Обзор по вероятностным тематическим моделям

В настоящее время графические модели являются основным инструментом для построения *вероятностных тематических моделей* (probabilistic topic model). Модели со скрытыми переменными оказались особенно эффективными для выявления скрытых структур в текстовых коллекциях. В данном обзоре рассматривается важный подкласс *ориентированных вероятностных тематических моделей* (directed probabilistic topic models, DPTM), которые осуществляют мягкую кластеризацию и применяются для выявления тематики текстов в больших коллекциях документов. В терминах кластерного анализа *тема* (topic) — это результат би-кластеризации, то есть одновременной кластеризации и слов, и документов по их семантической близости. При *мягкой кластеризации* (soft clustering) каждое слово и каждый документ относится к нескольким темам одновременно с определёнными вероятностями. Таким образом, сжатое семантическое описание слова или документа представляет собой вероятностное распределение на множестве тем. Процесс нахождения этих распределений и называется *тематическим моделированием*.

Тематические модели активно развивались последнее десять лет. Предложено много моделей для решения задач моделирования текстовых коллекций в таких приложениях, как классификация документов, поиск похожих документов, поиск экспертов, выявление сообществ и анализ временных трендов. В данном обзоре рассматриваются основные концепции, преимущества и недостатки различных моделей в хронологическом порядке, предлагается классификация существующих моделей по различным категориям, описываются алгоритмы оценивания параметров и критерии качества моделей. Также обсуждаются приложения, открытые проблемы и будущие направления в этой динамично развивающейся области исследований.

1 Введение

Графические модели могут быть разделены на две основные категории: ориентированные и неориентированные графические модели. Эти типы можно далее разбить на параметрические и непараметрические (рис. 1).

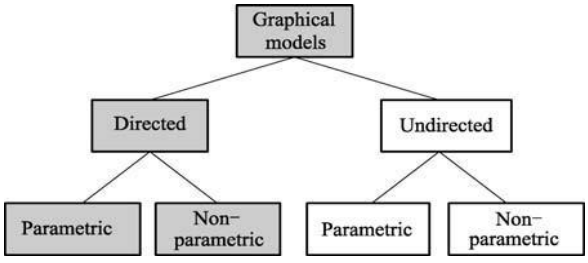

Рис 1. Классификация графических моделей.

Графические модели со скрытым тематическим слоем в последнее время успешно применялись в задачах поиска скрытых закономерностей в данных. Автоматическое выделение тематики текстов применялось в [1,2] для разбиения текстов по группам на основе семантической близости содержания. Эти модели позволяют классифицировать документы, но они ограничены предположением, что каждый документ относится только к одному кластеру. Модели мягкой кластеризации позволяют относить документ одновременно к нескольким кластерам, при этом каждый кластер ассоциируется с определённой темой, и каждый документ характеризуется оценками близости к каждой из тем. В *ориентированных вероятностных тематических моделях* (directed probabilistic topic model, DPTM) оценка близости документа к теме имеет вероятностный смысл и может интерпретироваться как доля содержимого документа, относящаяся к данной теме. DPTM — это относительно молодая область исследований в теории *самообучения* (обучения без учителя, unsupervised learning), представляющая в настоящее время значительный как теоретический, так и практический интерес. Одним из первых был предложен *вероятностный латентный семантический анализ* (probabilistic latent semantic analysis, PLSA), основанный на принципе максимума правдоподобия, как альтернатива классическим методам кластеризации, основанным на вычислении функций расстояния. Вслед за PLSA был предложен метод *латентного размещения Дирихле* (latent Dirichlet allocation, LDA) и его многочисленные обобщения. Применение тематических моделей позволяет получить ответ на целый ряд нетривиальных вопросов. Как выявлять смысл или тематику документов по их содержимому? Как осуществлять классификацию документов на основе этих скрытых тематических закономерностей? Как выявлять научные интересы авторов и находить экспертов в специальных областях знания? Как выявлять скрытые ассоциативные связи между отдельными исследователями или группами людей? Как подбирать коллаборации под проекты? Как выявлять тенденций в развитии научных направлений? Как выявлять роли людей в социальных сетях? Как осуществлять индексацию и автоматическое аннотирование документов?

Известны вводные обзоры [4,5] по вероятностным тематическим моделям и методам оценивания их параметров, однако они упускают некоторые детали. В данном обзоре на основе анализа обширного списка литературы предлагается классификация моделей DPTM по их функциональности, приводится описание основных моделей, сравниваются их достоинства и недостатки, обсуждаются приложения и перспективы развития. Основное внимание уделяется ориентированным моделям, называемым также байесовскими сетями. При этом остаются в стороне неориентированные вероятностные тематические модели, такие как «Harmoniums», основанные на марковских случайных полях [6,7].

Под *параметрическими моделями* будем понимать тематические модели, в которых темы изначально фиксированы и не меняются в процессе построения модели [8–10]. В *непараметрических моделях* число тем изначально не фиксировано, а сами темы настраиваются в процессе поиска наилучшего возможного модельного описания данных.

Изложение материала в обзоре организовано следующим образом. Во втором разделе описывается общая концепция и терминология DPTM. В третьем разделе обсуждаются ограничения ранних моделей, приводятся обоснования и обсуждаются ограничения более поздних моделей. В четвертом разделе описаны алгоритмы обучения и байесовского вывода. Пятый раздел посвящен способам оценивания качества тематических моделей. В шестом разделе описываются применения моделей в различных прикладных областях. Седьмой раздел посвящен открытым проблемам и перспективам развития моделей DPTM. В последнем, восьмом, разделе приводятся выводы и заключение.

Для большинства терминов и моделей приводятся их английские названия, чтобы упростить поиск англоязычных источников.

2 Концепции и терминология

В настоящем разделе описываются концепции и терминология, лежащие в основе теории тематического моделирования.

2.1 Основные концепции и терминология

2.1.1 Документ

Документ обычно состоит из множества слов, терминов (словосочетаний), специальных символов, таблиц, иллюстраций, и т.д. В исследованиях по тематическому моделированию типичными видами документов являются научные статьи и новостные сообщения.

Будем представлять документ d вектором (конкатенацией) слов $\boldsymbol{w}_d = \{w_{d1} + \cdots + w_{dn}\}$, где w_{di} — i-е слово в документе d.

Коллекцию из D документов будем представлять совокупностью векторов слов $D = \{w_1, \dots, w_D\}$, где $\boldsymbol{w}_d$ — вектор слов документа d.

2.1.2 Корпус текстов

Большую коллекцию текстовых документов принято называть *корпусом текстов* (text corpora). В исследованиях по тематическому моделированию часто используют общедоступные корпусы «NIPS proceedings» и «Cite seer». Оба содержат большое число научных статей и используются для тестирования

различных методов обнаружения знаний. Известные корпусы «TREC AP» и «Reuter's» используются для тестирования методов анализа новостей.

2.1.3 Тема

В литературе по тематическим моделям понятие *темы* (topic) определяется по-разному, в зависимости от научной школы: «скрытые паттерны», «компактные описания смысла документов», «вероятностные (нечёткие) кластеры семантически связанных терминов», «связующее звено между терминами и другими объектами (документами, авторами, организациями, конференциями, и т.д.), которое позволяет находить скрытые ассоциативные связи между ними».

Формально тема определяется как дискретное (мультиномиальное) вероятностное распределение в пространстве слов заданного словаря. Документ может состоять из огромного числа слов, однако эти слова могут порождаться небольшим числом тем, как смесью мультиномиальных распределений [12].

Таблица 1. Пример двух тем (искусство и образование).

Arts		Education	
Word	Probability	Word	Probability
New	0.03741	School	0.07344
Film	0.03626	Students	0.05702
Show	0.02753	Schools	0.04136
Music	0.02151	Education	0.02605
Movie	0.01854	Teachers	0.02465
Play	0.01124	High	0.02122
Musical	0.01109	Public	0.02026
Best	0.00989	Teacher	0.02006
Actor	0.00966	Bennett	0.01766
First	0.00899	Manigat	0.01746
York	0.00895	Namphy	0.01478
Opera	0.00870	State	0.0143
Theater	0.00854	President	0.01359
Actress	0.00817	Elementary	0.01219
Love	0.00806	Haiti	0.01211

В таблице 1 приведены примеры двух тем (искусство и образование). По каждой теме представлены пятнадцать наиболее употребительных слов и соответствующие им вероятности. Названия тем представляют собой только нашу интерпретацию.

2.1.4 Модель мешка слов

Предположение о том, что для целей анализа текстов (в нашем случае — для выявления тематики) важна только частота слов, но не их порядок, называется *моделью мешка слов* (bag of words). Когда не важен порядок предложений в документе или порядок документов в корпусе, соответствующим образом вводятся модели мешка предложений и мешка документов.

2.1.5 Тематические модели

Основная идея тематического моделирования заключается в описании документа смесью тем, т.е. в определении документа как выборки слов, порождаемой смесью вероятностных распределений [4].

Тематическим моделированием называется решение обратной задачи. Каждый документ в корпусе текстов рассматривается как наблюдаемая случайная независимая выборка слов (мешок слов), порождённая некоторым, как правило небольшим, латентным подмножеством тем. По этим данным требуется восстановить вероятностные распределения всех тем в корпусе и определить, каким именно подмножеством тем порождён каждый документ.

2.1.6 Синонимичные термины и обозначения

В области тематического моделирования много неустоявшихся терминов, и разные термины используются для определения одних и тех же понятий. Корпус текстов, коллекция документов (text corpora, corpus, large collections of documents) — синонимы. Модели текстовых коллекций, статистические модели языка (information discovery, statistical analysis, human language, learning and processing, modeling text corpora, statistical modeling of language, statistical language learning) — синонимы. Тема, скрытая тема, скрытый паттерн, скрытая закономерность, скрытая структура, краткое описание (topics, hidden topics, hidden patterns, latent topics, latent aspects, buried patterns, latent structure, and short descriptions) — синонимы.

В таблице 2 приводится список основных обозначений, используемых в данном обзоре.

Таблица 2. Основные обозначения.

Символ	Описание
D	число документов
N	число слов в коллекции
T	число тем
A	число авторов
V	число уникальных слов в словаре
N_d	число слов в документе d

$\boldsymbol{w}_d$	вектор слов документа d
$\boldsymbol{a}_d$	вектор авторов документа d
w_{di}	i-е слово в документе d
z_{di}	множество тем, присвоенных i-му слову в документе d
x_{di}	автор, соответствующий слову w_{di}
y_{di}	момент времени, соответствующий w_{di}
θ_d	мультиномиальное распределение в пространстве тем с параметром α
Φ_z	мультиномиальное распределение в пространстве слов для темы z с параметром β
Ψ_z	бета-распределение для темы z с учетом фактора времени
α	априорное распределение Дирихле на параметры θ
β	априорное распределение Дирихле на параметры Φ
ε	биномиальное распределение на Ω_i
rn	корневая вершина (корневая тема)
R	наблюдаемая целевая переменная, в тематических моделях с учителем
L	связь между документами
d	документ-источник
d'	целевой документ
τ	значение связи между документами
γ	распределение Дирихле на величину τ
λ	мультиномиальная порождающая модель для описания связей между документами
C	класс слова (именная группа или не именная группа)

2.2 Вспомогательные концепции и терминология

2.2.1 Неупорядоченность тем

Свойство *неупорядоченности* или *перестановочности* (exchangeability) тем состоит том, что порядок тем при различных запусках алгоритма может оказаться разным. В результате тема z_i с порядковым номером i при первом запуске может отличаться от темы с тем же номером при следующих запусках.

2.2.2 Выбор оптимального числа тем

Определение оптимального числа тем — важная подзадача в тематическом моделировании, поскольку её решение существенно влияет на осмысленность получаемого набора тем. Занижение числа тем приводит к чрезмерно общим результатам. Завышение приводит к невозможности разумной интерпретации. Оптимальное число тем зависит от числа документов в анализируемом корпусе: в малых корпусах оптимальным является, как правило, меньшее число тем.

Согласно [8], оптимальное число тем для корпуса из 16333 новостных статей составило 100, тогда как для корпуса из 5225 аннотаций научных статей — 50.

Существует несколько методов для определения оптимального количества тем. Во-первых, оптимизация оценки обобщающей способности модели путём разделения корпуса на обучающую и контрольную выборки. Модель строится по обучающей выборке и затем применяется для описания контрольной выборки. Для оценивания обобщающей способности вычисляется *мера неупорядоченности* (perplexity) на всех построенных подвыборках [8,9]. Во-вторых, в рамках байесовского подхода к выбору модели [10] могут оцениваться апостериорные вероятности моделей путём усреднения по всевозможным способам задания соответствия между темами и словами. Оптимальное количество тем выбирается по максимуму апостериорной вероятности модели. Метод оптимизации числа тем предложен в [11].

2.2.3 Полисемия

Полисемия (polysemy) — это важное языковое явление, широко обсуждаемое в области обработки естественного языка и состоящее в том, что слово может иметь множество значений. Неоднозначность интерпретации слова может быть снята с помощью контекста. Тематические модели позволяют реализовать эту возможность. В таблице 3 представлены две темы: спорт и развлечения. В десятке наиболее употребительных слов обеих тем встречается слово «играть». Налицо различие контекстов, в которых употребляется это слово, что позволяет эффективно обходить или устранять полисемию. Например, для решения задач классификации текстов могут быть выбраны другие слова. В то же время, слову «играть» может быть присвоена метка темы, найденная из контекста. Заметим, что названия тем «спорт» и «развлечения» являются лишь нашей интерпретации соответствующих тематических распределений слов.

Таблица 3. Полисемия слова «играть» (play) в двух темах

Sports	**Entertainment**
Game	Art
Play	Music
Ball	***Play***
Team	Part
Playing	Sing
Games	Like
Football	Poetry
Baseball	Band
Field	World
Sports	Rhythm

2.2.4 Мультиномиальное распределение

Мультиномиальное распределение является обобщением биномиального распределения. Биномиальное распределение — это распределение числа успехов в n независимых испытаниях по схеме Бернулли с постоянной вероятностью успеха. Мультиномиальное распределение описывает эксперимент с k возможными исходами. Мультиномиальная функция вероятности $f(y_1, \dots, y_k, p_1, \dots, p_k)$ равна вероятности получить j-й исход ровно y_j раз, $y_1 + \cdots + y_k = n$.

2.2.5 Распределение Дирихле

Распределение Дирихле $\mathrm{Dir}(\alpha)$ широко используется в байесовской статистике как сопряженное априорное распределение к мультиномиальному распределению. Его функция плотности вероятности $f(x_1, \dots, x_k, \alpha_1, \dots, \alpha_k)$ имеет k-мерный вектор неотрицательных вещественных параметров $\alpha = (\alpha_1, \dots, \alpha_k)$ и определяется как доверительная вероятность того, что вероятность каждого из k взаимоисключающих исходов равна x_i при условии, что каждое событие наблюдалось $\alpha_i - 1$ раз.

2.2.6 Бета-распределение

Бета-распределение представляет собой специальный случай распределения Дирихле с двумя параметрами и является сопряженным к биномиальному априорному распределению (вспомним, что распределение Дирихле есть сопряженное априорное распределение к мультиномиальному распределению). В байесовской статистике это есть апостериорное распределение параметра p биномиального распределения после наблюдения $\alpha - 1$ независимых событий с вероятностью p и $\beta - 1$ событий с вероятностью $1 - p$ при условии, что априорное распределение параметра p равномерно.

2.3 Концепция порождающих моделей

2.3.1 Плоское графическое представление

Плоские графические представления служат для наглядного описания тематических моделей. На графах затененный и белый круг обозначают скрытую и наблюдаемую переменную соответственно. Стрелка обозначает отношение условной зависимости между переменными. Прямоугольник, включающий в себя некоторый подграф, с указанным в правом нижнем углу числом N обозначает совокупность N экземпляров данного подграфа. Эти символы приведены на рис. 2. Детальная нотация плоских графических представлений описана в [13].

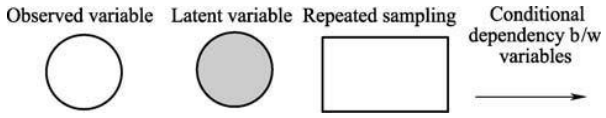

Рис.2. Условные обозначения на плоских графических представлениях.

2.3.2 Порождающие модели

Порождающая модель (generative model) описывает вероятностный закон генерации случайных значений наблюдаемых переменных. Порождающие модели используются как непосредственно для моделирования входных данных, так и для оценки соответствующего условного распределения с использованием формулы Байеса.

Тематическая порождающая модель описывает, как именно слова документа порождаются вероятностной смесью тем, определяемой значениями латентных переменных. Предполагается, что данные представляют собой реализацию случайной величины, распределенной согласно порождающей модели [4]. Задача заключается в том, чтобы по наблюдаемым данным оценить (восстановить) максимально правдоподобные значения латентных переменных.

2.3.3 Байесовская сеть

Байесовская сеть (Bayes network) или *ориентированная графическая модель* представляет собой вероятностную модель, в которой условные зависимости переменных описываются с помощью ориентированного графа. Например, байесовская сеть может моделировать зависимость между осадками и урожайностью: при заданном количестве осадков модель позволяет оценить распределение урожайности различных культур.

3 Ориентированные тематические модели

В данном разделе рассматривается история развития тематических моделей и предлагается их классификация на пять категорий по основным функциональным возможностям. Анализ преимуществ и недостатков моделей в порядке их появления позволяет разобраться, какие проблемы более ранних моделей были решены позднее. В конце каждого подраздела приводится краткое описание базовых предположений и основных характеристик соответствующей категории моделей.

Исторический анализ парадигм тематических моделей выявляет несколько интересных тенденций.

Таблица 4. Основные вероятностные тематические модели 1999–2009.

Year/Type	Basic PDPTMs	Inter-Document Correlated PDPTMs	Intra-Document Correlated PDPTMs	Temporal PDPTMs	Supervised PDPTMs
1999	PLSA				
2000					
2001		**PLSA**→A Joint Probabilistic Model			
2002	A probabilistic Approach				
2003	LDA, A Topic Model				**LDA** → Corr-LDA
2004	Discrete PCA	**LDA** →Mixed Membership Models, **LDA** →Author-Topic Model, **LDA**→ **Author-Topic Model** →ART			
2005			**LDA** → HMM-LDA		**LDA** →LLDA
2006		**LDA** → PAM, **LDA** → CTM, **LDA** →Statistical Entity-Topic Models	Bigram Topic Model, **PLSA** → CPLSA	**LDA** → TOT, **LDA** → DTM, **(PAM, TOT)** → Continuous Time Model	
2007		**LDA** → GWN-LDA, **LDA** → Citation Influence Model	**LDA** →HTMM, **LDA** → TNG	MTTM	**LDA** → sLDA
2008		**LDA** →LTHM, **LDA** → **Author-Topic Model** → ACT Model **(A Joint Probabilistic Model, LDA)** → Link-PLSA-LDA		**LDA** → **DTM** →cDTM	
2009		**LDA** → Generalized LDA, **LDA** → **Author-Topic Model** → **ACT** à Generalized ACT		**LDA** → **Author-Topic Model** → TAT	

Из таблицы 4 видно, что LDA безусловно лидирует среди вероятностных тематических моделей благодаря многочисленным обобщениям и приложениям к анализу коллекций текстовых документов. Из таблицы 4 можно также заметить, что тематические модели не ограничиваются предположением, что текст — это «мешок слов». В них также могут учитываться связи между документами, марковские зависимости внутри документа, временные тренды и дополнительные данные о метках (классификациях) документов. Некоторые модели имеют множество приложений, например, автор-тематическая модель, которая используется в основном для определения интересов авторов и поиска сообществ по тематикам, но также может быть использована для нахождения временных тематических трендов. Модель с непрерывным временем [15] имеет богатый набор функциональных возможностей, так как может использовать сразу и временную информацию, и марковские зависимости в документах для обнаружения временных тематических трендов.

В таблице 4 стрелки (→) показывают, что модель, выделенная жирным шрифтом, была расширена до модели, выделенной обычным шрифтом. Здесь более поздние модели представлены как улучшения более ранних, и можно полагать, что они являются лучшими для решения задач в своих категориях.

3.1 Базовые тематические модели

Вероятностный латентный семантический анализ (probabilistic latent semantic analysis, PLSA) [3] основан на введении слоя скрытых переменных для

описания тематик в коллекции текстовых документов. В PLSA каждый документ представляется числовым вектором, каждая компонента которого равна доле соответствующей темы в документе. Однако вероятностная модель не описывает ни закон распределения этих долей, ни вероятности самих документов. В результате число параметров модели линейно растёт с ростом размера текстовой коллекции, что может приводить к переобучению. Кроме того, не понятно, как оценивать вероятности новых документов, не входивших в состав обучающей выборки [8]. Другими словами, модель задаёт закон порождения слов, но не закон порождения документов.

В [16] предложен метод *смеси униграмм* (mixture of unigrams), основанный на униграммной модели языка. В униграммной модели предполагается, что слова каждого документа выбираются случайно и независимо из одного и того же мультиномиального распределения. Таким образом, модель основана на предположении, что каждый документ представляет только одну тему, что является существенным ограничением при моделировании текстовых коллекций [8]. При дополнении униграммной модели дискретной случайной величиной z, соответствующей темам, получается смесь униграммных моделей.

Модель PLSA является важной вехой в развитии вероятностного моделирования текстов, и она, несомненно, полезна для задач информационного поиска. Однако она имеет довольно существенные ограничения. Поэтому в [8] была предложена модель *латентного размещения Дирихле* (latent Dirichlet allocation, LDA), лишённая недостатков PLSA. В LDA предполагается, что каждое слово в документе порождено некоторой латентной темой, при этом в явном виде моделируется распределение слов в каждой теме, а также априорное распределение тем в документе. Темы всех слов в документе предполагаются независимыми. В LDA, как и в PLSA, документ может соответствовать не одной теме. Но LDA задаёт модель порождения как слов, так и документов, поэтому появляется дополнительная возможность оценивать вероятности документов вне текстовой коллекции с помощью алгоритма вариационного вывода и семплирования Гиббса. В отличие от PLSA, в LDA число параметров не увеличивается с ростом числа документов в коллекции. Многочисленные расширения модели LDA (см. таблицу 4) устраняют некоторые её ограничения и улучшают производительность для конкретных задач.

Предлагались и другие похожие подходы к тематическому моделированию [17,18], выявляющие вероятностные семантические связи между словами и документами. Эти подходы также основаны на идее описания документов как вероятностных смесей тем, где каждая тема описывается распределением вероятностей на всех словах в коллекции текстовых документов.

Влияние LDA на исследования по моделированию текстовых коллекций трудно переоценить. В то же время, в [19] сделана попытка обобщения более ранних моделей в форме дискретного варианта *метода главных компонент* (discrete principle component analysis, discrete PCA), ориентированного на анализ больших массивов данных. Утверждается, что PLSA, LDA и expectation-propagation [3,8,20] — это похожие подходы, отличающиеся, главным образом, мотивацией и обозначениями. Все они могут быть выражены в терминах дискретного PCA. Все они позволяют повышать обобщающую способность на контрольных данных при моделировании текстовых коллекций.

Итак, базовые вероятностные тематические модели позволяют выявлять скрытую тематику документов на основе модели документа как мешка слов. В них также предполагается существование скрытых взаимосвязей между различными объектами (документами, авторами, конференциями или журналами, организациями, пользователями), которые могут проявляться в структуре словоупотребления. Семантическая близость различных объектов может оцениваться путём сравнения их тематических векторов.

Основные базовые вероятностные тематические модели более подробно рассматриваются ниже.

3.1.1 Вероятностный латентный семантический анализ (PLSA)

PLSA [3] — это первая вероятностная тематическая модель с латентными переменными, имеющая строгие статистические обоснования. В [21] она также называется *моделью аспектов* (aspect model). Эта модель основана на предположении, что совместное появление пар (документ, слово) обусловлено латентными переменными $z \in T = \{z_1,...,z_t\}$. Совместное распределение на парах $d \times w$ определяется как смесь распределений:

$$p(d,w) = p(d)\,p(w \mid d), \quad \text{где} \quad p(w \mid d) = \sum_{z \in T} p(w \mid z)\,p(z \mid d). \qquad (1)$$

Предполагается, что каждая пара (d,w) появляется независимо, в соответствии с предположением о том, что документ — это «мешок слов». Слова *w* появляются в документе в зависимости от тем *z*, но независимо друг от друга. Таким образом, модель PLSA описывает порождение слов в документе, но не описывающая порождение самих документов. Соответствующее графическое представление модели приведено на рис.3.

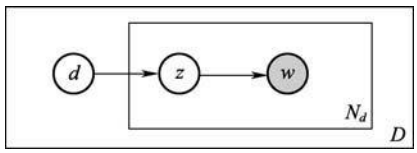

Рис. 3. Вероятностная модель PLSA.

Для оценивания параметров модели используется *EM-алгоритм* (expectation-maximization) — стандартная итерационная процедура идентификации скрытых переменных путём максимизации функционала правдоподобия.

3.1.2 Латентное размещение Дирихле (LDA)

PLSA является порождающей моделью только для слов, но не для документов. Модель LDA обходит это ограничение. Основная идея LDA заключается в том, что документы представляются смесью распределений латентных тем, где каждая тема определяется вероятностным распределением на множестве слов. Модель LDA выявляет скрытые связи между словами посредством тем. Она также позволяет присваивать вероятности новым документам, не входившим в обучающую выборку, используя алгоритм вариационного вывода.

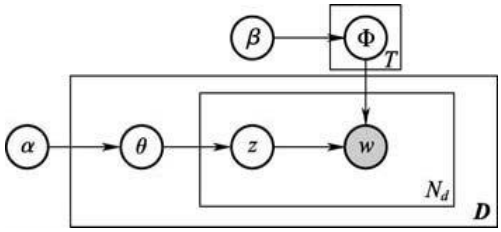

Рис. 4. Вероятностная модель LDA.

Графическое представление модели LDA показано на рис.4. Фактически LDA является трехуровневой байесовской сетью, которая порождает документ из смеси тем. На первом шаге для каждого документа d выбирается случайный вектор θ_d из распределения Дирихле с параметром α (обычно α принимается равным $50/T$). На втором шаге выбирается тема z_{di} из мультиномиального распределения с параметром θ_d. Наконец, согласно выбранной теме z_{di} выбирается слово w_{di} из распределения Φz_{di}, которое является распределением Дирихле с параметром β (обычно параметр $\beta = 0.1$, увеличение β ведёт к более разреженным тематикам) [10]. Таким образом, порождающая модель слова w из документа d представляется в виде:

$$p(w \mid d, \theta, \Phi) = \sum_{z=1}^{T} p(w \mid z, \Phi_z)\, p(z \mid d, \theta_d). \qquad (2)$$

Для оценки параметров LDA используются простой вариационный EM-алгоритм и семплирование Гиббса [8,10].

3.2 Тематические модели, учитывающие отношения между документами

PLSA описывает семантическую связь между документами и словами с помощью скрытого тематического слоя. Наряду с этим, между научными

статьями существует естественная связь в виде цитирования. Модель, учитывающая как тематики, так и цитирование, называется *совместной вероятностной моделью* (joint probabilistic model) [22] и представляет собой обобщение PLSA. Эта модель описывает тематическую структуру документов, но в то же время определяет порождающий процесс не только для текста, но и для цитирования. Ей присущи те же проблемы большого числа параметров и переобучения, характерные для ее предшественника PLSA [3]. Зависимости между словами в документе и связи между документами, получаемые, например, на основании цитирования, рассматриваются также в [16,22] при анализе смесей униграмм и совместной вероятностной модели. Расширение этой модели с учётом марковских зависимостей реализуется в рамках *скрытых тематических марковских моделей* AHMM [23]. Однако эти модели строятся на эвристическом предположении о том, что каждый документ соответствует только одной теме, как в модели смеси униграмм.

Опыт применения совместной вероятностной модели [22] показывает как важность учета связей цитирования между документами, так и возникающие на этом пути ограничения. Для преодоления этих ограничений были предложены порождающие *модели смешанного членства* (mixed-membership models) [24]. Они могут рассматриваться как развитие LDA со встроенным учётом ссылок или цитирований. Их ограничение состоит в неспособности учитывать тематические отношения между текстами [25].

Ключевая идея тематических моделей — использование скрытого слоя тематических переменных для описания взаимосвязей между терминами и документами — естественным образом переносится на задачу описания взаимосвязей между терминами и авторами [14] и выявления тематики интересов авторов. Для этого используется дополнительная информация об авторстве документов и строятся *автор-тематические модели* (author-topic model). Эти модели могут быть также использованы для анализа развития тем во времени, поиска наиболее релевантных авторов по теме, выявления нетипичных для данного автора текстов.

В задачах выявления связей между участниками социальных сетей автор-тематические модели оказалось менее перспективным. Для этих задач были разработаны специальные тематические *модели автор-получатель* (author-recipient-topic models) [26]. Они отличаются от автор-тематических моделей тем, что в них распределение тем в каждом документе (сообщении) зависит не только от автора, но и от получателя этого сообщения. Таким образом, при выявлении тем и ролей учитывается структура социальной сети, то есть связей между авторами и получателями сообщений.

Модель LDA, как и многие другие, не способна в явном виде учитывать взаимосвязи между темами (сходство или отношение тема–подтема). В то же время, для многих текстовых коллекций предположение о наличии таких взаимосвязей вполне естественно. *Корреляционные тематические модели* позволяют учитывать взаимосвязи между темами [27]. В модели CTM оценивание корреляций приводит к квадратичному по числу тем росту числа оцениваемых параметров. В [28] предлагается более гибкая модель PAM, позволяющая описывать произвольные вложенные разреженные взаимосвязи между темами с помощью ориентированного ациклического графа.

Описанные выше модели выявляли тематическую структуру документа, оставляя в стороне вопрос об отношениях между представленными в тексте объектами. В [29] предлагается статистическая *объектно-тематическая модель* (statistical entity-topic model), в которой не только слова, но также и объекты представляются как смеси тем. Эта модель применялась для анализа корпуса новостных сообщений.

В [14] автор-тематическая модель используется для выявления групп авторов относительно тем. Похожая задача рассматривается в [30], где выявляются вероятностные профили сообществ в социальных сетях и предлагается обобщённая *взвешенная сетевая модель* GWN-LDA. Здесь сообщества описываются скрытыми переменными, которые представляются как распределения в пространстве участников социальной сети. Проводится анализ эффективности GWN-LDA в сравнении с метрическими алгоритмами кластеризации, не учитывающими скрытых структур документа.

При рассмотрении цитирования и прочих подобных отношений между документами, естественно предположить, что если одна статья цитируется в другой, то между этими статьями существует также и тематическая связь. В ранее обсуждавшихся моделях [22,24] уже возможно учитывать дополнительную информацию о связях, что позволяет более точно выявлять тематику и зависимости между документами. Не так давно была предложена *модель влияния цитирований* (citation influence model) для прогнозирования влияния документа на цитирующие его тексты [31]. В этой модели структура цитирования описывается ориентированным графом. При этом рассматривается только совокупное влияние цитирований, без разделения влияний по темам. Для решения данной проблемы в [25] предложено расширение данной модели, названное Link-PLSA-LDA. Эти модели используют тематические отношения между цитируемыми и цитирующими документами, рассматривая цитируемые документы как мешки, которые необходимо заполнить словами. Ограниченность данной модели в том, что цитирующие и цитируемые документы порождаются по отдельности, так что отдельный документ не может

одновременно быть и цитированным, и цитировать другие. Кроме того, тематическое распределение определяется по фиксированной коллекции документов, то есть данная модель является порождающей на уровне слов, но не на уровне документов.

В работе [32] анализируются недостатки моделей [24,31] (неспособность учитывать тематические отношения между обоими связанными цитированием текстами) и предлагается *скрытая тематическая модель гипертекста* LTHM, способная описывать гипертекстовый корпус с учётом как ссылок внутри документов, так и ссылок на другие документы. На примере анализа корпусов текстов Википедии и webkb показано, что данная модель более эффективна по сравнению с другими моделями благодаря сокращению числа параметров.

Автор-тематические модели использовались для одновременного описания авторов и документов с целью выявления тематики авторов. В работе [33] утверждается, что авторы и конференции взаимозависимы через тематику, и потому могут моделироваться совместно. Была предложена соответствующая *тематическая модель автор-конференция* (author-conference topic model, ACT), позволяющая выявлять структуру научной социальной сети на основе семантических взаимосвязей терминов и авторов по корпусу текстов докладов на научных конференциях. В [34] предложено обобщение этого метода для решения задачи поиска экспертов на основе информации о семантике и времени (semantic and temporal information based maven search, STMS). В основе STMS лежат две идеи: 1) авторы, публикующиеся на конференциях мирового уровня, вероятнее являются влиятельными экспертами в своей области и 2) что прошедшие строгий отбор статьи наилучшим образом отвечают соответствующим подтемам, следовательно, данные статьи наиболее типичны и сильнее связаны тематически. Модель STMS рассматривает совокупность статей и авторов определенной конференции как виртуальный документ и учитывает тематическую структуру слов, связи между авторами и хронологию конференций, в отличие от модели ACT, рассматривавшей отдельные документы без учета времени [33].

В [35] утверждается, что выявление тематики на уровне конференций даёт более интересные результаты, чем на уровне документов, и предлагается подход к *интеллектуальному анализу данных о конференциях* (conference mining) на основе *обобщённого LDA* (generalized LDA). Совокупность статей конференции рассматривается как один супер-документ, при этом семантическая структура слов определяется без учета информации об авторстве. Такой подход строит более информативные модели по сравнению с моделями ACT [33], и потому оказывается предпочтительнее для ранжирования конференций и выявления связей между конференциями.

Итак, совместное использование информации об употреблении слов (терминов) в документах и о связях между документами (цитирование, авторство, объединение в рамках одной конференции или сборника и др.) повышает качество тематических моделей. Ниже некоторые из моделей этой категории рассматриваются более подробно.

3.2.1 Автор-тематическая модель

Автор-тематическая модель (author-topic model) [14] представляет собой расширение LDA для совместного описания документов и авторов. Идея моделирования слов и авторов для выявления области интересов авторов является непосредственным развитием идеи PLSA и LDA о моделировании слов документа с помощью скрытого тематического слоя для поиска зависимостей между документами. В действительности, приступая к написанию текста, автор сначала определяется с темой, затем подбирает слова сообразно с этой темой. Основанная на этом предположении порождающая модель оказалась в состоянии успешно выявлять тематические интересы авторов.

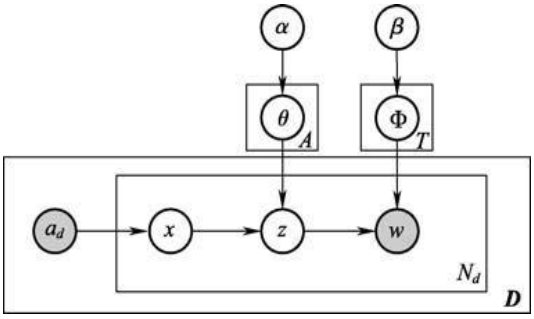

Рис. 5. Автор-тематическая модель.

Графическое представление автор-тематической модели приведено на рис.5. В рамках этой модели каждой теме соответствует мультиномиальное распределение Φ в пространстве слов. Каждому автору из множества A соответствует мультиномиальное распределение θ в пространстве тем. Для параметров обоих распределений Φ и θ введены симметричные априорные распределения Дирихле с параметрами α и β. Для каждого слова в документе автор x выбирается независимо из множества соавторов a_d, затем выбирается тема z из мультиномиального распределения θ, соответствующего автору x, наконец слово w выбирается из тематического распределения Φ, соответствующего теме z:

$$p(w|a,d,\Phi,\theta) = \sum_{z=1}^{T} p(w|z,\Phi_z)p(z|a,\theta_a). \qquad (3)$$

Для обучения модели используется сэмплирование Гиббса, являющееся известным методом из семейства марковских цепей Монте-Карло.

3.2.2 Скрытая тематическая модель гипертекста

Скрытая тематическая модель гипертекста (latent topic hypertext model, LTHM) описывает закон порождения ссылок в корпусе гипертекстов. В этой модели совместно используется информация о словах и о ссылках, что позволяет более точно восстанавливать тематику, приписывать темы ссылкам и генерировать новые ссылки, имеющую высокие вероятности в пределах темы.

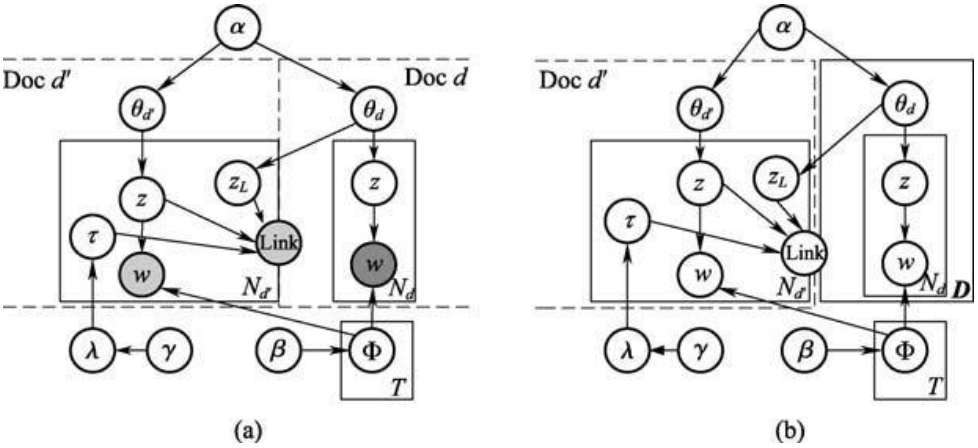

Рис. 6. Модель LTHM.

На рис. 6 приведена иллюстрация использования LTHM в двух сценариях: (a) как модели, порождающей ссылки из документа d' в заданный целевой документ d и (b) как модели, порождающей ссылки из d' в произвольный документ корпуса.

Порождение ссылок происходит в два шага. На первом шаге используется порождающий механизм, аналогичный LDA. На втором шаге по порожденным словам, производится порождение ссылок. Для порождения ссылки τ от слова w_i на документ τ_i используется мультиномиальное распределение λ с параметром γ, имеющим распределение Дирихле. Вероятность ссылки от слова на документ d зависит от частоты появления соответствующей этому слову темы в документе d, а также от частоты ссылок на данный документ в коллекции. Вероятность порождения ссылки из документа d на документ d' зависит от темы слова, от которого исходит ссылка, частоты ссылок на документ d' и тематического распределения, соответствующего документу d'. Для обучения и вывода в описанной модели авторы используют EM-алгоритм.

3.3 Тематические модели, учитывающие зависимости между словами документа

Идея описывать зависимости между словами документа с помощью *скрытых марковских моделей* (HMM) была реализована в тематической модели AHMM [23], но при ограничении, что каждый документ может относиться только к одной теме. Это ограничение было снято позже в очередном

расширении LDA, названном HMM-LDA [36]. В этой модели каждое предложение разбивается на функциональные слова, за генерацию которых отвечает скрытая марковская модель, и на термины, генерируемые тематической моделью LDA. Единственное ограничение этой модели состоит в том, что при выделении тематики никак не используется дополнительная информация, заключённая в локальной структуре текста [37].

Другой способ добавить учёт марковских зависимостей между словами в существующие тематические модели был найден в [38], где была предложена *биграммная тематическая модель* (bigram topic model). В рамках этой модели для порождения слова наряду со скрытым тематическим слоем используются предшествующие слова (т.е. дополнительно вводится марковская структура на уровне слов) Эта модель обеспечивает большую точность в сравнении с моделями, основанными на предположении мешка слов. Однако марковское предположение, что появление слова зависит только от предшествующего слова, слишком обременительно для многих реальных текстовых корпусов.

Обе модели [36,38] описывают локальные синтаксические и глобальные семантические зависимости между словами в документах. В [39] предлагается *модель контекстных смесей* (contextual mixture, CPLSA), которая обобщает PLSA [3] путём введения контекстных переменных для моделирования синтаксических зависимостей в документе и позволяет описывать скрытые взаимосвязи между темами и объектами в динамике.

Предположение о том, что каждое предложение целиком относится к одной теме, и соседние предложения, скорее всего, также относятся к этой теме, лежит в основе применения скрытых марковских моделей в задаче выявления тематики. Реализация этой идеи [37] в рамках *марковской модели скрытых тем* (hidden topic Markov model, HTMM) превосходит LDA по качеству выделяемых тем и снятия полисемии. В том же году была предложена *n-граммная тематическая модель* (topical *N*-gram model, TNG) [41], которая также описывает марковские зависимости. Она выделяет семантически связанные фразы произвольной длины, в отличие от семантически связанных униграмм (отдельных слов), как биграммная модель [38].

Тематические модели, учитывающие междокументные связи, чаще всего используют формализм марковских цепей для описания локальных синтаксических связей внутри документов. Они рассматривают документ не как мешок слов, а как мешок биграмм, *n*-грамм, предложений или абзацев. Такой подход имеет свои обоснования в области анализа последовательностей и анализа естественного языка. Учёт локальных синтаксических зависимостей улучшает интерпретируемость выделяемых тем.

3.3.1 Композитная модель HMM-LDA

В рамках композитной модели [36] HMM-LDA строится совместное описание синтаксиса и семантики текста. Скрытая марковская модель (HMM) описывает локальные закономерности между соседними словами, тогда как модель LDA даёт глобальное тематическое описание документа в целом.

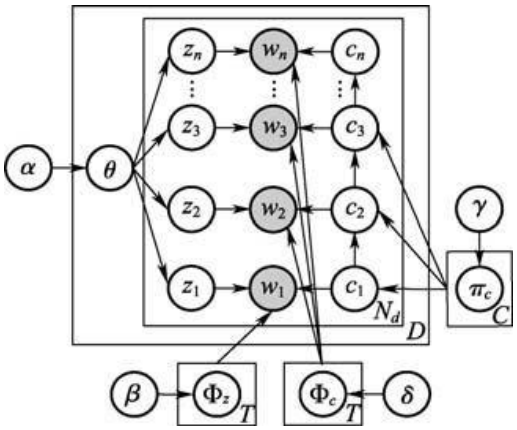

Рис. 7. Модель HMM-LDA.

Графическое представление модели приведено на рис. 7. Модель состоит из последовательности переменных-слов $\boldsymbol{w} = (w_1, \dots, w_n)$, тематических переменных $\boldsymbol{T} = (z_1, \dots, z_n)$ и последовательности бинарных классификаций $\boldsymbol{C} = (c_1, \dots, c_n)$. Если $c_i = 1$, то соответствующее слово анализируется в семантическом аспекте, т. е. на основании тематического распределения Φ_z. Если же $c_i \neq 1$, то слово порождается распределением Φ_C и не несёт смысловой нагрузки. Каждому документу соответствует распределение в пространстве тем θ_d и матрица вероятностей переходов для описания переходов между классами c_{i-1}, c_i в марковской цепи. Апостериорное распределение тем z_i может быть выписано в следующем виде:

$$p(z_i|z_{i-1}, c, \boldsymbol{w}) \propto p(z_i|z_{i-1})p(w_i|z, c, w_{i-1}) \propto \begin{cases} n_{zi}^{(di)} + \alpha, & c_i \neq 1; \\ \left(n_{zi}^{(di)} + \alpha\right)\dfrac{n_{wi}^{(zi)} + \beta}{n_{wi}^{(zi)} + V\beta}, & c_i = 1; \end{cases}$$

где $n_{zi}^{(di)}$ — число слов в документе d_i, относящихся к теме z_i; $n_{wi}^{(zi)}$ — общее число слов, относящихся к теме z_i. Подсчёт слов производится только для слов i, для которых $c_i = 1$. Для оценивания параметров модели применяется байесовский вывод с использованием сэмплирования Гиббса.

3.3.2 Скрытая тематическая марковская модель

Скрытая тематическая марковская модель (hidden topic Markov model, HTMM) [37] основана на предположении, что последовательность тем в документе является марковской цепью. Каждое предложение целиком

относится к одной теме, и соседние предложения, скорее всего, относятся к той же теме. Это позволяет правильнее относить слова к темам с учётом полисемии и приводит к более интерпретируемым темам, в сравнении с LDA [10].

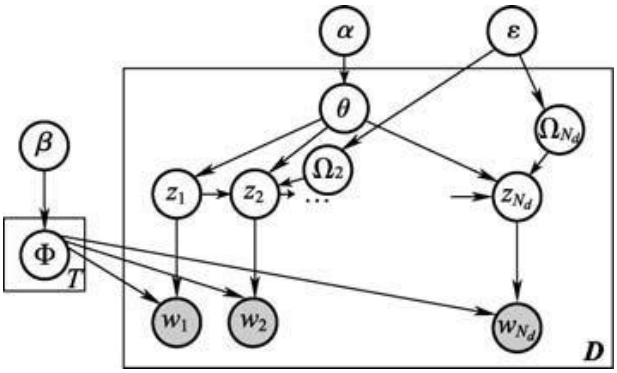

Рис. 8. Модель HTMM.

Из рис. 8 показано, что темы в документе образуют марковскую цепь с вероятностями переходов, зависящими от θ и переходной переменной Ω_{n}, имеющей биномиальное распределение с параметром ε, где $\Omega_n = 0$, если тема текущего слова совпадает с темой предыдущего и $\Omega_n = 1$ в противном случае. Ненулевые Ω_{n} может иметь только первое слово предложения, таким образом, всем словам предложения присваивается одна и та же тема. Через T обозначено число тем, через N_d — длина документа d. Порождающее распределение вероятностей вычисляется для каждого предложения в тексте и имеет вид

$$p(z_n, \Omega_n | d, w_1, \dots, w_n; \theta, \Phi, \varepsilon). \quad (5)$$

Для обучения модели используются стандартные для скрытых марковских моделей EM-алгоритм и алгоритм *вперёд-назад* (forward-backward) [42]

3.4 Темпоральные тематические модели

Особый интерес представляет анализ изменения тематики в корпусе текстов с течением времени. Для этого в [43] было предложено обобщение LDA, названное *динамической тематической моделью* (dynamic topic model, DTM). Эволюция тем прослеживается внутри корпуса текстов, организованного в виде последовательности. Время предполагается дискретным, поэтому мелкие изменения игнорируются, но выявляются интервалы повышения и понижения популярности терминов. Одной из проблем в DTM является то, что в качестве сопряжённого к мультиномиальному распределению берётся нормальное, что затрудняет оценивание параметров и вероятностный вывод в рамках модели.

Альтернативой DTM является *многомасштабная томографическая модель* (multiscale-topic tomography model, MTTM) [44], более естественная для анализа последовательных корпусов. В ней используются сопряжённые распределения и

одновременно несколько различных масштабов времени. Всё это обеспечивает в результате лучшую интерпретируемость тем.

Ещё одно расширение DTM — *динамическая тематическая модель с непрерывным временем* (continuous time dynamic topic model, cDTM) предложено в [45]. Она основана на использовании законов броуновского движения [46] для моделирования эволюции тем в непрерывном времени, что позволяет также снять некоторые ограничения по памяти и производительности, присущие DTM.

В модели тематики во времени (topics over time, TOT) [47] время создания документа является наблюдаемой переменной. Все слова документа получают одну и ту же отметку времени. Каждой теме ставится в соответствие бета-распределение, зависящее от времени, так что тема порождает одновременно и слово, и отметку времени, игнорируя, однако, временные паттерны и возможные взаимосвязи между темами, такие, как сходство или вложенность (отношение тема–подтема). Поэтому в [15] было предложено расширение этой модели — *модель непрерывного времени* (continuous time model, CTM), основанная на использовании ориентированного ациклического графа. Она не имеет недостатков TOT и позволяет обнаруживать как зависимости между темами, так и их изменения во времени в корпусе научных статей.

TOT моделирует только изменение тем во времени, но не выявляет интересы авторов. Ясно, что интересы авторов меняются во времени, и в разные годы одной и той же темой могут заниматься различные авторы. Для совместного описания динамики тематики и интересов авторов была предложена *темпоральная автор-тематическая модель* (temporal-author-topic model, TAT) [48], которая обобщает автор-тематические модели путём введения мультиномиального распределения для описания объектов, меняющихся во времени. Модель TAT позволяет ранжировать авторов по годам и темам и находить тематические связи между авторами в тот или иной период времени.

Итак, темпоральные тематические модели описывают изменения тем во времени на основе отметок о времени создания документов, как правило, по годам. Основная идея состоит в том, что семантическая структура множества слов и отношения между сущностями (например, авторами) меняются из года в год, поэтому возникает необходимость выявления тематических структур слов и других объектов (в частности, авторов) как функций от времени.

3.4.1 Модель с непрерывным временем

Модель с непрерывным временем CTM [15] объединяет в себе достоинства двух более ранних моделей, PAM [28] и TOT [47] с целью описания временных изменений как в самих темах, так и в отношениях между темами. Модель

способна выявлять произвольные зависимости между темами, равно как и локальные изменения темы.

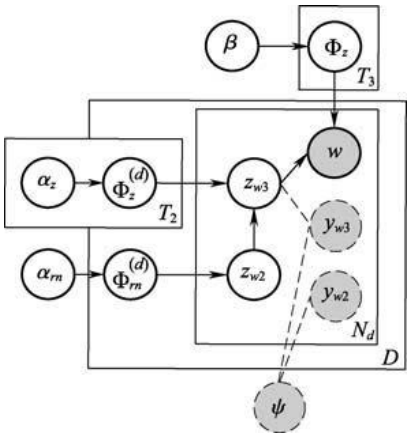

Рис. 9. Модель с непрерывным временем.

На рис. 9 приведено графическое представление модели с непрерывным временем. Здесь через $T = (z_1, \dots, z_n)$ обозначается множество тем. Корневая вершина обозначается через *rn*. Она не имеет вершин-предков, и любой теме соответствует путь от корня до вершины, соответствующей данной теме. Темы являются внутренними вершинами графа, темы — листьями. С каждой темой z_i связано распределение Дирихле с параметром α_i размерность которого равна числу потомков вершины z_i.

В этой модели для каждого слова w документа d тематический путь z_w выбирается из мультиномиальных распределений $\Phi_{z_1}, \dots, \Phi_{z_n}$. Отметка времени y_{wi} выбирается для каждой темы z_{wi} в тематическом пути из соответствующего бета-распределения $\psi_{z_{wi}}$ с параметрами Ψ_z. Наконец, совместное вероятностное распределение корпуса документов с отметками времени определяется как произведение вероятностей для отдельных документов:

$$p(d, y \mid \alpha, \psi) = \prod_d p(d, y^{(d)} | \alpha, \psi)). \quad (6)$$

Описанный порождающий процесс связывает каждое слово со множеством отметок времени, полученных от разных тем, тогда как исходно каждая отметка времени соответствует некоторому документу из обучающей коллекции и одинакова для всех слов этого документа. Для оценивания параметров модели применяется сэмплирование Гиббса.

3.5 Вероятностные тематические модели, обучаемые с учителем

Модель LDA является важнейшей вехой в истории тематического моделирования и представляет собой метод мягкой кластеризации документов по темам, то есть относится к методам *самообучения* или *обучения без учителя* (unsupervised learning). Задачи и методы, в которых требуется классифицировать

объекты на основе обучающей выборки с известными классификациями объектов, относятся к области *обучения с учителем* (supervised learning).

Модель соответствий (correspondence LDA, Corr-LDA) [49] была исходно предложена для решения задачи аннотирования изображений, когда каждому изображению из обучающей выборки ставится в соответствие некоторое множество слов, и требуется сгенерировать список слов, подходящих к новому, ещё не аннотированному изображению. Модель Corr-LDA является обобщением LDA и основана на смеси гауссовских и мультиномиальных распределений (гауссовские распределения в пространстве признаков изображения, мультиномиальные — в пространстве слов).

Аналогичная вероятностная модель labeled LDA (LLDA) предложена в [50] для мягкой кластеризации генов. В ней каждый ген может относиться более чем к одному кластеру, что даёт более гибкую классификацию генов. Модель LLDA может также включать в себя аннотацию известных генов, что переводит её в класс тематических моделей, обучаемых с учителем.

Модели самообучения имеют ограниченную применимость в тех задачах, где предсказание или классификация является конечной целью. Поэтому в [12] была предложена модель supervised LDA (sLDA) для классификации документов. Каждому обучающему документу соответствует метка класса, и эти метки существенно используются для выявления скрытых тем, имеющих наибольшую предсказательную силу при классификации документов. Модель sLDA использовалась для предсказания рейтингов фильмов по текстам отзывов, а также для предсказания популярности веб-страниц по текстовым описаниям.

Таким образом, в тематических моделях, обучаемых с учителем, совместно с основными текстами используется дополнительные данные о классификациях или рейтингах, привлечение которых позволяет более точно выявлять тематику. Возможно также использование частичного обучения, когда лишь некоторые документы коллекции имеют метки классов.

3.5.1 Тематическая модель, обучаемая с учителем sLDA

В модели sLDA [9] каждому документу соответствует метка, что отличает её от большинства тематических моделей, обучаемых без учителя. В качестве метки можно рассматривать рейтинг, поставленный фильму пользователем. Модель sLDA может работать с метками в любых шкалах — вещественными, порядковыми или номинальными.

На Рис. 10 показана порождающая модель sLDA. Для определённости рассматривается случай вещественных меток $r \in \mathbb{R}$. Главное отличие sLDA от базовой модели LDA (Рис. 4) в том, что появляется скрытая переменная метки R

с математическим ожиданием η и дисперсией σ^2. В порождающем процессе sLDA сначала генерируется документ d, затем по нему генерируется значение r метки r. Таким образом, метка r зависит от частоты тем, которые появлялись в данном документе. Параметры α, Φ, η и σ^2 оцениваются по обучающему корпусу текстов. Для приближённой максимизации правдоподобия в sLDA используется вариационный EM-алгоритм.

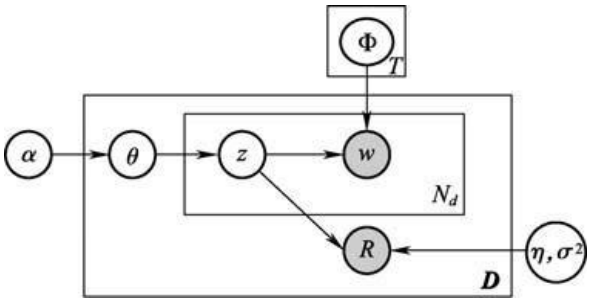

Рис. 10. Модель sLDA.

4 Алгоритмы обучения и байесовского вывода

Задача *оценивания модели* (parameter estimation) заключается в том, чтобы найти значения параметров модели, при которых наблюдаемая обучающая выборка максимально правдоподобна.

Задача *вывода по модели* (inference making) состоит в определении значений скрытых переменных (латентных вероятностей тем) для нового документа, изначально не входившего в состав обучающей выборки.

Если в графической модели корневые вершины соответствуют наблюдаемым документам, и требуется спрогнозировать значения скрытых переменных в листьях, то такой процесс называется *предсказанием* или *нисходящим выводом*. Если же наблюдаемым документам соответствуют листья, и требуется сделать предсказание для скрытых переменных в корнях, то такой процесс называется *диагностированием* или *восходящим выводом*. Формализм байесовских сетей подходит для решения обеих упомянутых задач [51]. В базовых тематических моделях основными характеристиками документа являются совместное распределение Φ в пространстве тем и слов и распределение θ в пространстве тем. Ниже приводится краткий обзор основных алгоритмов оценивания этих параметров и вывода по тематическим моделям.

EM-алгоритм (expectation-maximization) [52] используется для непосредственного точного оценивания параметров Φ и θ. В частности, он применяется для обучения модели PLSA [3]. Недостатками EM-алгоритма является, во-первых, билинейный рост числа параметров по числу слов и числу тем, а также по числу документов и числу тем; во-вторых, большое число

локальных максимумов у функции правдоподобия. Вариационные методы [8,53] являются специальной версией EM-алгоритма для случаев, когда вычисление апостериорного распределения скрытых переменных для заданного документа оказывается вычислительно неэффективным. Эти методы позволяют оценивать узкие нижние и верхние доверительные границы для значений скрытых переменных в наблюдаемом документе. К числу вариационных методов относятся: expectation-propagation [20], вариационные обобщения EM-алгоритма [54] и *вариационный EM-алгоритм* (variational expectation-maximization) [8].

Методы Монте-Карло для марковских цепей (Markov chain Monte Carlo, MCMC) [55,56] широко используются как эффективные приближенные процедуры генерации значений из распределений высоких размерностей. Одной из таких процедур является *сэмплирование Гиббса*, в котором строится марковская цепь, сходящаяся к апостериорному распределению темы z, по которой далее строятся оценки параметров Ф и θ. Такой подход позволяет эффективно находить скрытые темы в больших корпусах текстов [10]. В большинстве случаев он оказывается более эффективным, чем вариационные методы [5,10,19,57]. Детали реализации и применения этих методов могут быть найдены в соответствующих работах.

Вариационный подход и методы MCMC часто используются в рамках EM-парадигмы. При сверхбольших объемах обучающих выборок исключается возможность применения даже таких эффективных методов, как вариационные байесовские (variational Bayes) и сэмплирование Гиббса. Поэтому в [58] предложен эффективный, простой для реализации и существенно более точный метод *свёрнутого вариационного байесовского вывода* (collapsed variational Bayes, CVB) для модели LDA. Для упрощения вычислений в нём используется гауссовская аппроксимация. Основная идея метода состоит в том, чтобы отказаться от предположения независимости параметров от латентных переменных и учесть эту зависимость в явном виде. Благодаря более слабым предположениям факторизации, чем в обычном вариационном байесовском подходе, аппроксимация получается более точной.

5 Критерии качества моделей

Для анализа качества тематических моделей используется несколько критериев. Среди них наиболее важным является *степень неопределенности* (perplexity) [8]. Этот критерий используется также для оценивания обобщающей способности модели на контрольных данных и нахождения оптимального числа различных тем в коллекции документов. Он не требует предварительной категоризации документов. Изначально он использовался в языковых моделях [59], при этом параметры модели оценивались по подмножеству текстов из

коллекции, затем построенная модель применялась для предсказания ранее неизвестных тестовых данных. Чем меньше *степень неопределенности* в тестовой коллекции документов, тем лучше обобщающая способность построенной модели. Для контрольных данных из M документов степень неопределенности определяется равенством (7):

$$Perplexity = \exp\left\{\frac{\sum_{d=1}^{M} \log p(\boldsymbol{w}_d)}{\sum_{d=1}^{M} N_d}\right\}. \quad (7)$$

Энтропия является другим критерием качества, который характеризует чистоту полученного набора тем. Поскольку энтропия — это мера беспорядка в системе, то считается, что чем меньше внутритематическая энтропия, тем лучше.

$$Entropy(topic) = -\sum_z P(z) log_2[P(z)]. \quad (8)$$

Ещё один критерий *строится* на основе *симметричного расстояния Кульбака-Лейблера* (sKL) [9], которое характеризует расстояние между темами. Чем больше межтематические расстояния, тем лучше.

$$sKL(\text{тема } i, \text{тема } j) = \sum_{z=1}^{T} \left[\theta_{iz} \log \frac{\theta_{iz}}{\theta_{jz}} + \theta_{jz} \log \frac{\theta_{jz}}{\theta_{iz}}\right]. \quad (9)$$

Меньшие значения энтропии или большие значения sKL-расстояния свидетельствуют о большей сконцентрированности тем, лучшей обобщающей способности модели и, как следствие, более эффективном ранжировании объектов при информационном поиске. В [34] это показано на примере задачи поиска экспертов.

Степень неопределенности, энтропия и sKL-расстояние могут быть использованы для улучшения обобщающей способности тематических моделей на контрольных данных и, соответственно, генерации лучших наборов тем (кластеров). Однако при оценивании качества ранжирования в информационном поиске эти критерии не являются статистически значимыми [60]. В работе [34] показано, как низкая степень неопределенности влияет на ранжирование объектов при информационном поиске.

Предварительная разметка данных (классификация или категоризация документов) позволяет оценивать качество моделей с помощью стандартных в информационном поиске характеристик *точности* (precision) и *полноты* (recall) [59], как было показано в [60] для задачи поиска экспертов:

$$Precision = \frac{\text{Правильные ответы}}{\text{Полученные ответы}}. \quad (10)$$

$$Recall = \frac{\text{Правильные ответы}}{\text{Максимальное возможное число правильных ответов}}. \quad (11)$$

6 Применения

С помощью тематических моделей решаются разнообразные задачи анализа коллекций текстовых документов, в частности, задачи выявления тем, классификации и индексирования документов, выявления взаимосвязей между объектами, обнаружения тематических трендов и выявления сообществ. Таблица 5 содержит краткую сводку приложений тематических моделей в различных областях.

Выявление тематики — это задача построения списка латентных тем, характеризующего документ в дополнение к его названию или явно указанному списку ключевых слов. Модель мешка слов позволяет обнаруживать неявные взаимосвязи между словами с учетом полисемии. Методы выявления тем [8,16–19,50] направлены на поиск неявных зависимостей и понимание семантики документов. Некоторые методы учитывают марковские зависимости [37,38,41], тогда как другие [27] основаны на выявлении корреляций между темами.

Классификация текстов — это задача разделения документов на два или более взаимно исключающих класса. При индексировании документов задача заключается в том, чтобы найти документы, наиболее соответствующие запросу и ранжировать их в порядке близости к запросу. Модели, предложенные в [3,8,19,28,36,49] оказались достаточно эффективными в задачах классификации и индексирования текстовых коллекций из разных предметных областей. Эти модели выявляют структуру коллекции документов, основываясь на восстановлении скрытых тематических связей между словами, тогда как стандартные методы классификации и кластеризации в основном используют функции расстояния, наподобие евклидовой метрики. В результате они оказываются неспособны в полной мере учесть семантику документов.

Таблица 5. Сводка приложений метода DPTM

Модели	название	Метод оценивания	Прикладная область	Набор(ы) данных
PLSA	BDPTMs	EM	Topic Discovery, Ranking (automatic document indexing) Document Classification, Collaborative Filtering	LOB corpus, MED abstract dataset, CRAN abstracts dataset, CACM abstracts dataset, CISI abstracts dataset
A Joint Probabilistic Model	IrCDPTMs	EM	Document Classification, Relationship between Topics and Links	Webkb web pages dataset (http://www.cs.cmu.edu/~webkb/), Cora abstracts dataset (http://www.cora.justresearch.com)
A probabilistic Approach	BDPTMs	Gibbs Sampling	Topic Discovery (semantics of words)	TASA corpus “a collection of children reading”

LDA	BDPTMs	Variational EM	Topic Discovery, Document Classification, Collaborative Filtering	TREC AP newswire articles corpus, Reuters news articles dataset (http://www.daviddlewis.com/resources/testcollections /reuters21578/), C Elegants Literature (http://elegans.swmed.edu/wli/cgcbib), EachMovie collaborative filtering dataset
A Topic Model	BDPTMs	Gibbs Sampling	Topic Discovery (semantics of words)	TASA corpus "a collection of children reading"
Corr-LDA	SuDPTMs	Variational EM	Automating Annotation, Text-based Image Retrieval	Corel images and caption dataset
discrete (PCA)		Gibbs Sampling	Text classification, Information Retrieval	20 Newsgroup dataset (http://www.ai.mit.edu/_jrennie/20Newsgroups/), Reuters news articles dataset (http://www.daviddlewis.com/resources/testcollections /reuters21578/)
Mixed-Membership Models	IrCDPTMs	EM	Topic Discovery, Document Classification	PNAS scientific articles dataset (http://www.pnas.org)
Author-Topic Model	IrCDPTMs	Gibbs Sampling	Entities and Topics Correlations, Topics Evolution over Time	Cite seer dataset (http://citeseer.ist.psu.edu/oai.html)
ART Model	IrCDPTMs	Gibbs Sampling	Topic and Role Discovery	Enron email dataset (http://www.cs.cmu.edu/~enron/), Researchers email achieve
A Composite Model (HMM-LDA)	IaCDPTMs	Gibbs Sampling	Document Classification, Part-of-Speech Tagging	Brown and TASA corpus "a collection of children reading" datasets, NIPS00-12 Proceedings dataset (www.cs.toronto.edu/~roweis/data.html)
LLDA Model	SuDPTMs	Variational EM	Topic Discovery	Microarray dataset (http://genomics.lbl.gov/~patrickf/llda.html)
CTM	IrCDPTMs	Variational EM	Topics Correlations	JSTOR science articles dataset (http://www.jstor.org)
DTM	TDPTMs	Variational Kalman Filtering	Topics Evolution over Time	JSTOR science articles dataset (http://www.jstor.org)
Statistical Entity-Topic Models	IrCDPTMs	Gibbs Sampling	Entities and Topics Correlations	New York Times dataset (http://www.ldc.upenn.edu), Foreign broadcast information service FBIS dataset (http://www.fbis.gov)
Bigram Topic Model	IaCDPTMs	Gibbs EM	Topic Discovery	Psychological review abstracts dataset (http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm), Newsgroup dataset (http://people.csail.mit.edu/_jrennie/20Newsgroups/)
PAM	IrCDPTMs	Gibbs Sampling	Super and Sub Topic Discovery, Document Classification	NIPS00-12 Proceedings dataset (www.cs.toronto.edu/~roweis/data.html) , Newsgroup dataset (http://www.cs.cmu.edu/~textlearning/), Rexa research paper search engine (http://Rexa.info)
TOT Model	TDPTMs	Gibbs Sampling	Topics Evolution over Time	State of the Union Addresses dataset (http://www.gutenberg.org/dirs/etext04/suall11.txt), Researchers Email Achieve, NIPS00-12 Proceedings dataset (www.cs.toronto.edu/~roweis/data.html)
Continues-Time Model	TDPTMs	Gibbs Sampling	Topics Evolution over Time and their Correlations	Rexa research paper search engine (http://Rexa.info)
CPLSA	IrCDPTMs	EM	Temporal (Entities-Topic) Correlations, Topics Evolution over Time, Event Impact Analysis	Abstracts of 282 papers of two Data Mining researchers, from ACM Digital library, MSN Space documents, Abstracts of 28 years' SIGIR conferences from ACM Digital Library
HTMM	IaCDPTMs	EM and Forward-backward algorithm	Topic Discovery	NIPS00-12 Proceedings dataset (www.cs.toronto.edu/~roweis/data.html), used dataset (http://www.cs.huji.ac.il/~amitg/htmm.html)

MTTM	TDPTMs	Variational EM	Topics Evolution over Time	JSTOR science articles dataset (http://www.jstor.org)
sLDA Model	SuDPTMs	Variational EM	Ranking Movies and Web Pages	News paper movie reviews dataset (http://www.cs.cornell.edu/people/pabo/movie review-data/), Digg Links (digg.com)
Citation Influence Model	IrCDPTMs	Gibbs Sampling	Citation Influence	Cite seer dataset (http://citeseer.ist.psu.edu/oai.html)
GWN-LDA Model	IrCDPTMs	Gibbs Sampling	Entities and Topics Correlations	NanoSci articles dataset (2000-2006) taken from (http://scientific.thomson.com/products/sci/), Cite seer dataset (http://citeseer.ist.psu.edu/oai.html)
TNG Model	IaCDPTMs	Gibbs Sampling	Topic Discovery, Information Retrieval	TREC dataset, NIPS00-12 Proceedings dataset (www.cs.toronto.edu/~roweis/data.html),
Link-PLSA-LDA	IrCDPTMs	Variational EM	Blogs Influence	Nielsen Buzz metrics blogs postings dataset (http://www.nielsenbuzzmetrics.com)
cDTM	TDPTMs	Variational Kalman Filtering	Topics Evolution over Continuous Time	TREC-1 AP newswire articles corpus, “Election 08” dataset (digg.com)
LTHM	IrCDPTMs	EM	Relationship between Topics and Links	Webkb web pages dataset (http://www.cs.huji.ac.il/~amitg/lthm.html), Wikipedia (http://www.cs.cmu.edu/~webkb/)
TAT	TDPTMs	Gibbs Sampling	Temporal Authors Interests and Correlations	Computer science research papers taken from http://www.informatik.uni-trier.de/~ley/db/
ACT	IrCDPTMs	Gibbs Sampling	Expertise Search in Academics Social Network	Computer science research papers taken from http://www.arnetminer.org/
STMS	IrCDPTMs	Gibbs Sampling	Expert Finding	Computer science research papers taken from http://www.informatik.uni-trier.de/~ley/db/
GLDA	IrCDPTMs	Gibbs Sampling	Conference Mining	Computer science research papers taken from http://www.informatik.uni-trier.de/~ley/db/

Между объектами (авторами, организациями, конференциями, и т.д.) существуют определённые взаимосвязи, на основе которых формируются научные сообщества и социальные сети, причём косвенная информация об этих взаимосвязях содержится в структурах документов и в ссылках. Выявление этой информации является нетривиальной задачей. Тематические модели, предложенные в [14,22,29,30,33,35], позволяют выявлять взаимосвязи объектов и сообществ с тематикой документов.

Модель ART [26] позволяет выявлять роли субъектов в организационной структуре предприятия, с учетом выполняемых ими работ, на основе сообщений электронной почты и данных об отправителе и получателе. В более ранних исследованиях подобного рода системы ролей выявлялись без анализа семантики текстовых сообщений. Модель Citation Influence [31] строит отношения между документами на основе графа цитирования научных публикаций. Эта модель, как и ART, основана на совместном учёте как самих текстов, так и связей между текстами. Она неприменима к таким коллекциям, в которых нет информации о цитировании, например, к корпусу научных статей JSTOR.

Тематическая модель, предложенная в [34], позволяет совместно учитывать семантическую и временную информацию для поиска экспертов в

академических социальных сетях. В модели STMS фактор влиятельности конференций и фактор времени моделируются совместно, поскольку оба они имеют большое значение для поиска экспертов. Авторы, публикующиеся в журналах и конференциях высокого ранга, с большей вероятностью являются лучшими экспертами, но лишь на определённых промежутках времени. Если рассматривать каждый год в отдельности без моделирования процессов во времени, то может возникать проблема поиска соответствий между темами, как следствие перестановочности тем.

Выявление временных тенденций или эволюции тем — это интересная задача, дающая представление о развитии направлений исследований внутри научного сообщества или, скажем, вкусов потребителей в отношении еды или одежды. При исследовании научных сообществ строились динамические модели отношений автор–тема и анализировалось влияние на них определённых событий [39]. Проблема эволюции тематик во времени подробно изучалась в [14,43—45], где было предложено несколько решений. Модель с непрерывным временем [15] позволяет выявлять эволюцию тем во времени. Модель TAT позволяет ранжировать авторов разных лет по интересам [48]. Эти модели используют информацию о годе и/или месяце публикаций. Эти модели использовались также для решения задач оценивания влияния блогов [25], автоматического аннотирования документов [61] и коллекций документов [62], спам-фильтрации [63].

Кроме моделирования текстов, тематические модели применяются также для кластеризации изображений [64], распознавания объектов [65,66], распознавания рукописного текста [67], а также в других областях компьютерного зрения. Кроме того, они применяются в биоинформатике для анализа данных ДНК-микрочипов, клеточных данных, данных о нуклеотидных и полипептидных последовательностях [50].

7 Открытые проблемы и направления исследований

В этой части рассматриваются направления дальнейших исследований и открытые проблемы в области тематических моделей. Задачи моделирования текстовых коллекций можно разделить на четыре типа.

Первый тип задач заключается в выявлении скрытых структур с помощью марковских цепей в контекстах документов, что должно улучшать выделение тем в текстовой коллекции. Согласно [37], внедрение скрытой марковской модели в модель LDA связано с другими расширениями LDA, такими как в [14] и [36]. Таким образом, значительный интерес представляет соединение

различных расширений для создания наиболее полной модели текстовой коллекции. В частности, в автор-тематической модели [14] не учитывается стилистика авторского письма. Удачно соединив стилистические особенности текста с его тематикой, можно было бы достичь более лучших рещультатов при классификации авторов. Марковские зависимости могут быть очень полезны, но они требует глубокого понимания статистических методов обработки естественного языка для создания новых высокоэффективных решений.

Второй тип задач связан с использованием явных связей (например, цитат) между документами для лучшего определения отношений между ними, исследователями и социальными сетями. Несомненно в [25,37] получен достигнут значительный прогресс в использовании корреляций между документами, однако метод Link-PLSA-LDA [25] сталкивается с проблемой роста числа параметров, как и PLSA [3], а метод HTMM [37] не учитывает текстовое содержимое связанных документов. Эти проблемы, возможно, будут решены в новой модели с лучшим средством обработки связей между документами.

Третий тип задач связан с совместным использованием содержимого документов, данных о связанных с документом объектах и информации о времени создания документов для поиска временных трендов. Важно непрерывное временное моделирование, и одно из последних решений, явно игнорирующее синтаксические зависимости, предложено [43]. Кроме того, важно выявлять многозначность (полисемию) слов, при этом учитывать дискретность времени, которое вынужденно моделируется как непрерывное. Модель непрерывного времени [15] учитывает временные и синтаксические зависимости, используя ориентированные ациклические графы, что не вполне адекватно реальности. Необходима новая модель, которая учитывала бы временные и синтаксические зависимости, но была бы свободна от ограничений, налагаемых использованием ориентированного ациклического графа. Учёт явных связей между документами посредством корреляций между темами и временем создания документов мог бы быть хорошим продвижением вперёд.

Четвёртый тип задач связан с использованием классифицированных коллекций документов наряду с неявной информацией, содержащейся в контексте документов, для улучшения точности моделей. Модель sLDA [12] является важным шагом в этом направлении, но она не позволяет извлекать семантическую информацию из синтаксиса. При анализе временных трендов время создания также должно быть представлено в расширенной модели.

С практической точки зрения, использование моделей смеси распределений при решении вышеописанных четырёх проблем может приводить к практичным

решениям, приносящим превосходные результаты, как это случалось в прошлом. Однако для этого требуется более глубокое понимание статистических моделей языка. Проблема в том, чтобы выбрать и соединить различные подходы, объединив их преимущества в одном компактном решении.

С теоретической точки зрения, для построения тематических моделей по большим корпусам текстов необходимо иметь высокоэффективные методы оценивания параметров и байесовского вывода. Предложенные в [58] свёрнутые вариационные алгоритмы Байесовского вывода существенно повышают вычислительную эффективность LDA. Вопросы сходимости алгоритмов и выбора гиперпараметров для вероятностных тематических моделей до сих пор нуждаются в более детальной проработке. Вопросы разработки специализированных моделей, таких как ART [26] или Citation Influence [31], и интерпретации результатов, полученных различными моделями, также требуют дальнейшего изучения.

8 Выводы

Итак, в обзоре рассмотрены основные вероятностные тематические модели и проведена их классификация на пять категорий в соответствии с их функциональными возможностями. Рассмотрены методы оценивания параметров и извлечения латентных тем, проблемы полисемии, оценивания качества и производительности тематических моделей. Рассмотрены некоторые приложения тематических моделей к анализу текстовых коллекций. Представлено текущее состояние исследований и результаты последних десяти лет в области вероятностных тематических моделей и их приложений.

Литература

1. Popescul A, Flake G W, Lawrence S, Ungar L H, Giles C L. Clustering and identifying temporal trends in document databases. IEEE ADL, 2000, 173–182

2. McCallum A, Nigam K, Ungar L H. Efficient clustering of high-dimensional data sets with application to reference matching. In: Proceedings of the 6th ACM SIGKDD, 2000, 169–178

3. Hofmann T. Probabilistic latent semantic analysis. In: Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, July 30-August 1, 1999

4. Steyvers M, Griffiths T. Probabilistic topic models. In: Landauer T, Mcnamara D, Dennis S, Kintsch W (Eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2007

5. Heinrich G. Parameter Estimation for Text Analysis. Technical report, Version 2, February 2008

6. Smolensky P. Information processing in dynamical systems: foundations of harmony theory. In: Rumehart D E, McClelland J L (Eds), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations. McGraw- Hill, New York, 1986

7. Welling M, Rosen-Zvi M, Hinton G. Exponential family harmoniums with an application to information retrieval. In: Advances in Neural Information Processing Systems (NIPS). Cambridge, MA, MIT Press, 2004

8. Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993–1022

9. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI), Banff, Canada, July 7–11, 2004

10. Griffiths T L, Steyvers M. Finding scientific topics. In: Proceedings of the National Academy of Sciences. USA, 2004, 101: 5228–5235

11. Teh Y W, Jordan M I, Beal M J, Blei D M. Hierarchical Dirichlet Processes. Technical Report 653, Department of Statistics, UC Berkeley, 2004

12. Blei D M, McAuliffe J. Supervised topic models. In: Advances in Neural Information Processing Systems (NIPS) 21. Cambridge, MA, MIT Press, 2007, 121–128

13. Buntine W L. Operations for learning with graphical models. Journal of Artificial Intelligence Research, 1994, 2: 159–225

14. Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T. Probabilistic author-topic models for information discovery. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, Washington, August 22–25, 2004

15. Wang X, Li W, McCallum A. A continuous-time model of topic co-occurrence trends. In: AAAI Workshop on Event Detection. Boston, Massachusetts, USA, July 16–20, 2006

16. Nigam K, McCallum A K, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Journal of Machine Learning, 2000, 39(2–3): 103–134

17. Griffiths T L, Steyvers M. A probabilistic approach to semantic representation. In: Proceedings of the 24th Conference of the Cognitive Science Society. USA, 2002

18. Griffiths T L, Steyvers M. Prediction and semantic association. In: Advances in Neural Information Processing Systems (NIPS) 15. Cambridge, MA, MIT Press, 2003

19. Wray L, Buntine, Jakulin A. Applying discrete PCA in data analysis. In: Proceedings of 20th Conference on Uncertainty in Artificial Intelligence (UAI), Banff, Canada, July 7–11, 2004, 59– 66

20. Minka T, Lafferty J. Expectation-propagation for the generative aspect model. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI), Alberta, Canada, August 1–4, 2002, 352–359

21. Hofmann T, Puzicha J, Jordan M I. Learning from dyadic data. In: Advances in Neural Information Processing Systems (NIPS) 11. Cambridge, MA, MIT Press, 1999

22. Cohn D, Hofmann T. The missing link- a probabilistic model of document content and hypertext connectivity. In: Advances in Neural Information Processing Systems (NIPS) 13. Cambridge, MA, MIT Press, 2001

23. Blei D M, Moreno P J. Topic segmentation with an aspect hidden Markov model. In: Proceedings of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans. LA USA, September 9-13, 2001, 343–348

24. Erosheva E, Fienberg S, Lafferty J. Mixed-membership models of scientific publications. In: Proceedings of the National Academy of Sciences, USA, 2004, 101: 5220–5227

25. Nallapati R, Cohen W. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In: Proceedings of International Conference for Weblogs and Social Media, Seattle, Washington, USA, March 30-April 2, 2008

26. McCallum A, Corrada-Emmanuel A, Wang X. The author-recipient- topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. Technical Report UM-CS-2004-096, 2004

27. Blei D M, Lafferty J. Correlated topic models. In: Advances in Neural Information Processing Systems (NIPS) 18. Cambridge, MA, MIT Press, 2006, 147–154

28. Li W, McCallum A. Pachinko allocation: Dag-structured mixture models of topic correlations. In: Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, June 25-29, 2006, 577–584

29. Newman D, Chemudugunta C, Smyth P, Steyvers M. Statistical entity-topic models. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, August 20–23, 2006, 680–686

30. Zhang H, Giles C L, Foley H C, Yen J. Probabilistic community discovery using hierarchical latent Gaussian mixture model. In: Proceedings of 22nd AAAI Conference on Artificial Intelligence, Vancouver, British Columbia, Canada, July 22–26, 2007, 663–668

31. Dietz L, Bickel S, Scheffer T. Unsupervised prediction of citation influences. In: Proceedings of 24th International Conference on Machine Learning (ICML), Corvallis, Oregon, USA, June 20–24, 2007

32. Gruber A, Rosen-Zvi M, Weiss Y. Latent topic models for hypertext. In: Proceedings of Uncertainty in Artificial Intelligence (UAI), Helsinki, Finland, July 9–12, 2008

33. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z. ArnetMiner: extraction and mining of academic social networks. In: Proceedings of ACM SIGKDD, 2008

34. Daud A, Li J, Zhu L, Muhammad F. A generalized topic modeling approach for maven search. In: Proceedings of International Asia- Pacific Web Conference and Web-Age Information Management (APWEB-WAIM), Suzhou, China, 2009

35. Daud A, Li J, Zhu L, Muhammad F. Conference mining via generalized topic modeling. In: Proceedings of European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML PKDD), Bled, Slovenia, 2009

36. Griffiths T L, Steyvers M, Blei D M, Tenenbaum J B. Integrating topics and syntax. In: Advances in Neural Information Processing Systems (NIPS) 17. Cambridge, MA, MIT Press, 2005, 537–544

37. Gruber A, Rosen-Zvi M, Weiss Y. Hidden topic Markov models. In: Proceedings of Artificial Intelligence and Statistics (AISTATS), San Juan, Puerto Rico, USA, March 21–24, 2007

38. Wallach J M. Topic modeling: Beyond bag-of-words. In: Proceedings of 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, USA, June 25–29, 2006

39. Mei Q, Zhai C X. A mixture model for contextual text mining. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, August 20–23, 2006, 649–655

40. Deerwester S, Dumais S T, Furnas G W, Landauer T K, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41(6): 391–407

41. Wang X, McCallum A, Wei X. Topical N-grams: phrase and topic discovery, with an application to information retrieval. In: Proceedings of the 7th IEEE International Conference on Data Mining (ICDM), Omaha NE, USA, October 28–31, 2007

42. Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. In: Proceedings of the IEEE, 1989, 77(2): 257–286

43. Blei D M, Lafferty J. Dynamic topic models. In: Proceedings of 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, USA, June 25–29, 2006

44. Nallapati R, Cohen W, Ditmore S, Lafferty J, Ung K. Multiscale topic tomography. In: Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12–15, 2007

45. Wang C, Blei M D, Heckerman D. Continuous time dynamic topic models. In: Proceedings of Uncertainty in Artificial Intelligence (UAI), Helsinki, Finland, July 9–12, 2008

46. Uhlenbeck G E, Ornstein L S. On the theory of Brownian motion. Physics Reviews, 1930, 36: 823–841

47. Wang X, McCallum A. Topics over time: A non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, August 20–23, 2006

48. Daud A, Li J, Zhu L, Muhammad F. Exploiting temporal authors interests via temporal-author-topic modeling. In: Proceedings of 5th International Conference on Advance Data Mining and Applications (ADMA), Beijing, China, 2009

49. Blei D M, Jordan M. Modeling annotated data. In: Proceedings of 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, July 28-August 1, 2003, 127–134

50. Flaherty P, Giaever G, Kumm J, Jordan M, Arkin A. A latent variable model for chemogenomic profiling. Bioinformatics, 2005, 21(15): 3286–3293

51. Murphy K. An Introduction to Graphical Models. Technical report, University of California, Berkeley, May 2001

52. Bilmes J A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov modals. Berkeley, ICSI TR-97-021, 1997

53. Jordan M I, Ghahramani Z, Jaakkola T S, Saul L K. An introduction to variational methods for graphical models. In: Jordan M (Eds), Learning in Graphical Models. MIT Press, 1998

54. Buntine W. Variational Extensions to EM and Multinomial PCA. In: Elomaa T et al. (Eds.): ECML, LNAI 2430, Springer-Verlag, Berlin, 2002, 23–34

55. Gilks W R, Richardson S, Spiegelhalter D J. Markov Chain Monte Carlo in Practice. London: Chapman & Hall, 1996

56. Andrieu C, Freitas N D, Doucet A, Jordan M. An introduction to MCMC for machine learning. Journal of Machine Learning, 2003, 50: 5–43

57. Erosheva E A. Grade of membership and latent structure models with applications to disability survey data. Unpublished doctoral dissertation, Department of Statistics, Carnegie Mellon University, 2002

58. Teh Y W, Newman D, Wellingm M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: Advances in Neural Information Processing Systems (NIPS). Cambridge, MA, MIT Press, 2006

59. Azzopardi L, Girolami M, Risjbergen K V. Investigating the relationship between language model perplexity and IR precision-recall measures. In: Proceedings of the 26th ACM SIGIR, Toronto, Canada, 2003

60. Zhang J, Tang J, Liu L, Li J. A mixture model for expert finding. In: Proceedings of the PAKDD, Washio T et al. (Eds). LNAI, 2008, 5012: 466–478

61. Chang Y L, Chien J T. Latent Dirichlet learning for document summarization. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009

62. Arora R, Ravindran B. Latent Dirichlet allocation based multidocument summarization. In: Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Rext Data, 2008

63. Bíró I, Szabó J, Benczúr A A. Latent Dirichlet allocation in web spam filtering. In: Proceedings of the Adversarial Information Retrieval on the Web (AIRWeb'08), 2008

64. Elango P K, Jayaraman K. Clustering images using the latent Dirichlet allocation model, 2005

65. Wang Y, Mori G. Human action recognition by semi-latent topic models. IEEE Transactions on Pattern Analysis and Machine Intelligence Special Issue on Probabilistic Graphical Models in Computer Vision (T-PAMI), 2009

66. Wang Y, Sabzmeydani P, Mori G. Semi-latent Dirichlet allocation: A hierarchical model for human action recognition. In: 2nd Workshop on Human Motion Understanding, Modeling, Capture and Animation (ICCV), 2007

67. Rath T M, Lavrenko V, Manmatha R. A Statistical Approach to Retrieving Historical Manuscript Images Without Recognition. Technical Report, 2003