маркировка тематических моделей. В *материалах Ассоциации компьютерной лингвистики* , 2011 г.

Джей Хан Лау, Дэвид Ньюман и Тимоти Болдуин. Машинное чтение чайных листьев: автоматическая оценка согласованности темы и качества модели темы. В *материалах Европейского отделения Ассоциации компьютерной лингвистики* , 2014 г.

Виктор Лавренко и У. Брюс Крофт. Релевантные языковые модели. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 2001 г.

Юре Лесковец, Ларс Бэкстром и Джон Клейнберг. Мем-трекинг и динамика новостного цикла. В *книге «Обнаружение знаний и интеллектуальный анализ данных* », 2009 г.

|  |  |
| --- | --- |
| *использованная литература* | 141 |

Аарон К. Ли, Амр Ахмед, Суджит Рави и Александр Дж. Смола. Уменьшение сложности выборки тематических моделей. Обнаружение *знаний и интеллектуальный анализ данных* , 2014.

Перси Лян и Дэн Кляйн. Структурированные байесовские непараметрические модели с вариационным выводом (учебник). В *Трудах Ассоциации компьютерной лингвистики* , 2007 г.

Шансонг Лян, Чжаочунь Рен и Маартен де Рийке. Персонализированная диверсификация результатов поиска с помощью структурированного обучения. Обнаружение *знаний и интеллектуальный анализ данных* , 2014.

Чэнхуа Линь и Юйлан Хэ. Совместная модель настроений/тем для анализа настроений. В *материалах Международной конференции ACM по управлению информацией и знаниями* , 2009 г.

Тяньи Линь, Вэнтао Тянь, Цяочжу Мэй и Хун Ченг. Модель с двойной разреженной темой: извлечение сфокусированных тем и сфокусированных терминов в коротком тексте. В *материалах конференции World Wide Web* , 2014 г.

Ян Лю, Александру Никулеску-Мизиль и Войцех Грыц. Ссылка на тему LDA: Совместные модели тематического и авторского сообщества. В *материалах Международной конференции по машинному обучению* , 2009 г.

Юэ Лу и Чэнсян Чжай. Интеграция мнений посредством тематического моделирования с полуучителем. В *материалах конференции World Wide Web* , 2008 г.

Юэ Лу, Цяочжу Мэй и Ченгсян Чжай. Исследование выполнения задач вероятностных тематических моделей: эмпирическое исследование PLSA и LDA. *Информационный поиск* , 14(2):178–203, 2011.

Джефф Лунд, Коннор Кук, Кевин Сеппи и Джордан Бойд-Грабер. Тандемная привязка: многословный подход к интерактивному тематическому моделированию. В *Ассоциации компьютерной лингвистики* , 2017.

Дэвид Дж. К. Маккей и Линда К. Бауман Петой. Иерархическая модель языка Дирихле. *Инженерия естественного языка* , 1:1–19, 1995.

Гидеон Манн, Дэвид Мимно и Эндрю МакКаллум. Библиометрические измерения влияния, использующие тематический анализ. На *совместной конференции по электронным библиотекам* , 2006 г.

Сянь-Лин Мао, Чжао-Янь Мин, Чжэн-Цзюнь Чжа, Тат-Сэн Чуа, Хунфэй Янь и Сяомин Ли. Автоматическая маркировка иерархических тем. В *материалах Международной конференции ACM по управлению информацией и знаниями* , 2012 г.

Митчелл П. Маркус, Беатрис Санторини и Мэри А. Марцинкевич. Создание большого аннотированного корпуса английского языка: The Penn Treebank. *Компьютерная лингвистика* , 19(2):313–330, 1993.

142 *ссылки*

Гириш Маскери, Сантону Саркар и Кеннет Хифилд. Добыча бизнес-тем в исходном коде с использованием скрытого распределения дирихле. *Индийская конференция по разработке программного обеспечения* , 2008 г.

Спирос Мацукас, Антти-Вейкко И. Рости и Бинг Чжан. Дискриминационная оценка веса корпуса для машинного перевода. В *Трудах эмпирических методов обработки естественного языка* , 2009 г.

Эндрю МакКаллум, Сюэруи Ван и Андрес Коррада-Эммануэль. Открытие тем и ролей в социальных сетях с экспериментами на enron и академической электронной почте. *Журнал исследований искусственного интеллекта* , 30(1):249–272, октябрь 2007 г.

Эндрю Кахитес МакКаллум. Маллет: машинное обучение для языкового инструментария, 2002 г. http://www.cs.umass.edu/mccallum/mallet.

Ришаб Мехротра, Скотт Саннер, Рэй Бантин и Лексин Се. Улучшение тематических моделей lda для микроблогов с помощью объединения твитов и автоматической маркировки. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 2013 г.

Цяочжу Мэй и Чэнсян Чжай. Обнаружение эволюционных шаблонов темы из текста: исследование временного анализа текста. В *книге «Обнаружение знаний и интеллектуальный анализ данных* », 2005 г.

Цяочжу Мэй, Чао Лю, Ханг Су и Чэнсян Чжай. Вероятностный подход к исследованию пространственно-временных шаблонов тем в блогах. В *материалах конференции World Wide Web* , 2006 г.

Цяочжу Мэй, Сюй Линг, Мэтью Вондра, Ханг Су и Чэнсян Чжай. Смесь настроений тем: моделирование аспектов и мнений в блогах. В *материалах конференции World Wide Web* , 2007a.

Цяочжу Мэй, Сюэхуа Шэнь и Ченгсян Чжай. Автоматическая маркировка полиномиальных тематических моделей. Обнаружение *знаний и интеллектуальный анализ данных* , 2007b.

Цяочжу Мэй, Дэн Цай, Дуо Чжан и Чэн Сян Чжай. Тематическое моделирование с сетевой регуляризацией. В *материалах конференции World Wide Web* , 2008 г.

Массимо Мелуччи. Контекстный поиск: вычислительная структура. *Основы и тенденции информационного поиска* , 6: 257–405, 2012 г.

Алессандро Микарелли, Фабио Гаспаретти, Филиппо Шарроне и Сьюзен Гоуч. Персонализированный поиск во всемирной паутине. В *«Адаптивной сети* », том 4321, 2007 г.

Ян Мэтью Миллер. Восстание, преступность и насилие в цинском Китае, 1722–1911 годы: подход к моделированию тем. *Поэтика* , 41(6):626–649, декабрь 2013 г.

|  |  |
| --- | --- |
| *использованная литература* | 143 |

Дэвид Мимно. Вычислительная историография: интеллектуальный анализ данных в классических журналах века. *Journal on Computing and Cultural Heritage* , 5(1):3:1–3:19, апрель 2012 г.

Дэвид Мимно и Дэвид М. Блей. Байесовская проверка тематических моделей. В *Трудах эмпирических методов обработки естественного языка* , 2011.

Дэвид Мимно и Эндрю МакКаллум. Тематические модели, обусловленные произвольными признаками, с полиномиальной регрессией Дирихле. В *материалах конференции 2008 г. по неопределенности в искусственном интеллекте (UAI)* , 2008 г.

Дэвид Мимно, Ханна Уоллах, Джейсон Нарадовски, Дэвид Смит и Эндрю МакКаллум. Полиязычные тематические модели. В *Трудах эмпирических методов обработки естественного языка* , 2009 г.

Дэвид Мимно, Ханна Уоллах, Эдмунд Талли, Мириам Линдерс и Эндрю МакКаллум. Оптимизация семантической согласованности в тематических моделях. В *Трудах эмпирических методов обработки естественного языка* , 2011.

Дэвид Мимно, Мэтью Хоффман и Дэвид М. Блей. Разреженный стохастический вывод для скрытого распределения Дирихле. В *материалах Международной конференции по машинному обучению* , 2012 г.

Том Минка, Джон Винн, Джон Гивер и Дэвид Ноулз. Infer.NET 2.6, 2014.

Исследовательский центр Майкрософт в Кембридже. http://research.microsoft.com/infernet.

Франко Моретти. Бойня литературы. *Modern Language Quarterly* , 61 (1): 207–227, 2000.

Франко Моретти. *Дистанционное чтение* . Оборот, 2013а. URL -адрес [https://books.google.](https://books.google.com/books?id=YKMCy9I3PG4C)

[com/книги?id=YKMCy9I3PG4C](https://books.google.com/books?id=YKMCy9I3PG4C) .

Франко Моретти. Операционализация, или функция измерения в теории литературы. *New Left Review* , 84, ноябрь/декабрь 2013b.

Фредерик Мостеллер и Дэвид Л. Уоллес. *Вывод и оспариваемое авторство:*

*Федералист* . Аддисон-Уэсли, Рединг, Массачусетс, 1964.

Кристоф Мюллер и Ирина Гуревич. Исследование семантической связи терминов запроса и документа при поиске информации. В *Трудах эмпирических методов обработки естественного языка* , 2009 г.

Рамеш Наллапати и Уильям Коэн. Link-PLSA-LDA: новая неконтролируемая модель тем и влияния блогов. На *Международной конференции по блогам и социальным сетям* , 2008 г.

Шраван Нараянамурти. Яху! LDA, 2011. URL [https://github.com/](https://github.com/shravanmn/Yahoo_LDA/wiki) [shravanmn/Yahoo\_L DA/wiki](https://github.com/shravanmn/Yahoo_LDA/wiki) .

Рэдфорд М. Нил. Вероятностный вывод с использованием методов Монте-Карло цепи Маркова. Технический отчет CRG-TR-93-1, Университет Торонто, 1993 г.

144 *ссылки*

Дэвид Ньюман, Артур Асунсьон, Падраик Смит и Макс Веллинг. Распределенный вывод для скрытого распределения Дирихле. В *материалах достижений в области нейронных систем обработки информации* , 2008 г.

Дэвид Ньюман, Джей Хан Лау, Карл Гризер и Тимоти Болдуин. Автоматическая оценка связности темы. На *конференции Североамериканского отделения Ассоциации компьютерной лингвистики* , 2010 г.

Дэвид Дж. Ньюман и Шэрон Блок. Вероятностная декомпозиция темы американской газеты восемнадцатого века. *Журнал Американского общества информационных наук и технологий* , 18(1):753–767, 2006 г.

Герман Ней, Уте Эссен и Рейнхард Кнезер. О структурировании вероятностных зависимостей в стохастическом языковом моделировании. *Компьютерная речь и язык* , 8:1–38, 1994.

Танг Нгуен, Джордан Бойд-Грабер, Джефф Лунд, Кевин Сеппи и Эрик Ринггер.

Ваш якорь поднимается или опускается? Быстрые и точные контролируемые тематические модели.

В *Североамериканской ассоциации компьютерной лингвистики* , 2015a.

Вьет-Ан Нгуен, Джордан Бойд-Грабер и Филип Резник. Лексическая и иерархическая тематическая регрессия. В *материалах достижений в области систем обработки нейронной информации* , 2013 г.

Вьет-Ан Нгуен, Джордан Бойд-Грабер, Филип Резник и Джонатан Чанг. Изучение иерархии понятий из документов с несколькими метками. В *системах обработки нейронной информации* , 2014.

Вьет-Ан Нгуен, Джордан Бойд-Грабер, Филип Резник и Кристина Милер. Чаепитие в доме: иерархическая идеальная точечная модель темы и ее применение к законодателям-республиканцам на 112-м Конгрессе. В *Ассоциации компьютерной лингвистики* , 2015b.

Сяочуань Ни, Цзянь-Тао Сунь, Цзянь Ху и Чжэн Чен. Добыча многоязычных тем из Википедии. В *материалах конференции World Wide Web* , 2009 г.

Брендан О'Коннор, Рамнат Баласубраманян, Брайан Р. Рутледж и Ной А. Смит. От твитов к опросам: связь настроений в тексте с временными рядами общественного мнения. В *материалах Международной конференции AAAI по блогам и социальным сетям* , 2010 г.

Лоуренс Пейдж, Сергей Брин, Раджив Мотвани и Терри Виноград. Рейтинг цитируемости PageRank: наведение порядка в сети. Рабочий документ Стэнфордской цифровой библиотеки SIDL-WP-1999-0120, Стэнфордский университет, 1999.

Бо Панг и Лилиан Ли. *Изучение мнений и анализ настроений* . Сейчас Издатели Inc, 2008.

|  |  |
| --- | --- |
| *использованная литература* | 145 |

Христос Х. Пападимитриу, Прабхакар Рагхаван, Хисао Тамаки и Сантош Вемпала. Скрытое семантическое индексирование: вероятностный анализ. *Журнал компьютерных и системных наук* , 61(2):217–235, 2000.

Лоуренс А. Парк и Котагири Рамамоханарао. Чувствительность скрытого распределения Дирихле для поиска информации. В *материалах Европейской конференции по машинному обучению и обнаружению знаний в базах данных* ,2009.

Майкл Пол и Роксана Гирью. Двумерная тема-аспектная модель для обнаружения многогранных тем. В *материалах Ассоциации по развитию искусственного интеллекта* , 2010 г.

Джеймс Питкоу, Хинрих Шютце, Тодд Касс, Роб Кули, Дон Тернбулл, Энди Эдмондс, Эйтан Адар и Томас Брюэль. Персонализированный поиск. *Сообщения ACM* , 45(9):50–55, сентябрь 2002 г.

Джей М. Понте и В. Брюс Крофт. Подход языкового моделирования к поиску информации. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 1998 г.

Джонатан К. Притчард, Мэтью Стивенс и Питер Доннелли. Вывод о структуре популяции с использованием данных о многолокусных генотипах. *Генетика* , 155:945–959, 2000.

Дэниел Рэймидж, Дэвид Холл, Рамеш Наллапати и Кристофер Мэннинг. Labeled LDA: контролируемая тематическая модель для атрибуции кредита в мультимаркированных корпусах. В *Трудах эмпирических методов обработки естественного языка* , 2009 г.

Дэниел Рэймидж, Сьюзан Т. Дюме и Дэниел Дж. Либлинг. Характеристика микроблогов с помощью тематических моделей. На *Международной конференции по блогам и социальным* сетям, 2010a.

Дэниел Рэймидж, Кристофер Д. Мэннинг и Дэниел А. Макфарланд. Какие вузы лидируют, а какие отстают? к рейтингам университетов, основанным на научных результатах. На *семинаре NIPS по вычислительным социальным наукам и мудрости толпы* , 2010b.

Раджеш Ранганат, Линпэн Танг, Лоран Чарлин и Дэвид М. Блей. Глубокие экспоненциальные семейства. В *Трудах искусственного интеллекта и статистики* , 2015.

Филип Резник и Эрик Хардисти. Выборка Гиббса для непосвященных. Технический отчет, Мэрилендский университет, 2009 г. URL [http://www.umiacs.umd.edu/](http://www.umiacs.umd.edu/~resnik/pubs/gibbs.pdf) [~resnik/pubs/gibbs.pdf](http://www.umiacs.umd.edu/~resnik/pubs/gibbs.pdf) .

Лия М. Роди. Тематическое моделирование и образный язык. *Журнал цифровых гуманитарных наук* , 2(1), 2012 г.

146 *ссылок*

Аллен Бей Ридделл. *Как читать 22 198 журнальных статей: Изучение истории германистики с помощью тематических моделей* , страницы 91–113. Камден Хаус, 2012.

Маргарет Э. Робертс, Брэндон М. Стюарт и Дастин Тингли. STM: Пакет R для структурных тематических моделей, 2014. URL [http://www.](http://www.structuraltopicmodel.com) [www.структурнаяtopicmodel.com](http://www.structuraltopicmodel.com) . Пакет R версии 1.0.8.

Джозеф Джон Роккио. *Обратная связь по релевантности в информационном поиске* , страницы 313–323. Энглвуд Клиффс, Нью-Джерси: Прентис-Холл, 1971.

Михал Розен-Цви, Томас Л. Гриффитс, Марк Стейверс и Падраик Смит. Авторо-тематическая модель для авторов и документов. В *материалах о неопределенности в искусственном интеллекте* , 2004 г.

Ник Руис и Марчелло Федерико. Адаптация темы для перевода лекций с помощью двуязычных латентных семантических моделей. На *семинаре WMT по статистическому машинному переводу* , 2011 г.

Джерард. Солтон. *Автоматическая организация и поиск информации.* Текст Макгроу Хилл, 1968.

Родриго Л. Т. Сантос, Крейг Макдональд и Иад Оунис. Диверсификация результатов поиска. *Основы и тенденции информационного поиска* , 9(1):1–90, март 2015 г.

Тобиас Шнабель, Игорь Лабутов, Дэвид М. Мимно и Торстен Йоахимс. Методы оценки неконтролируемых вложений слов. В *Трудах эмпирических методов обработки естественного языка* , 2015.

Кристи Сеймор и Рональд Розенфельд. Использование сюжетных тем для адаптации языковой модели. На *Европейской конференции по речевым коммуникациям и технологиям* , 1997 г.

Кристи Сеймор, Стэнли Ф. Чен и Рональд Розенфельд. Нелинейная интерполяция тематических моделей для адаптации языковой модели. На *Международной конференции по обработке разговорной речи* , 1998 г.

Элисон Смит, Джейсон Чуан, Юнин Ху, Джордан Бойд-Грабер и Лия Финдлейтер. Параллельная визуализация отношений между словами и темами в тематических моделях. На *семинаре ACL на семинаре по интерактивному изучению языков, визуализации и интерфейсов* , 2014 г.

Элисон Смит, Сана Малик и Бен Шнейдерман. *Визуальный анализ тематического развития в неструктурированном тексте: Дизайн и оценка тематического потока* , страницы159–175. Спрингер, 2015.

Элисон Смит, Так Ён Ли, Форуг Пурсабзи-Сангде, Лия Финдлейтер, Джордан Бойд-Грабер и Никлас Эльмквист. Оценка визуальных представлений для понимания темы и их влияния на созданные вручную метки. *Труды Ассоциации компьютерной лингвистики* , 2016.

|  |  |
| --- | --- |
| *использованная литература* | 147 |

Ричард Сочер, Броуди Хувал, Кристофер Д. Мэннинг и Эндрю Ю. Нг. Семантическая композиционность через рекурсивные матрично-векторные пространства. В *Pro-ceedings of Empirical Methods in Natural Language Processing* , 2012.

Фей Сонг и В. Брюс Крофт. Общая языковая модель для поиска информации. На *Международной конференции по управлению информацией и знаниями* , 1999 г.

Вэй Сун, Юй Чжан, Тин Лю и Шэн Ли. Объединение тематического моделирования и персонализированного поиска. В *материалах Международной конференции по компьютерной лингвистике* , 2010 г.

Команда разработчиков Стэн. Стэн: библиотека C++ для вероятности и выборки, версия 2.5.0, 2014 г. URL <http://mc-stan.org/>.

Марк Стейверс, Падраик Смит, Михал Розен-Цви и Томас Гриффитс. Вероятностные авторско-тематические модели для обнаружения информации. В *книге «Обнаружение знаний и интеллектуальный анализ данных* », 2004 г.

Цзиньсон Су, Хуа Ву, Хайфэн Ван, Идун Чен, Сяодун Ши, Хуайлинь Дун и Цюнь Лю. Адаптация модели перевода для статистического машинного перевода с одноязычной тематической информацией. В *материалах Ассоциации компьютерной лингвистики* , 2012 г.

Ичжоу Сунь, Цзявэй Хань, Цзин Гао и Иньтао Ю. iTopicModel: Тематическое моделирование, интегрированное в информационную сеть. На *Международной конференции по интеллектуальному анализу данных* , 2009 г.

Кристиан Сегеди, Войцех Заремба, Илья Суцкевер, Джоан Бруна, Думитру Эрхан, Ян Гудфеллоу и Роб Фергус. Интересные свойства нейронных сетей. На *Международной конференции по обучающим представлениям* , 2014 г.

Рик Шостак. Классифицирующая наука. *Классификация науки: явления, данные, теория, метод, практика* , страницы 1–22, 2004 г.

Эдмунд М. Тэлли, Дэвид Ньюман, Дэвид Мимно, Брюс В. Херр, Ханна М. Уоллах, Галли А.П.С. Бернс, А. Г. Мириам Лендерс и Эндрю МакКаллум. База данных грантов NIH с использованием категорий с машинным обучением и графической кластеризации. *Nature Methods* , 8(6):443–444, май 2011 г.

Йик-Чунг Там, Ян Лейн и Таня Шульц. Адаптация фильма на основе двуязычного ИСА для устного языкового перевода. В *Трудах Ассоциации компьютерной лингвистики* , 2007 г.

Цзянь Тан, Чжаоши Мэн, СюаньЛонг Нгуен, Цяочжу Мэй и Минг Чжан. Понимание ограничивающих факторов тематического моделирования с помощью апостериорного анализа сокращения. В *материалах Международной конференции по машинному обучению* , 2014 г.

148 *ссылок*

Тимоти Р. Тангерлини и Питер Леонард. Блуждание в море великого непрочитанного: Подкорпусное тематическое моделирование и гуманитарные исследования. *Поэтика* , 41(6): 725–749, декабрь 2013 г.

Команда разработчиков Теано. Theano: платформа Python для быстрого вычисления математических выражений. *arXiv e-prints* , abs/1605.02688, май 2016 г.

Иван Титов и Райан Макдональд. Совместная модель оценок текста и аспектов для обобщения настроений. В *Трудах Ассоциации компьютерной лингвистики* , 2008 г.

Кристина Тутанова и Марк Джонсон. Модель на основе байесовского LDA для полуконтролируемой маркировки частей речи. В *материалах достижений в области нейронных систем обработки информации* , 2008 г.

Дэвид Валлет и Пабло Кастельс. Персонализированная диверсификация результатов поиска. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 2012 г.

Фернанда Б. Вьегас и Мартин Ваттенберг. Облака тегов и необходимость народной визуализации. *Взаимодействия* , 15(4):49–52, 2008.

Максимилиан Вирмец, Михал Скубач, Кай-Николас Циглер и Дитмар Зайпель. Отслеживание эволюции темы в новостной среде. На *Международной конференции IEEE по технологиям электронной коммерции* , 2008 г.

Эллен М. Вурхиз. Обзор TREC 2003. В *материалах конференции Text REtrieval* , страницы 1–13, 2003 г.

Эллен М. Вурхиз и Донна К. Харман. *TREC: Эксперимент и оценка в информационном поиске* . Массачусетский технологический институт, 2005.

Ян Восецки, Кеннет Вай-Тинг Люнг и Уилфред Нг. Совместный персонализированный поиск в Твиттере с тематически-языковыми моделями. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 2014 г.

Иван Вулич и Мари-Франсин Моэнс. Обнаружение высокоточных переводов слов из сопоставимых корпусов без каких-либо предварительных знаний. В *материалах Европейского отделения Ассоциации компьютерной лингвистики* ,2012.

Иван Вулич и Мари-Франсин Моэнс. Вероятностные модели межъязыкового семантического сходства в контексте, основанные на скрытых межъязыковых понятиях, полученных на основе сопоставимых данных. В *Трудах эмпирических методов обработки естественного языка* , 2014.

Иван Вулич, Вим Де Смет и Мари-Франсин Моэнс. Поиск информации на разных языках с использованием скрытых тематических моделей, обученных на сопоставимом корпусе. В *азиатских информационно-поисковых обществах* , 2011a.

|  |  |
| --- | --- |
| *использованная литература* | 149 |

Иван Вулич, Вим Де Смет и Мари-Франсин Моэнс. Идентификация переводов слов из сопоставимых корпусов с использованием латентных тематических моделей. В *материалах Ассоциации компьютерной лингвистики* , 2011b.

Иван Вулич, Вим Де Смет и Мари-Франсин Моэнс. Межъязыковые модели поиска информации, основанные на моделях скрытых тем, обученных с помощью сопоставимых корпусов, выровненных по документам. *Информационный поиск* , 16(3):331–368, 2013.

Иван Вулич, Вим Де Смет, Джи Тан и Мари-Франсин Моэнс. Вероятностное тематическое моделирование в многоязычных условиях: обзор его методологии и приложений. *Обработка информации и управление* , 51(1), 2015.

Ханна Уоллах, Дэвид Мимно и Эндрю МакКаллум. Переосмысление LDA: почему приоритеты имеют значение. В *материалах достижений в области нейронных систем обработки информации* , 2009a.

Ханна М. Уоллах, Иэн Мюррей, Руслан Салахутдинов и Дэвид Мимно. Методы оценки тематических моделей. В *материалах Международной конференции по машинному обучению* , 2009b.

Чонг Ван, Дэвид М. Блей и Дэвид Хекерман. Динамические тематические модели с непрерывным временем. В *материалах о неопределенности в искусственном интеллекте* , 2008 г.

Цюань Ван, Цзюнь Сюй, Ханг Ли и Ник Красуэлл. Регулярная скрытая семантическая индексация: новый подход к крупномасштабному тематическому моделированию. *ACM Transactions on Information Systems* , 31(1):5:1–5:44, январь 2013 г.

Шилианг Ван, Дж. Майкл Пол и Марк Дредзе. Социальные сети как датчик качества воздуха и реакции общественности в Китае. *Журнал медицинских интернет-исследований* , 17(3):e22, март 2015 г.

Син Ван, Дейи Сюн, Мин Чжан, Ю Хун и Цзяньминь Яо. Тематическая модель переупорядочивания для статистического машинного перевода. *Обработка естественного языка и китайские вычисления* , 496:414–421, 2014.

Син Вэй. *Тематические модели в информационном поиске* . Кандидат наук. диссертация, Массачусетский университет в Амхерсте, 2007 г.

Син Вэй и В. Брюс Крофт. Модели документов на основе LDA для оперативного поиска. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 2006 г.

Цзяньшу Вэн, И-Пэн Лим, Цзин Цзян и Ци Хэ. TwitterRank: Поиск влиятельных пользователей Твиттера, чувствительных к теме. В *материалах Международной конференции ACM по поиску в Интернете и интеллектуальному анализу данных* , 2010 г.

Тереза Уилсон и Дженис Вибе. Аннотирование атрибуций и частных состояний. В *CorpusAnno '05: Материалы семинара по границам в аннотациях корпуса II* , 2005 г.

150 *ссылок*

Фрэнк Вуд и Йи Уай Тех. Иерархический непараметрический байесовский подход к адаптации предметной области статистической языковой модели. В *материалах Международной конференции по искусственному интеллекту и статистике* ,том 12, 2009.

Синьян Сяо, Дейи Сюн, Мин Чжан, Цюнь Лю и Шоусунь Линь. Модель подобия тем для иерархического перевода на основе фраз. В *материалах Ассоциации компьютерной лингвистики* , 2012 г.

Дейи Сюн и Мин Чжан. Тематическая модель согласованности для статистического машинного перевода. В *материалах Ассоциации по развитию искусственного интеллекта* , 2013 г.

Дейи Сюн, Цюнь Лю и Шоусунь Линь. Модель переупорядочивания фраз на основе максимальной энтропии для статистического машинного перевода. В *Трудах Ассоциации компьютерной лингвистики* , 2006 г.

Цзе-И Ян, Эндрю Дж. Торгет и Рада Михалча. Моделирование темы на исторических газетах. На *семинаре ACL-HLT по языковым технологиям для культурного наследия, социальных и гуманитарных наук* , 2011 г.

Лимин Яо, Дэвид Мимно и Эндрю МакКаллум. Эффективные методы вывода тематической модели для потоковых коллекций документов. В *книге «Обнаружение знаний и интеллектуальный анализ данных* », 2009 г.

Син И и Джеймс Аллан. Сравнительное исследование использования тематических моделей для поиска информации. В *материалах Европейской конференции по информационному поиску* , том 5478, 2009 г.

Чжицзюнь Инь, Лянлян Цао, Цзявэй Хань, Чэнсян Чжай и Томас Хуан. Открытие и сравнение географических тем. В *материалах конференции World Wide Web* , 2011 г.

Хэн Юй, Цзиньсон Су, Яджуан Лв и Цюнь Лю. Тематическая языковая модель для статистического машинного перевода. На *Международной совместной конференции по обработке естественного языка* , 2013 г.

Цзя Цзэн, В. К. Чеунг и Джиминг Лю. Изучение тематических моделей путем распространения убеждений. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 35(5):1121–1134, 2013.

Цин Т. Цзэн, Дуг Редд, Томас К. Риндфлеш и Джонатан Р. Небекер. Синоним, тематическая модель и расширение запроса на основе предикатов для извлечения клинических документов. На *ежегодном симпозиуме Американской ассоциации медицинской информатики* , 2012 г.

|  |  |
| --- | --- |
| *использованная литература* | 151 |

Чэн Сян Чжай и Джон Лафферти. Изучение методов сглаживания языковых моделей применительно к поиску информации. В *материалах конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска* , 2001a.

Чэн Сян Чжай и Джон Лафферти. Обратная связь на основе модели в подходе языкового моделирования к поиску информации. В *материалах Международной конференции ACM по управлению информацией и знаниями* , 2001b.

Чэн Сян Чжай, Атулья Веливелли и Бэй Ю. Модель смешанной коллекции для сравнительного анализа текстов. В *книге «Обнаружение знаний и интеллектуальный анализ данных* », 2004 г.

Ке Чжай, Джордан Бойд-Грабер, Нима Асади и Мохамад Альхуджа. Г-н LDA: Гибкий пакет крупномасштабного тематического моделирования с использованием вариационного вывода в mapreduce. В *материалах конференции World Wide Web* , 2012 г.

Дуо Чжан, Цяочжу Мэй и Чэн Сян Чжай. Межъязыковое извлечение латентной темы. В *материалах Ассоциации компьютерной лингвистики* , 2010 г.

Бин Чжао и Эрик П. Син. BiTAM: двуязычные модели смешения тем для выравнивания слов. В *Трудах Ассоциации компьютерной лингвистики* , 2006 г.

Уэйн Синь Чжао, Цзин Цзян, Хунфэй Янь и Сяомин Ли. Совместное моделирование аспектов и мнений с помощью гибрида MaxEnt-LDA. В *Трудах эмпирических методов обработки естественного языка* , 2010.

Уэйн Синь Чжао, Цзин Цзян, Цзяньшу Вэн, Цзин Хэ, И-Пэн Лим, Хунфэй Янь и Сяомин Ли. Сравнение Twitter и традиционных СМИ с использованием тематических моделей. В *материалах Европейской конференции по информационному поиску* , 2011 г.

Дин Чжоу, Сян Цзи, Хунъюань Чжа и К. Ли Джайлз. Эволюция темы и социальные взаимодействия: как авторы влияют на исследования. В *материалах Международной конференции ACM по управлению информацией и знаниями* , 2006 г.

Джун Чжу, Амр Ахмед и Эрик П. Син. MedLDA: контролируемые тематические модели с максимальной маржой для регрессии и классификации. В *материалах Международной конференции по машинному обучению* , 2009 г.

**Индекс**

*k* - означает кластеризацию, 50

Джейстор, 55 лет

Американская революция, 50 лет

рейтинг одобрения, 81

модель аспекта, 86

Двуязычная местная смесь, 110

Китайский язык, 53, 95

внимательное чтение, 68

сопоставимые корпуса, 94

динамическая тема непрерывного времени

модель, 64

подражатель модели, 65

декодирование (машинный перевод), 99 глубокое обучение, 116, 128 дневник, 55

Диккенс, 72 года

словарь, 95, 97

Распределение Дирихле, 10

альтернативы, 64

сопряжение, 115

параметр, 10

роль в лда, 18

сглаживание в языковых моделях, 25

дерево, 96

неоднозначность, 104

дискретное распределение, 9, 64, 115

дистанционное чтение, 69

распределенное представительство, 128

моделирование языка документа, 23

модель динамического воздействия, 66

динамическая тематическая модель, 63

максимизация ожидания, 117

Фейсбук, 83

полностью факторизованное распределение, 118

Распределение Гаусса, 9, 63, 85, 115

Немецкий язык, 53, 55, 94 Выборка Гиббса, 14, 121

для лда, 15

152

*Индекс*  153

обоснованная теория, 83 романа, 69

скрытая тема Марковская модель, 108

Инфер.Нет, 117

грипп, 81

интерактивный Модель TOpic и метаданные (том), 44

интерактивность

информационный поиск, 32

тематические модели, 46

интерпретируемость, 45, 82, 116, 124, 128

Японский язык, 53

с надписью лда, 42

языковая модель

машинный перевод, 105

расширение запроса, 28

латентное распределение Дирихле, 7, 11 модель языка документа, 25 порождающий процесс, 13 реализации, 18

латентный семантический анализ, 7, 50, 107 оценка правдоподобия, 45, 124 ссылка, латентное распределение Дирихле, 89

Маллет, 19

Марта Баллард, 54 года, блочная модель со смешанным членством, 88 лет Ассоциация современного языка, 57 лет

имена вымышленных персонажей, 71 Национальный институт здоровья, 61 вложенный процесс Дирихле, 87 газета, 49

латентное онлайн-распределение Дирихле, 127

PageRank, 41

параллельный корпус, 97

фраза (машинный перевод), 101 табличка, 17

персонализированный поиск, 35 стихов, 76

загрязнение в Китае, 81 многоязычное латентное распределение Дирихле-

ция, 93

многоязычный Скрытое распределение Дирихле на основе дерева, 97

апостериорные прогностические проверки, 112 прогнозов и интерпретаций, 81 вероятностный латентный семантический анализ.

исис, 7, 25, 50

расширение запроса, 27, 94

модель релевантности, 29

изменение порядка (машинный перевод), 109

персонализация поиска, 34 анализа тональности, 83 сглаживания, 24 спектрального обучения, 128

Стэн, 117 лет

статистический машинный перевод, 91 доменная адаптация, 98

стохастическая блочная модель, 88 стилометрия, 77

154 *Индекс*

контролируемое латентное распределение Дирихле-

ция, 85, 123

опрос, 68, 80

синтетические данные, 122

Термит, 44

Теано, 117 лет

связность темы, 46

обнаружение и отслеживание темы, 7

маркировка темы, 40

Механизм визуализации тематической модели,

44

тематический гид, 44

Факел, 117

латентное распределение Дирихле на основе дерева

ция, 95

Твиттер, 90

восходящие и нисходящие модели,

85

вариационный вывод, 117

Википедия, 41, 92

Ворднет, 95