Урок 7

Тематическое моделирование

Этот модуль будет посвящен вероятностному тематическому моделированию — направление, которое развивается в машинном обучении очень активно последние 15 лет. Постоянно появляются новые постановки задач, новые прикладные задачи и новые методы.

7.1. Введение в тематическое моделирование

7.1.1. Понятие темы в тематическом моделировании

Тематическое моделирование — современное направление статистического анализа текстов. Методы тематического моделирования призваны ответить на вопрос, какой теме посвящена большая коллекция текстовых документов.

Интуитивно под темой понимается специальная терминология определенной предметной области. Более точно, **тема** — набор **терминов**, то есть слов или словосочетаний, которые совместно часто встречаются в документах. Интуитивное понимание исходит из того факта, что документ по математике легко можно отличить от документа по биологии за счет узнаваемых терминов даже без знания этих предметных областей. Более того, если термины в документах будут случайным образом переставлены, все равно можно будет понять тему этого документа.

Более формально эти соображения можно изложить на языке теории вероятностей и статистики. В этом случае речь пойдет о частотах встречаемости слов во коллекции документов.

Каждая тема характеризуется своим словарем и своими вероятностями терминов из этого словоря, $p(w|t)$ — вероятность (частота) термина w в теме t. Так как документ может относиться одновременно к нескольким темам (как говорят, к вероятностной смеси тем), существует понятие тематики документа. **Тематика документа** — условное распределение $p(t|d)$, где t — тема, а d — документ.

Тематическая модель должна автоматически выявлять латентные темы по наблюдаемым частотам терминов в документе $p(w|d)$. Поскольку неизвестно, о какой теме думал автор, когда писал документ, тема документа не наблюдается непосредственно и по этой причине называется латентной.

7.1.2. Цели тематического моделирования

Автоматический анализ текста

Тематическое моделирование может, в том числе, использоваться для автоматического анализа текстов для решения целого множества задач:

- **Классификация и категоризация документов**: необходимо, опираясь на тематику текстов, «разложить по папкам» большую коллекцию или поток документов. Тематическое моделирование обладает особенностью, что позволяет относить документ сразу ко многим темам.
- **Автоматическое аннотирование документов:** необходимо выделить в документе наиболее важные фразы и составить на их основе его краткий обзор. Тематическое моделирование позволяет, определив тематику документа, выделить наиболее «тематичные» фразы и сделать из них выборку таким образом, чтобы были покрыты все имеющиеся в документе темы.
- **Автоматическая суммаризация коллекций** очень похожа на предыдущую задачу, но в этом случае необходимо подготовить обзор не одного документа, а большой коллекции или какой-то ее части.

- **Тематическая сегментация документов:** необходимо разбить длинный документ на тематически однородные фрагменты и определить тематику каждого такого фрагмента.

Общая идея решения всех представленных задач состоит в том, что распределение тем $p(t|d)$ в документе становится признаковым описанием документа d. Часто при анализе текстов используются векторные модели, в которых каждый частота каждого слова или термина дает свой признак. Число признаков в таких моделях очень велико и совпадает с числом терминов в словаре. Тематическое моделирование позволяет перейти к сжатому признаковому описанию, в котором каждый признак соответствует одной теме.

Систематизация больших объёмов информации

Другой большой класс задач составляют задачи, связанные с новыми методами поиска текстовой информации и помощи в ее понимании. Это особенно необходимо людям, которые постоянно занимаются самообразованием, чтобы быстро находить нужные знания в новых для себя областях. Такой тип поиска называется семантическим или разведочным (название еще не устоялось). Для решения такого рода задач необходимо структурировать общие представления об устройстве предметной области. В частности определить, как тема делится на подтемы.

Тематическое моделирование для цели систематизации больших объемов информации позволяет решать следующие задачи:

- **Семантический (разведочный) поиск информации**
- **Визуализация тематической структуры коллекции**
- **Анализ динамики развития тем**, особенно, если к каждому документу привязана временная метка.
- **Тематический мониторинг новых поступлений**, который предоставляет возможность автоматически сообщать пользователям о появлении в сети Интернет или какой-то библиотеке новых тематических документов без необходимости постоянно производить повторный поиск.
- **Рекомендация документов пользователям**, то есть когда необходимо построить рекомендательную систему, которая использует данные о прошлой активности пользователя и данные об активности других пользователей.

7.1.3. Приложения тематического моделирования

Можно кратко перечислить основные приложения тематического моделирования:

- **Поиск научной информации, трендов, фронта исследований**. По разным оценкам множество статей и публикаций ежегодно увеличивается примерно на 10 миллионов документов.
- **Подбор экспертов, рецензентов, исполнителей проектов**. Для автоматизации деятельности экспертных советов необходимо анализировать огромное количество заявок на инновационные проекты и научные гранты, чтобы быстро подобрать экспертов и распределить пришедшую коллекцию документов на экспертизу.
- **Агрегирование новостных потоков**: необходимо в приходящем из разных источников новостном потоке определять тематику каждого документа, находить дубликаты и отслеживать каждую тему во времени. С этой задачей сталкиваются многие компании, которые занимаются социологическими исследованиями, аналитические компании, информационные агентства и так далее. Тематическое моделирование позволяет автоматизировать эту деятельность так, чтобы с этой задачей справлялось меньшее число сотрудников.
- **Аннотирование и поиск изображений** — задача, не связанная с текстовой аналитикой, но при решении которой применять те же методы для совершенно другого типа данных. Действительно, если существуют методы выделения тех или иных элементов изображений, то такие элементы можно рассматривать в качестве терминов, а сами изображения — в качестве документов.
  Более того, некоторые изображения сопровождаются текстом. Появляется возможность сразу строить два представления документа: основанное на выделении графических элементов и основанное на анализе сопровождающего текста. Это позволяет для такого рода многомодальных коллекций реализовывать поиск текстов по изображениям, изображения по тексту, а также аннотировать изображения и так далее.
- **Анализ видеопоследовательностей**, **аннотация генома и другие задачи биоинформатики**, **анализ дискретизированных биомедицинских сигналов**, **мониторинг состояния технических**

систем — еще несколько примеров задач, в которых все чаще используются методы тематического моделирования. В некоторых задачах сигнал был исходно дискретным, в некоторых — изначально вещественным. Во втором случае его необходимо дискретизировать, чтобы после дискретизации была возможность представить его в виде символьной последовательности и использовать методов символьной лингвистики, машинного обучения и тематического моделирования для ее анализа.

7.1.4. Примеры использования тематического моделирования

Мультиязычная модель Википедии

При тематизации википедии была построена мультиязычная модель (на русском и английском языка), которая смогла сама, без помощи эксперта, собрать темы на обоих языках. Было обработано 216 175 русско-английских пар статей Википедии и собрано 400 тем.

тема 68				**тема 79**			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14
тема 88				**тема 251**			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Таблица 7.1: Первые 10 слов (с их вероятностями $p(w|t)$ в %) в каждой из представленных тем.

Эти темы оказались легко интерпретируемыми. Более того, модель выявляет двуязычные темы без выравнивания, без словарей, даже когда тексты не являются точными переводами. В этом эксперименте независимый эксперт оценил 396 тем из 400 как хорошо интерпретируемые.

Биграммная модель термины – словосочетания

Оказывается, что интерпретируемость тем, понятность для человека резко возрастает, если в качестве терминов использовать не отдельные слова, а словосочетания. Например, была построена тематическая модель для коллекции научных статей. Это статьи конференций «Математические методы распознавания образов» и «Интеллектуализация обработки информации» на русском языке.

распознавание образов в биоинформатике	
unigrams	bigrams
объект	задача распознавания
задача	множество мотивов
множество	система масок
мотив	вторичная структура
разрешимость	структура белка
выборка	распознавание вторичной
маска	состояние объекта
распознавание	обучающая выборка
информативность	оценка информативности
состояние	множество объектов
закономерность	разрешимость задачи
система	критерий разрешимости
структура	информативность мотива
значение	первичная структура
регулярность	тупиковое множество

теория вычислительной сложности	
unigrams	bigrams
задача	разделять множества
множество	конечное множество
подмножество	условие задачи
условие	задача о покрытии
класс	покрытие множества
решение	сильный смысл
конечный	разделяющий комитет
число	минимальный аффинный
аффинный	аффинный комитет
случай	аффинный разделяющий
покрытие	общее положение
общий	множество точек
пространство	случай задачи
схема	общий случай
комитет	задача MASC

Таблица 7.2: Несколько тем, построенные по 850 статьям конференций ММРО, ИОИ на русском языке.

Модель, в которой в качестве терминов используются отдельные слова, легко интерпретируема, но понять авторов и научную школу по списку терминов практически невозможно. Другое дело в случае биграммной модели, где словарь состоит из пар слов: разбираясь в тематике конференции, можно даже соотнести конкретных авторов с некоторой темой.

Поиск этно-релевантных тем в социальных сетях

В целях поддержка социологических исследований в области межэтнических отношений была поставлена задача создать систему разведочного поиска для систематизации и мониторинга этно-релевантных тем.

Заранее специалистами был создан словарь этнонимов (более 800 слов), который потом использовался, чтобы выделить в социальной сети контент, который относится к обсуждению подобного рода тем (его оказалось не более 1%). После этого, уже без участия экспертов, в выделенном контенте были выделены этнорелевантные темы. Например:

русские: русский, князь, россия, татарин, великий, царить, царь, иван, император, империя, грозить, государь, екатерина, акция, организация, митинг, движение, активный, мероприятие, пикет, русский, участник, москва, оппозиция.

сирийцы: сирийский, асад, боевик, район, террорист, уничтожать, группировка, дамаск, оружие, алесио, оппозиция

таджики и узбеки: мигрант, страна, россия, миграция, азия, нелегальный, миграционный, таджикистан, гастарбайтер

канадцы: команда, игра, игрок, канадский, сезон, хоккей, сборная, играть, болельщик, победа, кубок, счет, чемпионат

норвежцы: дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, опека, норвежский, сын

китайцы: китайский, россия, производство, китай, страна, продукция, предприятие, компания, технология, военный

По этим данным можно понять, что, если обсуждаются русские, то тема либо связана с историей России, либо с текущей политикой. Сирийцев обсуждают в связи с войной в Сирии, таджиков и узбеков — в связи с трудовой миграцией. Разговоры про канадцев в нашей социальной сети идут про то, как они играют в хоккей. Когда говорят про норвежцев, обсуждают чаще всего проблематику ювенальной юстиции, а когда говорят про китайцев, то производство и деловые отношения.

7.2. Постановка задачи тематического моделирования

В этом видео речь пойдет о формальной постановке задачи тематического моделирования и о вероятностной порождающей модели текста.

7.2.1. Подготовка данных для тематического моделирования

Прежде чем применять формальные методы, тексты необходимо подготовить. Предварительная обработка и очистка текстов производится по-разному для разных типов текста и может включать в себя:

- **Удаление форматирования и переносов**
- **Удаление обрывочной и нетекстовой информации**. Например, если коллекция была получена из PDF файлов, в файлах могут содержаться артефакты: разрывы слов на месте перевода строки, колонтитулы, остатки графиков, таблиц, а также различные нечитаемые символы и так далее. Все это должно быть удалено из исходных текстов.
- **Исправление опечаток**
- **Слияние слишком коротких текстов**. Тексты из социальных сетей, например сообщения из twitter, часто очень короткие, поэтому необходимо применять разные методики для того, чтобы сливать короткие тексты в один более длинный и более удобный для тематического моделирования. Как именно это делается, зависит от цели исследования. В частности, можно слить в один все тексты автора за некоторый день. Или же объединять в один вообще все тексты данного автора.

После того, как это сделано, текст представляет собой последовательность слов без лишнего. После это необходимо выполнить формирование словаря:

- **Выделение терминов (term extraction)**. В научных текстах существует хорошо устоявшаяся терминология, причем чаще всего термины представляют собой словосочетания, а не отдельные слова. Поэтому важно сделать выделение терминов, используя специальные методики. Некоторые из них, так называемые методы предобработки, выполняются отдельно от тематического моделирования. Другие можно прямо встроить в тематические модели, но о таких сложных моделях речь идти не будет.
- **Приведение слов к нормальной форме** позволяет решить проблему со словоформами и может быть выполнено используя **стемминг** (лучше подходит для английских текстов) или **лемматизацию** (лучше для русских текстов).
- **Удаление стоп-слов и слишком редких слов**. Слова, которые встречаются слишком часто, не помогают определить тематику текста и называются **стоп-словами**. Также не помогут слишком редкие слова. Действительно, тема — это некое статистическое явление, набор слов, которые совместно часто встречаются в определенных текстах. Если слово встречается очень редко, как правило, ни о какой статистике речи идти не может. Поэтому слова, которые встретились реже примерно десяти раз, просто удаляются из коллекции. Удаление стоп-слов и слишком редких слов позволяет сократить словарь и уменьшить вычислительную сложность задачи.

7.2.2. Базовые предположения простых тематических моделей

Базовые предположения простых тематических моделей включают в себя следующие:

- **Порядок документов в коллекции не важен**
- **Порядок терминов в документе не важен (bag of words)**. Каждый текст предполагается последовательностью слов, которые генерируются из некоторого вероятностного распределения. В самых простых моделях используется предположение о «мешке слов», согласно которому, даже переставив в документе слова или выделенные словосочетания, можно определить его тематику.
- **Употребление каждого слова в каждом документе связано с некоторой темой**, то есть каждая пара (d, w) связана с некоторой темой $t \in T$. Следовательно, коллекция документов представляет собой последовательность троек (d, w, t), в которой темы являются латентными: они не видны и для их определения как раз используется тематическая модель.

- **Гипотеза условной независимости:** $p(w|t,d) = p(w|t)$ заключается в том, что вероятность слова документах определяется только темой, а не самим документом. Это предположение позволяет строить легко оцениваемые тематические модели.

Часто используются дополнительные предположения разреженности:

- Предположение, что документ относится к небольшому числу тем.
- Предположение, что тема состоит из небольшого числа терминов, лексического ядра, которое существенно отличает эту тему от остальных.

7.2.3. Вероятностный процесс порождения текстовой коллекции

В вероятностной порождающей модели документ d — это смесь распределений $p(w|t)$ с весами $p(t|d)$:

$$p(w|d) = \sum_{t \in T} p(w|t)\, p(t|d).$$

Условное распределение тем в документе $p(t|d)$ — важный параметр модели, который и необходимо оценивать.

Таким образом, процесс порождения текста следующий. Для каждой словопозиции w сначала из распределения тем в документе выбирается тема, к которой это слово будет относиться. После этого из распределения слов в выбранной теме выбирается конкретное слово, которое будет записано в данную словопозицию. Слово за словом так появляется весь текст.

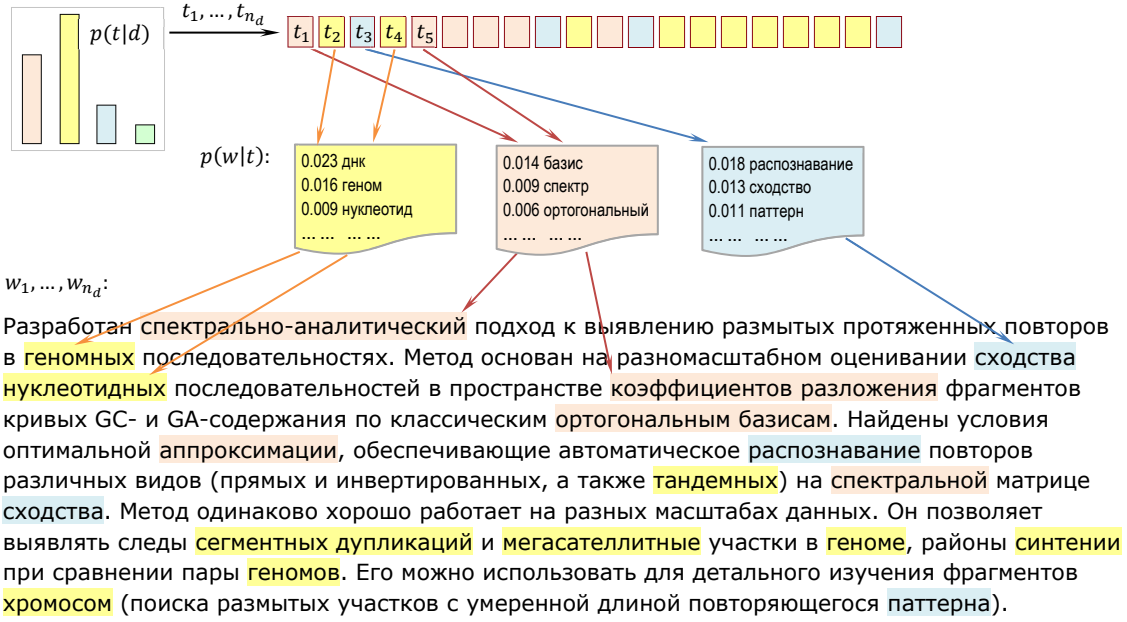
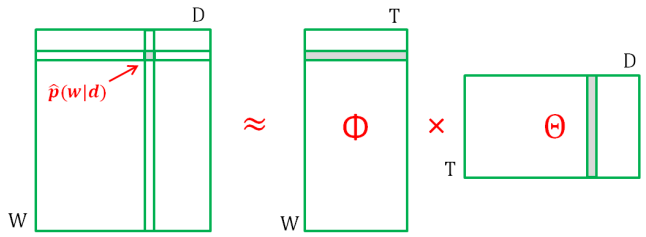


Рис. 7.1: Процесс порождения текстового документа вероятностной тематической моделью

Поскольку выполняется гипотеза «мешка слов» сгенерированный текст вряд ли будет осмысленным. Можно только говорить о том, что с точностью до произвольной перестановки слов, этот текст вполне мог бы нести в себе какую-то тематику. А именно тематику текста и нужно выявить. Другими словами, тематическое моделирование не обеспечивает понимание компьютером смысла текста, а только позволяет выполнить кластеризацию документов по темам.

7.2.4. Постановка задачи тематического моделирования

Формальная постановка задачи тематического моделирования следующая. Пусть зафиксирован словарь терминов W, из элементов которого складываются документы, и дана коллекция D документов $d \subset W$. Для каждого документа d известна его длина n_d и количество n_{dw} использований каждого термина w.



Требуется найти параметры вероятностной порождающей тематической модели, то есть представить вероятность появления $p(w|d)$ слов в документе в виде:

$$p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td},$$

где $\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t, $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d.

Порождающая модель описывает процесс построения коллекции по ϕ_{wt} и θ_{td}. Тематическое моделирование представляет собой обратную задачу: по наблюдаемой коллекции необходимо понять, какими распределениями ϕ_{wt} и θ_{td} она могла бы быть получена.

7.2.5. Задача тематического моделирования как задача матричного разложения

Фактически, эту задачу можно трактовать как задачу матричного разложения. Пусть Φ — матрица распределений терминов в темах, а Θ — матрица распределений тем в документах:

$$\Phi = \big(\phi_{wt}\big), \qquad \Theta = \big(\theta_{td}\big).$$

Матрицы называются стохастическими, если каждый их столбец преставляет собой дискретное распределение вероятностей, а ,следовательно, сумма значений по каждому столбцу равна 1 (условие нормировки) и каждое значение является неотрицательным (условие неотрицательности). Следует особо отметить, что стохастические матрицы — это НЕ такие матрицы, элементы которых генерируются случайно. Обе определенные выше матрцы Φ и Θ — стохастические. Согласно вероятностной тематической модели, произведение матриц Φ и Θ должно давать частотные оценки $p(w|d)$ условных вероятностей слов в документах коллекции. Наблюдаемые частоты терминов в документах известны:

$$\hat{p}(w|d) = \frac{n_{dw}}{n_d}.$$

Задача тематического моделирования, таким образом, стала задачей стохастического матричного разложения матрицы $\big(\hat{p}(w|d)\big)$ на стохастические матрицы Φ и Θ.

Теперь можно воспользоваться приниципом максимума правдоподобия с ограничениями, следующими из условий нормировки и неотрицательности на элементы стохастических матриц. Если максимизировать логарифм правдоподобия, получается:

$$\begin{cases} \sum\limits_{d \in D} \sum\limits_{w \in d} n_{dw} \ln \sum\limits_{t \in T} \phi_{wt} \theta_{td} \to \max\limits_{\Phi, \Theta}; \\ \sum\limits_{w \in W} \phi_{wt} = 1; \qquad \phi_{wt} \geq 0; \\ \sum\limits_{t \in T} \theta_{td} = 1; \qquad \theta_{td} \geq 0. \end{cases}$$

7.2.6. Принцип максимума регуляризованного правдоподобия

Задача матричного разложения некорректно поставлена, поскольку её решение в общем случае не единственно:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

С одной стороны, строящаяся математическая модель получается неустойчивой и невоспроизводимой (результат работы итерационных методов будет зависет от начального приближения), но, с другой стороны, это

дает свободу выбора дополнительных ограничений на матрицы Φ и Θ. В теории некорректно поставленных задач, то есть решение которых не единственно, такие ограничения принято называть регуляризаторами.

Чтобы из множества решений выбрать наиболее подходящее, вводится **критерий регуляризации** $R(\Phi, \Theta)$, который представляет собой некоторый функционал, построенный исходя из тех или иных содержательных соображений данной задачи. Теперь вместо задачи максимизации логарифма правдоподобия рассматривается задача максимизации регуляризованного правдоподобия, то есть суммы логарифма правдоподобия и регуляризатора $R(\Phi, \Theta)$:

$$\begin{cases} \sum\limits_{d\in D}\sum\limits_{w\in d} n_{dw} \ln \sum\limits_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max\limits_{\Phi,\Theta}; \\ \sum\limits_{w\in W} \phi_{wt} = 1; \qquad \phi_{wt} \geq 0; \\ \sum\limits_{t\in T} \theta_{td} = 1; \qquad \theta_{td} \geq 0. \end{cases}$$

7.3. Базовые тематические модели и EM-алгоритм

7.3.1. Регуляризованный EM-алгоритм

Выше было получено, что для построения тематической модели $p(w|d) = \sum_{t\in T} \phi_{wt}\theta_{td}$ по наблюдаемым частотам $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$ нужно решать задачу максимизации логарифма правдоподобия с регуляризатором $R(\Phi, \Theta)$:

$$\begin{cases} \sum\limits_{d\in D}\sum\limits_{w\in d} n_{dw} \ln \sum\limits_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max\limits_{\Phi,\Theta}; \\ \sum\limits_{w\in W} \phi_{wt} = 1; \qquad \phi_{wt} \geq 0; \\ \sum\limits_{t\in T} \theta_{td} = 1; \qquad \theta_{td} \geq 0. \end{cases}$$

Регуляризатор может быть произвольным, однако удобно выбрать его гладким, чтобы можно было применить условие Каруша — Куна — Таккера. В результате получается система уравнений относительно относительно ϕ_{wt}, θ_{td} и вспомогательных переменных $p_{tdw} = p(t|d,w)$ (выкладки не приводятся):

$$\begin{cases} p_{tdw} = \underset{t\in T}{\mathsf{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \phi_{wt} = \underset{w\in W}{\mathsf{norm}}\left(\sum\limits_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{t\in T}{\mathsf{norm}}\left(\sum\limits_{w\in d} n_{dw}p_{tdw} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

где операция нормировки вектора определена следующим образом:

$$\underset{t\in T}{\mathsf{norm}}(x_t) = \frac{\max\{x_t, 0\}}{\sum\limits_{s\in T}\max\{x_s, 0\}}.$$

Полученную систему удобно решать методом простых итерации, который в данном случае по совместительству оказывается EM-алгоритмом. Первое уравнение системы (**E-шаг**):

$$p_{tdw} = \underset{t\in T}{\mathsf{norm}}\left(\phi_{wt}\theta_{td}\right)$$

представляет собой вычисление с помощью формулы Байеса вспомогательных переменных $p_{tdw} = p(t|d,w)$ через параметры модели ϕ_{wt} и θ_{td}. Оставшиеся уравнения (**M-шаг**):

$$\phi_{wt} = \underset{w\in W}{\mathsf{norm}}\left(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \qquad \theta_{td} = \underset{t\in T}{\mathsf{norm}}\left(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$$

наоборот, выражают основные параметры модели через вспомогательные переменные. В случае отсутствия регуляризатора ($R(\Phi, \Theta) = 0$) эти формулы представляют собой частотные оценки условных вероятностей слов в темах и тем в документах.

ЕМ-алгоритм заключается в последовательном выполнении **E-шага** и **M-шага** до достижения требуемой точности, то есть является итерационным процессом.

Следует обратить особое внимание на то, что оператор нормировки переводи произвольный вектор x_t в вектор, координаты которого неотрицательны и нормированы, то есть сумма которых в точности равна 1. Другими словами, дискретное распределение можно задавать с помощью любого вектора нужной размерности, если сперва отнормировать его таким образом.

Два самых известных частных случая этой системы уравнений (при разном выборе регуляризатора):

- **PLSA, вероятностный латентный семантический анализ**. В этом случае регуляризатор не используется, то есть:
$$R(\Phi, \Theta) = 0.$$
Этот метод был придуман и опубликован Томасом Хофманом в 1999 году.

- **LDA, латентное размещение Дирихле**:
$$R(\Phi, \Theta) = \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}$$
где $\beta_w > 0$, $\alpha_t > 0$ — параметры регуляризатора. Этот метод был предложен четыре года спустя Дэвидом Блеем, Эндрю Энджи и Майклом Джорданом. Их статья по латентному размещению Дирихле — это, наверное, самая цитируемая работа в тематическом моделировании.

7.3.2. Байесовская интерпретация модели LDA

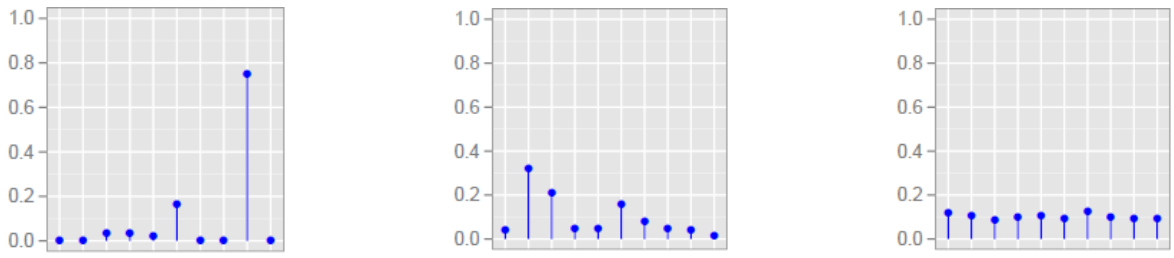
Авторы метода латентного размещения Дирихле рассматривали его с точки зрения байесовского обучения, то есть вычисляли апостериорные оценки параметров, а не использовали принцип максимизации правдоподобия. Использовалось предположение, что столбцы ϕ_t матрицы Φ порождаются $|W|$-мерным распределением Дирихле с вектором параметров $\beta = (\beta_w)$. (Фактически, имеет место двухступенчатая порождающая модель текста: сначала из распределения Дирихле порождаются столбцы матриц Φ и Θ, а потом на их основе порождается текстовая коллекция.)

Распределение Дирихле нужной размерности порождает нормированные и неотрицательные векторы, которые могут использоваться в качестве вероятностных распределений. Распределение Дирихле удобно использовать в качестве априорных распределений в байесовской интерпретации, поскольку распределение Дирихле позволяет генерировать как разреженные (при малых значениях параметров), так и такие вот сильно сглаженные дискретные распределения (при больших значениях параметров). Если все параметры распределения Дирихле в точности равны 1, получается равномерное распределение.

Апостериорные оценки доставляют максимум правдоподобия с использованием регуляризатора

$$R(\Phi, \Theta) = \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}.$$

Другими словами, вместо вычисления априорных оценок можно использовать принцип максимума правдоподобия с представленным регуляризатором. Именно в этой формулировке метод LDA был представлен выше по тексту.



$\beta_w = 0.1$ $\beta_w = 1$ (равномерное) $\beta_w = 100$

Рис. 7.2: Распределение $\phi \sim \mathsf{Dir}(\beta)$ при $|W| = 10$, $\phi, \beta \in \mathbb{R}^{10}$

7.3.3. Дивергенция Кульбака–Лейблера

Существует другой, более простой, взгляд на латентное размещение Дирихле. Прежде, чем перейти к нему, необходимо ввести понятие дивергенции Кульбака — Лейблера. Оно часто встречается в машинном обучении и теории вероятностей и является чем-то вроде расстояния между распределениями.

Пусть даны два дискретных распределения $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$, тогда дивергенция Кульбака — Лейблера

$$\mathsf{KL}(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i}.$$

(В случае непрерывных распределений вместо суммы должен стоять интеграл)

Дивергенция Кульбака — Лейблера обладает следующими основными свойствами:

1. Неотрицательность:
$$\mathsf{KL}(P\|Q) \geq 0; \quad \mathsf{KL}(P\|Q) = 0 \;\Leftrightarrow\; P = Q.$$

2. Несимметричность:
$$\mathsf{KL}(P\|Q) \neq \mathsf{KL}(Q\|P).$$
Дивергенция Кульбака — Лейблера не является симметричной, а следовательно, не совсем правильно ее называть функцией расстояния между распределениями. Более корректно говорить, что она меряет в некотором смысле степень вложенности распределения P в распределение Q.

3. Связь с принципом максимума правдоподобия:
$$\sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_\alpha \;\Leftrightarrow\; \sum_{i=1}^n p_i \ln q_i(\alpha) \to \max_\alpha.$$
То есть минимизация дивергенции Кульбака — Лейблера эквивалентна максимизации правдоподобия. Если P — эмпирическое распределение, а Q — это какая-то параметрическая модель распределения с параметром α, то чтобы определить такое значение α, при котором P как можно лучше соответствовало модели не имеет значения, минимизировать дивергенцию Кульбака — Лейблера или максимизировать правдоподобие.

Последнее свойство дает интуитивное понимание смысла дивергенции Кульбака — Лейблера. Далее она будет применена в роли регуляризатора.

7.3.4. Не-байесовская интерпретация модели LDA

Пусть $\beta = (\beta_w)$ — некоторый вектор над словарем W.

Если для некоторого слова $w \in W$ выполнено $\beta_w > 1$, то в модели LDA распределение вероятности ϕ_{wt} этого слова по темам темах будет сглаживаться, приближаясь к β_w^+:

$$\mathsf{KL}\big(\beta^+ \,\big\|\, \phi_t\big) \to \min, \qquad \beta_w^+ = \underset{w\in W}{\mathsf{norm}}(\beta_w - 1)$$

При $\beta_w < 1$, наоборот, значения ϕ_{wt} будут разреживаются, удаляясь от β_w^- к нулю:

$$\mathsf{KL}\big(\beta^- \,\big\|\, \phi_t\big) \to \max, \qquad \beta_w^- = \underset{w\in W}{\mathsf{norm}}(1 - \beta_w),$$

то есть среди ϕ_{wt} будет больше нулевых или почти нулевых элементов.

В отличие от байесовской точки зрения, такая интепретация не требует введения распределения Дирихле и априорных вероятностей. Здесь достаточно сказать, что столбцы матриц Φ и Θ либо приближаются к некоторому заданному распределению, либо удаляются от него, причем в одном из столбцов может происходить сглаживание, а в другом — разреживаение. Более того, так как априорные распределения Дирихле больше не используется, можно снять ограничения $\beta_w > 0,\ \alpha_t > 0$. Это позволяет сильнее разреживать матрицы Φ и Θ.

7.3.5. Онлайновый EM-алгоритм

В общем случае EM-алгоритм представляет собой систему уравнений, которая решается методом простых итераций, то есть по вспомогательным переменным $p_{tdw} = p(t|d,w)$ будут расчитываться основные параметры модели ϕ_{wt}, θ_{td}, а затем наоборот — по ϕ_{wt} и θ_{td} вычисляются значения p_{tdw}.

Можно, ничего не меняя в формулах, организовать вычислительный процесс метода простых итераций таким образом, чтобы он шел максимально быстро, причем именно на больших коллекциях текстовых документов. Дело в том, что матрицы Φ и Θ находятся в неравноправном положении: каждый столбец матрицы Φ относится ко всей коллекции целиком, а каждый столбец матрицы Θ — только к одному документу.

Поэтому, если процесс будет устроен таким образом, что матрицы Φ и Θ будут обновляться только после просмотра всей коллекции, матрица Φ будет обновляться слишком редко и весь процесс будет работать крайне медленно.

Другой подход заключается в обновлении матрицы Φ после просмотра очередного документа и простроения для него тематической модели. Можно также обрабатывать по несколько документов (по порциям, пакетам документов) и обновлять матрицу Φ после анализа очередного пакета документов. Оказалось, что такой способ организовать процесс является самым быстрым, и позволяет создавать онлайновые алгоритмы

Ввод: коллекция D, число тем $|T|$, параметры $i_{\max}$, $j_{\max}$, γ;
Вывод: матрицы терминов тем Θ и тем документов Φ;

для всех $i = 1, \ldots, i_{\max}$ (итерации по коллекции)
  для всех документов $d \in D$
    для всех $j = 1, \ldots, j_{\max}$ (итерации по документу)
      $p_{tdw} := \underset{t\in T}{\mathsf{norm}}\big(\phi_{wt}\theta_{td}\big)$
      $\theta_{td} := \underset{t\in T}{\mathsf{norm}}\Big(\sum\limits_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big)$
    $n_{wt} := \gamma n_{wt} + n_{dw}p_{tdw}$
  если пора обновить матрицу Φ **то**
    $\phi_{wt} := \underset{w\in W}{\mathsf{norm}}\Big(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big)$

Онлайновые алгоритмы появились не сразу, а примерно лет 5 назад, и стали эффективным инструментом анализа больших коллекций текстов. Оказалось, что матрица Φ может сходиться завершения просмотра всей коллекции: после просмотра нескольких первых десятков тысяч документов матрица Φ получается уже более-менее устоявшейся. После этого EM-алгоритм можно использовать, чтобы тематизировать остальные документы.

7.4. Регуляризация тематических моделей

7.4.1. Аддитивная регуляризация тематических моделей (ARTM)

Как было показано ранее, при решении задача тематического моделирования с помощью EM-алгоритма существует свобода выбора дополнительных критериев, которые называются регуляризаторами. Регуляризаторы бывают двух типов: для учёта дополнительных данных и для получения решения Φ, Θ с заданными свойствами.

На самом деле, запас неединственности решения основной задачи матричного разложения настолько большой, что можно на модель одновременно наложить несколько таких ограничений. Такой подход называется аддитивной регуляризацией и заключается в максимизации логарифма правдоподобия с k регуляризаторами R_i:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta},$$

где τ_i — коэффициенты регуляризации.

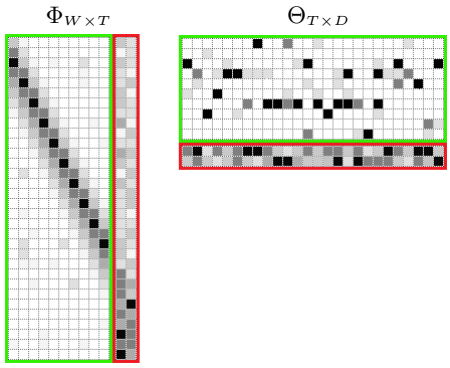


Рис. 7.3: Матрицы Φ и Θ

Такой подход дает возможность наложить сразу несколько условий, но также появляется проблема нахождения коэффициентов регуляризации. На данный момент, в основном, регуляризаторы добавляются по одному и у каждого регуляризатора оптимизируя этот коэффициент в ходе нескольких пробных запусков модели.

7.4.2. Разделение тем на предметные и фоновые

Продемонстрировать, как используя несколько регуляризаторов наделить модель нужными свойствами, можно на следующем примере. Наличие слов общей лексики в теме приводит к плохой интерпретируемости данной темы. Поэтому хотелось бы такие общеупотребительные слова выделить в отдельные темы, так называемые фоновые темы. Все остальные темы называются, соответственно, предметными, так как они описывают предметные области текстовой коллекции.

Предметные темы должны быть достаточно сильно разреженными, чтобы у каждой такой темы существовало свое лексическое ядро, существенно отличающее эту тему от остальных. Другими словами, требуется не только разреженность тем, но и их декоррелированность.

Эти требования можно выразить с помощью регуляризаторов. Пусть S — множество предметных тем, а B — множество фоновых. Поскольку для предметных тем ($t \in S$) матрицы $p(w|t)$ и $p(t|d)$ должны быть разреженными и существенно различными, а для фоновых ($t \in B$) — существенно отличными от нуля (больше половины слов в каждом документе — фоновые), имеет смысл применить регуляризатор, рассмотренный ранее в методе латентного размещения Дирихле. Единственное отличие состоит в том, что тогда он применялся для всего словаря, а в данном случае регуляризатор сглаживания необходимо применить только к фоновым темам:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in B}\sum_{w\in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d\in D}\sum_{t\in B} \alpha_t \ln \theta_{td} \to \max,$$

где β_0, α_0 — коэффициенты регуляризации. В этом случае распределения ϕ_{wt} будут близки к заданному распределению β_w, а распределения θ_{td} — к заданному распределению α_t. Распределения β_w и α_t вычисляются заранее. Например, в качестве β_w можно использовать распределение слов в используемом языке.

По аналогии можно построить разреживающий регуляризатор для предметных тем:

$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in S}\sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in S} \alpha_t \ln \theta_{td} \to \max\,.$$

где β_0, α_0 — коэффициенты регуляризации. В этом случае распределения ϕ_{wt} и θ_{td} будут как можно далеки от заданных распределений β_w и α_t. Определением параметров β_w и α_t занимается специалист, который занимается построением тематической модели. Часто в качестве β_w также используют распределение слов в используемом языке, а в качестве α_t — равномерное распределение.

7.4.3. Регуляризатор частичного обучения (semi-supervised learning)

Интересное обобщение этих двух регуляризаторов — сглаживающего и разреживающего — возникает в том случае, если векторы β_{wt} и α_{td} могут быть свои для каждого столбца:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T}\sum_{w\in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d\in D}\sum_{t\in T} \alpha_{td} \ln \theta_{td} \to \max,$$

Казалось бы, такое большое количество параметров дает лишнюю свободу. Но цель введения такого регуляризатора проста и может быть продемонстрирована с помощью следующего рассмотрения.

Для каждой темы можно задать свои подмножества типичных для этой темы слов и важных документов на эту тему. Такую информацию могут задавать эксперты. В таком случае говорят о semi-supervised learning, то есть о внесении частичной обучающей информации. Например, если в результате построения модели в ней остались некоторые ошибки (ключевое слово не попало в тему и так далее), то правки можно внести вручную, используя предложенный регуляризатор и следующие значения параметров:

- $\beta_{wt} = \big[w \in W_t\big]$ — белый список W_t терминов темы t
- $\alpha_{td} = \big[d \in D_t\big]$ — белый список D_t документов темы t
- $\beta_{wt} = -\big[w \in W_t\big]$ — чёрный список W_t терминов темы t
- $\alpha_{td} = -\big[d \in D_t\big]$ — чёрный список D_t документов темы t

7.4.4. Регуляризатор декоррелирования тем

Еще одно требование, которое уже было озвучено, состоит в том, чтобы в каждой теме выделялось свое лексическое ядро (множество терминов, отличающее её от других тем), то есть чтобы темы, как столбцы матрицы Φ, как можно меньше коррелировали друг с другом. Для этого вводится следующий ковариационный регуляризатор:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt}\phi_{ws} \to \max,$$

где τ — коэффициент регуляризации. Такой регуляризатор добавляет условие, чтобы попарно все столбцы матрицы Φ были далеки друг от друга.

Такой регуляризатор легко можно продифференцировать по ϕ и, если подставить в формулу для M-шага, вычисления получаются несложными. То же самое касается и других рассмотренных в данном разделе регуляризаторов: модификации формул M-шага включают добавление новой константы или легко вычислимой величины.

7.4.5. Регуляризатор для отбора тем

Пусть $p(t)$ — распределение тем в коллекции документов:

$$p(t) = \sum_d p(d)\theta_{td}.$$

Еще одним полезным на практике регуляризатором, который основан на представлениях о дивергенции Кульбака-Лейблера, является регуляризатор разреживающий распределение $p(t)$.

$$R(\Theta) = -\tau \sum_{t \in S} \ln \sum_{d \in D} p(d)\theta_{td} \to \max.$$

где τ — коэффициент регуляризации. Действительно, максимизация KL-дивергенции между $p(t)$ и равномерным распределением, фактически, есть требование, чтобы как можно больше вероятностей в $p(t)$ приняли нулевые значения.

Этот регуляризатор можно использовать для отбора тем: после введения регуляризации вероятности для наиболее незначительных тем обнулятся, то есть такая ненужная тема будет удалена из коллекции. Это можно использовать, в том числе, для определения полного количества тем: необходимо сначала взять число тем избыточным а затем, постепенно с помощью регуляризатора удалять незначительные темы.

Интересным побочным эффектом этого регуляризатора оказалось удаление линейно-зависимых, расщепленных темы. Такие избыточные темы получаются при разделении слов одной темы в две. Этот эффект оказался нетривиальным и до сих пор не имеет теоретического обоснования, хотя хорошо работает на практике.

7.5. Мультимодальные тематические модели

7.5.1. Понятие модальности

На практике часто встречаются коллекции документов, которые включают в себя метаинформацию, связывающую каждый документ с элементами (токенами) каких-то конечных множеств (не обязательно слов). Эти конечные множества называются модальностью.

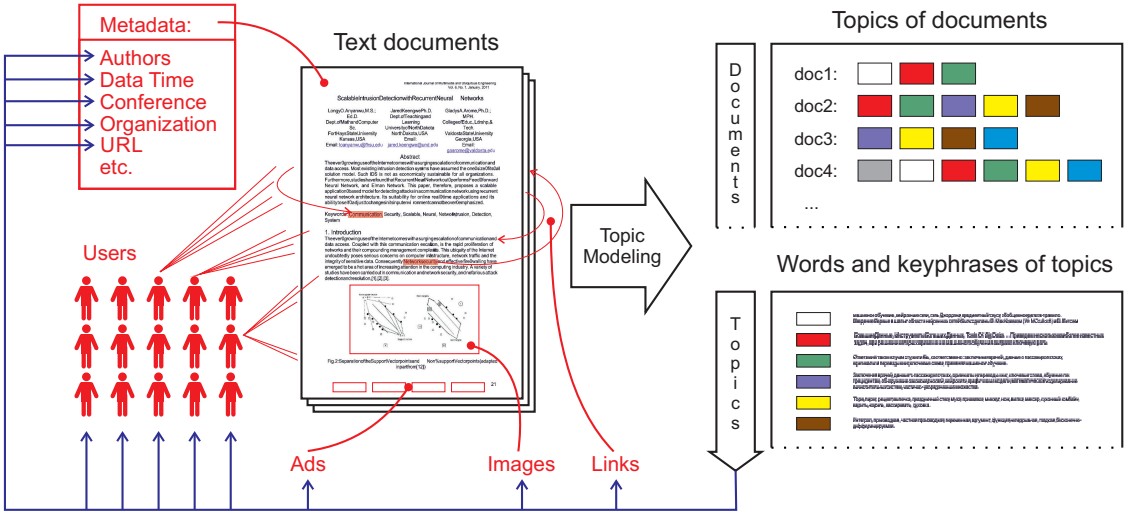


Рис. 7.4: Мультимодальная тематическая модель описывает появление элементов разных модальностей

Примеры модальностей:

- **Авторы, моменты времени и так далее:** в этом случае каждому документу приписывается соответственно метка автора, временная метка и так далее.
- **Элементы изображений**, содержащихся в документе. Изображение в таком случае можно мыслить как мини-документ, состоящий из псевдослов — элементов изображений.
- **Множество ссылок на другие документы**, в том числе гиперссылки в сети Интернет и цитирование других статей в научных трудах.
- **Множество рекламных баннеров, которые появились на данной странице**, а также **множество пользователей, которые кликнули на данные баннеры**, это два примера возможных модальностей.
- **Множество пользователей, сделавших определенное действие с документом (скачал, лайкнул, поставил оценку и так далее)**. После того, как операция выполнена, в системе остается запись о том, что данный пользователь сделал конкретную операцию. И поэтому можно считать, что в документ также включена и эта информация.

Чтобы иметь возможность пользоваться данными типами информации, необходимо строить тематические модели, которые описывают появление элементов разных модальностей в документе по известной тематике. Другими словами, благодаря тому, что документ относится к какой-либо теме, в нем появляются определенные слова из этой темы, на картинках изображены элементы, которые характерны для этой темы, а также его читают пользователи, которым эта тема интересна, и так далее.

7.5.2. Мультимодальная ARTM

Тематическая модель описывает появление элементов всех модальностей исходя из единого тематического профиля всего документа. Каждая модальность $m \in M$ описывается своим словарём токенов W^m, каждая тема имеет своё распределение $p(w|t)$ для каждой модальности $w \in W^m$.

Словарь всех модальностей — это объединение непересекающихся словарей отдельных модальностей:

$$W = W^1 \sqcup \cdots \sqcup W^M.$$

Можно построить вероятностную модель, как это делалось ранее, расписывая принцип максимума логарифма правдоподобия с регуляризацией (исходя из требуемых от модели свойств) отдельно для каждой модальности:

$$\sum_{m\in M} \tau_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta},$$

где τ_m — веса модальностей.

Такая постановка задачи оказывается не намного сложнее рассмотренной ранее и для нее легко выписывается модифицированный EM-алгоритм, представляющий собой метод простой итерации для системы уравнений относительно параметров модели ϕ_{wt}, θ_{td} и вспомогательных переменных $p_{tdw} = p(t|d,w)$:

$$\text{E-шаг:} \quad p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \operatorname*{norm}_{w\in W^m}\Big(\sum\limits_{d\in D} \tau_{m(w)} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum\limits_{w\in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

Всего модификации было сделано две: во-первых, частота слова в документе n_{dw} домножается на вес $\tau_{m(w)}$ соответствующей модальности $m(w)$ этого слова w. Во-вторых, при вычислении матрицы Φ она нормируется для каждой модальности по отдельности. Такое незначительное расширение математической модели позволяет решать разнообразные прикладные задачи с большим числом модальностей.

Пусть дана параллельная коллекция документов, то есть коллекция из оригинальных текстов и переводов этих текстов на другие языки. В этом случае различными модальностями будут различные языки. Если построить тематическую модель, учитывая все доступные модальности, появляется возможность делать кросс-язычный поиск, то есть получать результат на другом языке, нежели чем на котором был сделан запрос. Например, по тексту русскоязычному научной статьи можно будет в большой коллекции англоязычных текстов еще что-нибудь с похожей тематикой.

В качестве второй модальности можно использовать биграммы, то есть выделенные в тексте словосочетания. В таком случае получаются хорошо интерпретируемые модели, примеры которых были показаны в самом начале.