

复杂文本网数据的主题建模进展

曾 嘉^{1),2)} 严建峰¹⁾ 龚声蓉¹⁾

¹⁾(苏州大学计算机科学与技术学院 江苏 苏州 215006)

²⁾(复旦大学上海智能信息处理重点实验室 上海 200433)

摘 要 文中介绍了大规模文本网数据的主题建模研究的特点和近年来的重要进展. 主题建模方法吸引了世界范围的广泛兴趣, 并且促进了许多重要的数据挖掘、计算机视觉和计算生物应用系统的发展, 包括文本自动摘要、信息检索、信息推荐、主题检测和追踪、自然场景理解、人体动作识别以及微阵列基因表达分析等. 文中重点介绍文本网数据的4个主要特点以及对应的主题模型. 文本网数据含有动态、高阶、多通路及分布式的结构, 而之前的主题模型仅对部分的结构进行建模. 而文中讨论了在三维马尔可夫模型的框架下统一对文本网数据的4个结构特点进行建模, 并分析了结合三维马尔可夫模型和二型模糊系统对分布式单词计算和主题建模应用的可能性. 除了对文本网数据的结构建模之外, 还讨论了一些对三维马尔可夫模型能量最小化的机器学习算法.

关键词 主题模型; 文本网数据; 数据挖掘; 单词计算; 二型模糊系统

中图法分类号 TP391

DOI号: 10.3724/SP.J.1016.2012.02431

Advances in Topic Models for Complex Document Network Data

ZENG Jia^{1),2)} YAN Jian-Feng¹⁾ GONG Sheng-Rong¹⁾

¹⁾(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006)

²⁾(Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433)

Abstract This paper reviews important advances that have been made in the past decade for topic modeling of large-scale document network data. Interest in topic modeling is worldwide and touches a number of practical text mining, computer vision and computational biology systems that are important in text summarization, information retrieval, information recommendation, topic detection and tracking, natural scene understanding, human motion categorization and microarray gene expression analysis. The main focus of this review is on the recent advances of topic modeling techniques for document network data. We introduce the four major characteristics of document network data and the current state-of-the-art topic models, with descriptions of what they are, what has been accomplished, and what remains to be done. Document network data contain dynamic, higher-order, multiplex, and distributed structures. Prior efforts on topic models focus on modeling parts of these structures for topic detection and tracking. To handle all document network structures, we discuss a three-dimensional Markov model that solves dynamic, higher-order, multiplex and distributed structures within a unified framework. In addition, we also discuss the integration of three-dimensional Markov models with type-2 fuzzy logic systems for distributed computing with words. Besides document network structure modeling, we also discuss the inference and parameter estimation method in terms of energy minimization for three-dimensional Markov models.

Keywords topic models; document network data; data mining; computing with words; type-2 fuzzy logic systems

收稿日期: 2011-05-25; 最终修改稿收到日期: 2011-11-11. 本课题得到国家自然科学基金(61003154, 61003259, 61170124)、复旦大学上海智能信息处理重点实验室课题(HIPL-2010-009)以及江苏省高校自然科学研究重大项目基金(12KJA520004)资助. 曾 嘉, 男, 1980年生, 博士, 教授, 博士生导师, 研究领域为模式识别、数据挖掘、机器学习、计算生物学. 严建峰(通信作者), 男, 1978年生, 博士, 研究方向为模式识别、传感器网络. E-mail: yanjf@suda.edu.cn. 龚声蓉, 男, 1966年生, 博士, 教授, 研究领域为模式识别、图像与视频处理.

1 引 言

随着互联网和相关技术的快速发展,信息呈现爆炸式的增长,对大规模电子文档(例如海量科技文献、政府文件、社区网页和新闻存档)进行自动分析的需求与日俱增. 有效地组织、管理、理解和摘要大规模文档,才能充分发现并利用文档中潜在的知识,使得人们更加便利、准确地获取所需信息. 文本网数据(document network data)涉及文本相互引用网(citation networks)、文本共标签网(co-tag network)、超链接网页(hyperlinked Web pages)、社交网(social networks)等. 文本网数据中的节点通常表示文本或者图像,节点之间的链接包含引用、超链接、共标签和文本内容相似度等多种类型. 网数据是目前应用统计和机器学习的一个重要研究对象,对网数据的建模和分析方法凝聚了跨学科的努力,在自然科学、社会科学以及信息科学有广泛的应用^[1-3]. 统计分析大规模文本网数据不但能提供有用的预测模型,同时能够提供对大量数据的统计刻画. 预测模型能够同时预测网数据中节点的内容(例如文本的单词)以及节点和节点的链接关系(例如社交网中的朋友关系、科技论文的引用关系以及网页之间的链接关系等). 统计刻画可以直接揭示隐藏在网数据下面的聚类或者社区结构,因此有助于发现潜在的知识层次.

主题模型(topic models)是一种全新的统计模型,它利用快速的非监督机器学习算法来自动组织、理解并且摘要文本的主要内容,是解决这类问题的新方法. 主题模型是一种用于分析大规模文本的层次贝叶斯模型(hierarchical Bayesian models),它利用快速的非监督机器学习算法从高维稀疏的单词数据中提取低维的数学表示,从而对文本单词进行聚类,是当前数据挖掘和机器学习领域关注的热点之一^[4-15]. 探索高精度的主题模型及其快速学习算法有两个实际意义. 首先,高精度的主题模型可以解决许多传统文本或图像分析中的关键应用问题,例如文本

主题检测及追踪(topic detection and tracking)^[16]、自动文本摘要(text summarization)^[17]、信息检索(information retrieval)^[18]等. 其次,高精度的主题模型可以自动从大量的科技文献中发现并且预测潜在的重要的研究方向,节省研究人员宝贵的前期调研时间和经费,将在生物、物理和化学等研究领域发挥重大作用^[19]. 然而,传统的主题模型 Latent Dirichlet Allocation (LDA)^[7,10] 仅关注文本之间的内容相似度信息,却忽视了文本之间存在的大量外在联系,例如引用(citation)、超链接(hyperlink)和共标签(co-tag)等,从而会影响抽取的主题在语义上的精度和稳定度. 因此,本文将在分析文本网数据的主要特点的基础上,对当前主题模型和机器学习策略的发展进行综述,并讨论设计新型主题模型以及开发相应的快速机器学习算法的新思路,以期刻画并利用文本网数据丰富的结构特点来进一步提高主题建模的精度和稳定度.

2 复杂文本网数据的 4 个特点

复杂文本网数据主要有以下 4 个特点: 动态(dynamic)、高阶(higher-order)、多通路(multiplex)和分布式(distributed). 这 4 个特点对改进基于 LDA 的主题模型有如下 4 点意义:

(1) 文本网数据具有时间戳来表示文本发布的时间,这样主题会随着时间动态演化(evolution),通过网络结构相互影响,或者产生新的主题,或者消亡旧的主题. 对文本网数据的动态主题建模可以应用于新主题的预测或者在线主题爆发(online bursty topics)的检测. 图 1 是一个文本网数据在 2007~2009 年的变化的例子,其中黑色小人表示文本的作者. 可观察到每一个时间片段的网数据链接结构不同,因此需要刻画其中动态结构的细微变化,来反映主题分布随时间的演化或者依赖关系. 同时,刻画时间对主题分布的影响可以更好地建立预测模型,动态地揭示未来潜在的重要的研究方向,节省科研人员宝贵的文献调研时间.

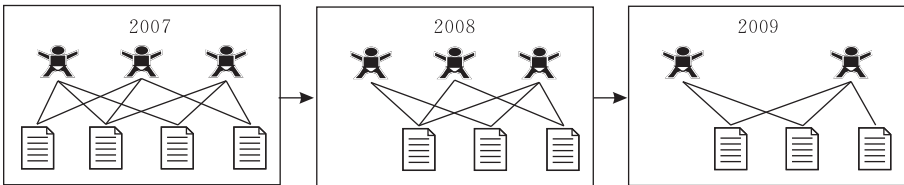


图 1 随时间演化的文本网数据示意图

(2) 共标签文本网数据 (tag-document network data) 广泛存在大量特定的高阶结构。标签通常可以注释文本某一方面的属性。例如“作者标签”描述了文本的作者属性,“时间标签”表明了文本何时发表。如果用视觉单词 (visual words) 来表示图像,那么图像也可以像文本一样具有层次结构。用户一般用标签手工注释图像的局部内容,例如标签“人物”、“建筑物”或者“树木”等。通常,共同标签的文本之间的主题结构不仅仅局限于二阶 (2-order or pairwise) 关系即两个文本之间存在链接,还广泛出现三阶甚至四阶的高阶关系,即 3 个文本或者 4 个文本中,每两个文本之间都共标签。对文本网数据的高阶结构建模可以描述更加丰富特定的主题结构特征,从而可以抽取不同层次的主题。图 2 显示了普遍存在于标签文本网数据的高阶结构,(a) 文本的作者作为标签;(b) 图片的注释作为标签;(c) 标签网数据的二分图表示;(d) 可以将文本 δ_2 的特定主题从文本并集的主题 $\{\delta_1, \delta_3, \delta_4\}$ 区分开的高阶结构。通常,一个文本拥有多个标签。通过多个标签对文本进行标注,多个文本形成了有别于二阶关系的高阶关系。可

以用二分图 (如图 2(c) 显示) 来抽象表示共标签文本网数据。高阶关系一般对主题建模有积极的影响。例如图 2(d) 显示,作者 $\{\gamma_1, \gamma_2\}$ 合作写了以“机器学习”为主题的文本 δ_1 ,表示为两个圆圈 $\{\gamma_1, \gamma_2\}$ 的交集。类似的,作者 $\{\gamma_1, \gamma_3\}$ 合作写了以“计算机视觉”为主题的文本 δ_3 ,作者 $\{\gamma_2, \gamma_3\}$ 合作写了以“数据挖掘”为主题的文本 δ_4 。他们 3 个人共同合作写了文本 δ_2 ,表示为 3 个圆圈 $\{\gamma_1, \gamma_2, \gamma_3\}$ 的交集。理论上,如果简单地将多个标签组成的高阶关系 $\{\gamma_1, \gamma_2, \gamma_3\}$ 分解表示成为 3 个二阶关系的并集 $\{\{\gamma_1, \gamma_2\}, \{\gamma_1, \gamma_3\}, \{\gamma_2, \gamma_3\}\}$,就可以得出一个结论:文本 δ_2 是关于 3 个主题“机器学习”、“计算机视觉”、“数据挖掘”的并集。然而,实际上文本 δ_2 很可能是关于一个全新的主题,例如“计算生物学”,是 3 个圆圈 $\{\gamma_1, \gamma_2, \gamma_3\}$ 的交集。因此,十分有必要对高阶关系建模,用来区分文本 δ_2 不同于文本 $\{\delta_1, \delta_3, \delta_4\}$ 并集的特定主题。除了标签文本网数据,具有超链接的网页以及文本之间的相互引用也存在类似的高阶结构。例如,3 个网页相互之间都存在超链接以及 3 个文本同时引用同一个文本等。

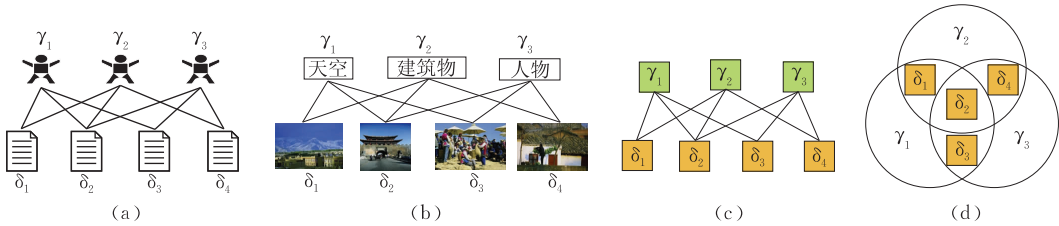


图 2 标签文本网数据中的高阶关系

(3) 文本节点之间通常存在多种链接类型,即从一个节点到另一个节点存在不同类型的通路。融合多种类型的通路信息进行主题聚类可以细致地反映不同通路的网结构对主题建模的影响,从而发现特定关联稳定的主题。图 3 是一个多链接的文本网数据示意图,其中含有两种文本之间常见链接类型:相互引用(箭头)和共同作者(黑色人形图标)。通常,不同链接类型对主题分布影响不同,因此需要分别刻画建模并且有效融合不同通路的影响。其相互引用文本之间通常显示出很强的主题依赖性,例如“数据挖掘”主题常常引用“机器学习”主题的文本。共同作者的文本却往往显示出很强的主题相似性,例如相同的作者在“机器学习”领域常常发表与“机器学习”相关的文本。很显然,在大多数情况下这两种链接对最终主题的形成有不同的影响。因此,需要对网数据的每一种通路分别进行刻画,以便达到最好的主题建模精度和稳定性。在实际应用中,当通路

类型繁多的情况下,如何平衡各个通路对主题分布的影响是一个极具挑战性的工作。目前,尚未有完备的理论支持多通路网数据分析,很多情况都是依靠设计者的先验知识,这往往导致对应主题模型和机器学习算法缺乏良好的泛化能力。

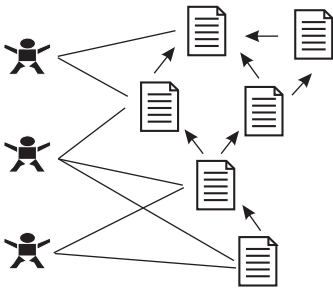


图 3 多通路文本网数据示意图

(4) 大规模文本网数据通常分布在不同的数据服务器上,每一个服务器仅存储部分网数据信息,从而需要高效的信息融合技术对分布式网数据进行主

题建模.

图 4 显示了网数据分布存储在 3 个服务器上, 在每一个服务器上的网数据都只能反映局部的结构信息, 因此要求我们对分布式网数据开发并行的主题学习算法. 因为在每一个服务器上网数据都反映局部信息, 我们需要有效的信息融合技术将不同服务器上的主题分布融合成达到或者接近全局最优的主题分布. 首先, 需要开发并行的分布式主题模型学习算法分别得到每个服务器上的网数据主题分布, 然后利用高效信息融合的技术, 将所有服务器上的主题分布融合, 使得最终的主题分布达到或者接近全局最优.

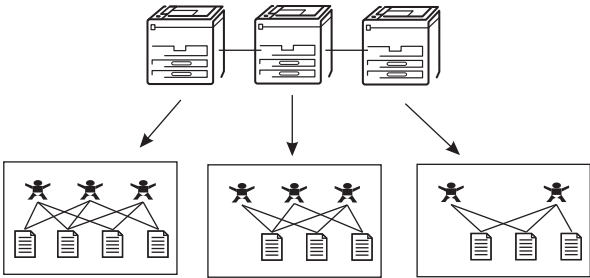


图 4 分布式网数据示意图

综上所述, 根据复杂文本网数据的结构特点来开发新型的主题建模及其相应的机器学习算法对改善主题建模的精度和稳定性有重要意义, 近期在数据挖掘和机器学习领域受到广泛的关注^[9,13,20-25]. 同时, 研究复杂文本网数据的主题建模可以启发和促进其它类型的复杂网数据的聚类分析, 例如基因调控网 (gene regulatory network)、蛋白质互动网 (protein-protein interaction network) 以及交通网 (transportation network) 等. 例如, 研究者利用文本网数据互相引用的特征对相关蛋白质互动网的全文文本进行分类, 从而发现蛋白质互动网的潜在联系^[26].

3 网数据主题建模的研究进展

根据复杂文本网数据动态、高阶、多通路和分布式这 4 个主要特点, 本节将在 LDA 的基础上, 对国内外文本网数据主题建模的研究进展进行分析.

3.1 Latent Dirichlet Allocation(LDA)简介

大规模文本集的复杂性使得人们广泛使用生成模型 (generative models) 将文本表示为一组混合主题 (mixtures of topics). 通常, 一个主题被定义为一个在单词表中所有单词上的多项概率分布, 用来刻

画高度关联的单词聚类 (word clusters), 反映文本特定方面的语义内容. 图 5 给出了目前广泛使用的一类主题模型 LDA^[7,10] 的图模型表示, 其中空心圆圈表示隐藏的随机变量或者参数, 阴影圆圈表示观测变量, 方框则代表重复的次数. 例如, 文本重复 D 次表示整个文本集, 单词在每个文本 d 中重复 I_d 次, 主题分布 ϕ 重复 J 次. LDA 将主题标签 z 分配给文本 d 的每一个单词 w , 这里 d 是文本在整个文本集中的索引 $1 \leq d \leq D$, w 是单词在单词表 (vocabulary) 中的索引 $1 \leq w \leq W$. 通常, 主题标签 $z=j$ 是隐藏的随机变量 (latent variables), 其中 $1 \leq j \leq J$ 是主题标签索引. 当 LDA 将主题标签分配给每一个单词之后, 实际上将所有单词分成了 J 类, 因此主题建模可以被视为一种单词聚类的范例. 理论上, LDA 基于两个参数来分配主题标签: 一个是文本中主题标签的比例 θ_d , 另一个是主题在单词表上的多项分布 ϕ . 这两个参数分别由两个基于超参数 (hyper parameters) α 和 β 的共轭 Dirichlet 分布生成. 实际中, 超参数 α 和 β 起到平滑多项分布参数 θ_d 和 ϕ 的作用. 图 5 显示 LDA 可以通过分享同一个主题分布 ϕ 来表示任意两个文本 $\{d, d'\}$ 之间隐含的联系. 本质上, 如果两个文本包含相似的单词, LDA 则鼓励两个文本具有相似的主题标签配置. 然而, 我们注意到 LDA 仅仅用了文本之间内容的相似程度来交换主题信息, 并未充分利用文本网数据中丰富的外部链接结构, 例如引用或者超链接. 这个缺点促使近来发展了许多 LDA 的变形, 利用网数据的结构来调整主题分布, 从而改善主题建模的精度和稳定性.

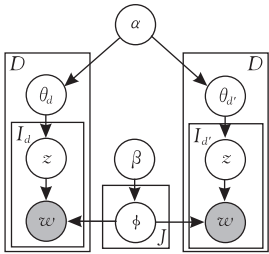


图 5 LDA 的图模型表示

3.2 动态主题模型

时间戳 (time stamp) 是文本网数据的一个重要特征, 可以揭示主题随时间演化的过程^[14,16,27-28]. 目前大多数时间序列文本数据的主题建模使用基于马尔可夫假设的两大类方法: 隐马尔可夫模型 (Hidden Markov Models, HMM) 和线性动态系统 (Linear Dynamic Systems, LDS)^[29], 分别针对离散和连续

的主题状态空间. 图 6 显示了 3 个典型的动态主题模型. 如图 6(a) 所示, dDTM^[14] 使用动态线性高斯模型来描述主题分布参数 ϕ_t 随时间的变化以及超参数 α_t 随时间的变化. 但是, dDTM 假设所有的主题在所有的时间片段都存在, 并没有考虑主题的产生和消亡过程. 此外, 类似于 HMM 的假设, dDTM 仅仅考虑 ϕ_t 在离散的状态空间变化. 如图 6(b) 所示, cDTM^[12] 进一步将 dDTM 的主题分布参数 ϕ_t 随时

间的演化扩展到连续的状态空间. 基于类似于 LDS 的假设, cDTM 利用布朗运动时间参数 s_t 来描述主题分布参数 ϕ_t 随时间的变化. 相比之下, 如图 6(c) 所示, TOT^[25] 并没有采用马尔可夫假设来描述主题在时间上的依赖关系, 而是采用一个基于主题标签 z 的连续 Beta 分布 ψ 来生成时间戳 t . 这样, 具有相同时间戳 t 的文本有很大的可能性具有相同的主题标签 z .

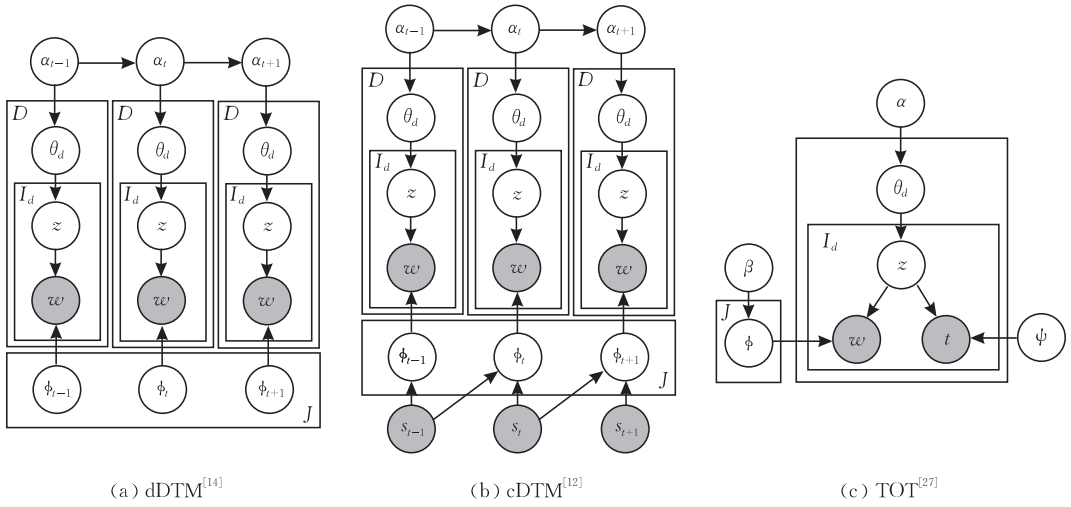


图 6 3 个动态主题模型范例

dLDA^[26] 与之前动态主题模型不同, 它假设主题分布参数 ϕ_t 不随时间改变, 但是文本的主题混合参数 θ_d 随时间改变. OLDA^[30] 自动适应文本流数据, 根据文本流数据在线更新参数, 从而学习并捕捉随时间变化的主题, 实现主题的实时跟踪和检测. 王金龙等人^[31] 采用两个独立的步骤, 首先利用 LDA 抽取主题, 然后分析主题随时间变化的强度来揭示主题的演化. 张小平等人^[32] 提出一种动态更新主题模型的超参数设置方法, 在中医临床诊疗数据的实验中, 获得领域专家解释性较好的分析结果. 有别于以上所有生成模型, DPM^[16] 利用概率判别模型 (discriminative model) 对在线主题进行检测, 着重最大化主题标签在给定文本下的条件概率而不是主题标签和文本的联合概率. 但是以上动态主题模型均未利用文本网数据丰富的结构特征, 例如引用、超链接和共标签等外部链接等. 因此, ITM^[3] 利用文本的相互引用信息来调节动态主题模型的参数, 从而检测出较为均衡的主题演化过程. 同时, 作者之间的合作关系也可能影响主题分布的演化, 并被模拟成为一个马尔可夫过程^[33], 来表示主题之间的跳转关系. 监督式主题模型 STM^[5] 可以被认为是一种特殊

的动态主题模型. 它利用文本的外部响应, 比如用户对文本的评价来调节主题分布. 如果这种外部响应是“时间戳”, STM 则鼓励具有相同时间戳的文本聚在一个主题内. 在这种情况下, STM 和如图 6(c) 所示的 TOT^[27] 模型十分相似.

3.3 二阶和高阶关系主题模型

静态复杂文本网数据的研究主要集中在如何刻画网结构信息对主题形成的影响. 其中二阶结构信息被广泛用来调节主题分布参数, 主要的建模方法大多可以归纳为无向图模型, 即马尔可夫随机场 (Markov Random Fields, MRF). 图 7 显示了 3 个典型的二阶关系主题模型. 如图 7(a) 所示, Link LDA^[34] 使用基于当前文本的主题混合比例 θ_d 生成一个主题标签 z , 然后用一个基于该主题标签 z 在所有其它文本的多项分布 ψ 来共同生成被引用的文本 d'' . 如果两个文本 $\{d, d'\}$ 共同引用同一个文本 $d''=d'''$, 它们就有较高的可能性分享类似的主题标签 z , 从而影响到主题分布 ϕ . 但是通过这种方式, Link LDA 仅仅描述了两个文本共同引用同一个文本的关系, 并未直接描述文本和文本的引用链接, 从而使得引用链接无法直接对主题的形成进行有效的

调节. 尽管 Link LDA 有简化的单个文本的图模型表示, 图 7(a)更加清晰地显示了两个相互关联文本的图模型表示. 如图 7(b)所示, PLDA^[23] 直接用基于主题标签 z 的伯努利分布 Ω 生成文本对文本的引用链接变量 y , 该变量是个二值变量, $y=1$ 表示观测到有链接, $y=0$ 表示没有观测到链接. 但是,

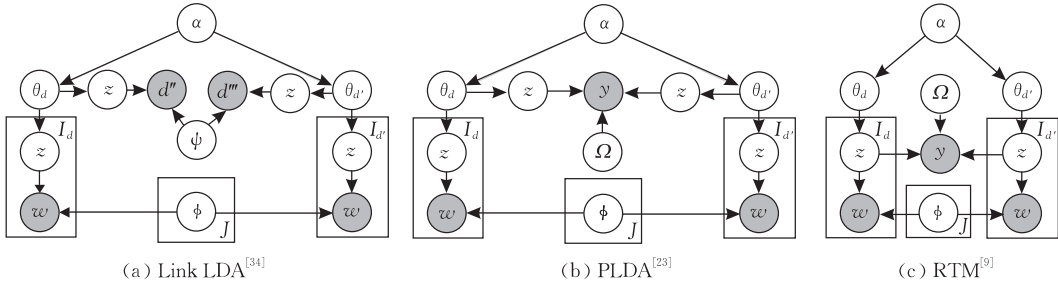


图 7 3 个二阶关系主题模型范例

LTHM^[20]假定文本部分的主题变量产生链接, 因此 LTHM 利用链接调节主题分布的作用介于 PLDA 和 RTM 之间. CIM^[25] 允许引用文本的主题依赖于自身主题混合参数和被引用文本的主题混合参数, 从而描述了引用和被引用文本之间的依赖关系. NetPLSA^[35] 利用文本网络的结构调控主题分布, 并鼓励减小主题分布之间的欧氏距离来调整主题分布. CTM^[4] 并未使用 Dirichlet 分布, 而是使用 logistic 正态分布来生成文本特定的主题比例, 从而达到调控主题之间关系的效果. 比 CTM 更进一步, PAM^[36] 使用有向无环图来表示主题之间的任意关系, 特别是稀疏的主题相互关联. BigramLDA^[37] 结合 bigram 语言模型和主题模型从而可以直接对单词之间的二阶关系而非主题之间的二阶关系进行建模, 部分地解除了传统的 unigram 主题建模的局限性. 但是, 以上所有二阶主题模型并未考虑到网数据中的高阶结构, 例如三阶或者四阶的关系, 导致常常遗漏特定的主题结构. 为了解决高阶关系, MTM^[13] 使用高斯马尔可夫随机场来描述多个会议文本集之间的关系, 这种高阶关系(在实验中使用了 6 个会议即六阶关系)使得发现的主题不仅捕捉每一个会议特定的内容, 而且还捕捉了多个会议之间的交叉主题. 陈江峰等人^[38] 利用高阶谱核测量 Web 服务文档的相似度, 实现 Web 服务结构化信息的发现.

如图 2 所示, 共标签的文本网数据存在大量的高阶结构, 例如一个文本有 3 个不同的标签, 那么这 3 个标签之间的关系(高阶关系)和每两个标签之间的关系(二阶关系)有很大区别. 因此, 以上所有针对

PLDA 仅仅随机地采样文本的其中一个主题标签 z 来产生链接 $y=1$, 从而限制链接调控文本主题分布 ϕ 的能力. 为了克服这个缺点, 如图 7(c)所示, RTM^[9] 使用文本全部的主题变量 $\tilde{z}$ 并且利用广义线性模型 Ω 来生成链接 $y=1$, 从而最大效能地利用链接对主题分布 ϕ 形成的影响.

二阶结构的主题模型无法精确描述细微的高阶结构. 图 8 显示了两个针对标签文本网数据的主题模型. 如图 8(a)所示, ATM^[39] 用一个文本特定的均匀分布 u_d 来产生一个标签 t , 然后用一个标签特定的主题比例 θ_t 来为每一个单词 w 生成一个主题标签 z . 所有和标签 t 联系的文本都共用同一个主题比例 θ_t , 因此间接地定义了多个文本之间的高阶关系. 如图 8(b)所示, L-LDA^[40] 是一类监督式主题模型. 它将隐藏的主题变量 z 表示成为观测到的标签 t , 这样每一个文本可以表示成为混合标签比例 θ_d . 在 L-LDA 中, 每一个标签 t 都对应着一个主题分布 ϕ , 来生成观测到的单词 w . 显然所有文本都通过标签 t 特定的主题分布 ϕ 来交换信息, 这跟图 5 所示的 LDA 模型相似. 然而, ATM 和 L-LDA 都没有充分考虑由多个标签引起的文本间的高阶关系, 因此不能充分刻画图 2 显示的普遍情况. 通常, 一个文本的多个作者会形成高阶的社交网络, 因此 CNT^[41] 使用高阶 MRF 刻画三阶和四阶的作者合作关系, 从而发现作者之间的共同研究兴趣, 并且发现作者之间潜在的社区结构.

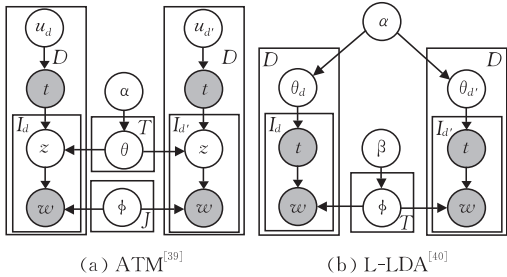


图 8 两个标签主题模型范例

3.4 多通路主题模型

多通路是复杂文本网数据的另一个主要特点. 不同类型的链接通常连接文本节点, 例如相互引用和共同作者两类链接. 如何利用不同通路的信息来调整高精度的主题分布目前获得广泛的关注. 本质上, 上述所有二阶或者高阶主题模型经过变形都适用于多通路文本网数据. 图 9 显示了 3 个典型的多通路主题模型. MRTM^[22] 用平行的 MRF 或者级联的 MRF 对相互引用和共同作者两种链接来联合建模, 通过平衡两种链接的关系来寻找更加精确的主题. 如图 9(a) 所示的平行结构 MRTM, 它本质上联合了之前讨论的两个简单的主题模型 ATM 和 LDA. MRTM 首先从一个文本特定的均匀分布 u_d 生成一个标签 t , 然后由标签特定的多项分布 ψ 生成一个主题变量 z . 同时, 该主题变量 z 还是由文本特

定的主题比例 θ_d 决定. 我们注意到两个文本之间的引用关系可以用 $\{\theta_d, \theta_{d'}\}$ 之间的链接来表示. 这样, MRTM 可以同时表示两类链接类型, 即文本之间的相互引用和共标签链接. 如图 9(b) 所示, MRTF^[21] 使用多个并列 MRF 分别对共同作者、共同杂志、文本引用、共同年份以及超级链接等多种链接进行建模. 它将所有的链接都表示成为文本特定主题比例 $\{\theta_d, \theta_{d'}\}$ 之间的关系. 大量实验表明使用各种不同的链接通常可以调整出较好的主题分布. 如图 9(c) 所示, TLLDA^[24] 利用图模型对博客共同作者的社交网络信息和博客之间的超链接信息来帮助搜索主题, 同时还能自动对博客作者的兴趣进行聚类. 它利用文本特定的主题比例 θ_d 来生成社交网结构 G . 这类类似于之前的许多二阶主题模型, 使用隐藏变量来生成网链接观测.

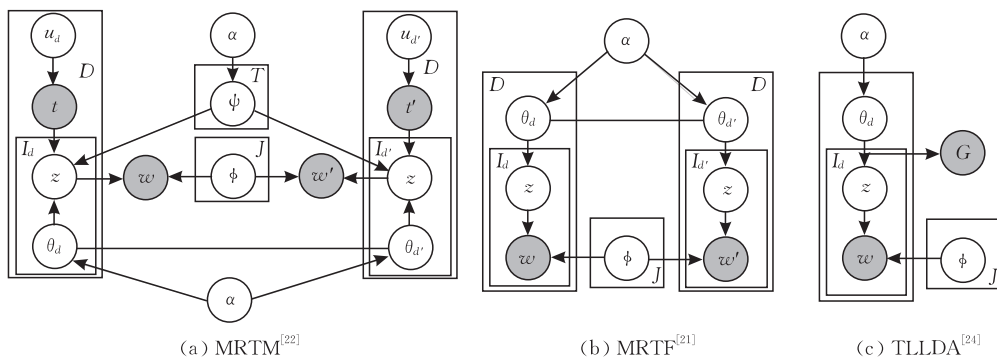


图 9 3 个多通路主题模型范例

事实上, 目前多约束的聚类算法大多可应用到多通路网数据信息. 例如, FICM^[42-43] 利用了文本之间多种信息通路, 比如标题、主题词以及语义相似度对文本聚类. MGC^[44] 使用广义凸集投影算法, 利用两类不同的约束 (基因表达序列和基因本体的语义相似度) 将基因聚类. 因此, 多通路的主题建模可以借鉴当前的多约束的聚类算法. 类似之前的监督主题模型 STM^[5], CTRF^[30] 使用条件随机场 (Conditional Random Fields, CRF) 使得主题的生成过程依赖于各种输入的特征. 如果把链接表示成为输入特征, 就可以用不同链接类型的输入特征来调节主题的生成. 例如, 张小平等^[41] 用文档中的词符合幂律分布的特征来调控主题分布.

3.5 分布式主题模型

分布式学习策略是机器学习和认知科学的一个重要基本问题. 首先, 随着多处理器和云计算的发展, 我们有必要发展分布式主题模型的学习算法, 来利用日趋强大的计算资源. 其次, 很多大规模网数据

都分布存储在不同的数据中心上. 每一个数据服务器只有局部的网数据信息, 全局的网数据信息通常无法获得. 因此, 分布式主题模型主要将计算任务和数据分配到多个局部的计算单元, 再利用信息融合的技术得到全局的主题模型参数. 然而, 将常用的主题模型 LDA 的学习算法 (例如 Gibbs sampling^[10] 或者 variational Bayes^[7]) 简单地并行化并不容易, 需要很多信息融合的技术. 为了对大规模独立文本集进行主题提取, Parallel-LDA^[45] 采用全局同步的思路, 首先基于全局初始化的主题模型参数对每一个分布的主题模型进行优化, 然后同步融合得到全局的主题模型参数. 这种方法要求每个分布的计算单元同步完成优化, 并不适用很多实际的情况. 因此, Async-LDA^[46] 采用异步通信的方式, 不需要等待所有局部的主题模型优化完毕, 只需要随机调整任意两个优化循环完毕的局部主题模型的参数. Async-LDA 等到所有局部主题算法收敛之后, 再做信息融合得到全局的主题模型参数.

3.6 计算复杂性

LDA 常用的两大类近似推理和参数估计方法包括 Gibbs 抽样 (Gibbs sampling) 和变分方法 (variational methods)^[7,10]. 这两类非监督学习算法的计算复杂度是 $O(JDI_d)$, 其中 J 是主题的数目, D 是文本集中文本的数目, I_d 是每个文本平均的单词数目. 动态主题模型的复杂度跟时间戳的总数 T 以及对相邻时间片段的主题依赖关系有关. 如果不考虑主题相互之间的时间依赖, 计算复杂度是 $O(JDI_dT)$. 如果考虑主题之间二阶时间相关性, 通常计算复杂度是 $O(J^2DI_dT)$. 事实上, 大多数二阶关系的主题模型的计算复杂度都是介于 $O(J^2DI_d)$ 和 $O(JDI_d)$ 之间, 取决于是否对任意二阶主题关系进行建模. 高阶主题模型的复杂度跟阶数相关, 例如 M 阶关系通常复杂度有 $O(J^MDI_d)$. 然而, 可以根据实际情况降低如此高的复杂度, 例如使用主题平滑 (smoothness) 的先验知识, 复杂度将降到只有 $O(JDI_d)$, 和传统 LDA 相当. 多通路的主题模型的计算复杂度跟通路的总数相关. 例如, 如果存在 M 种通路的一阶关系, 复杂度将是 $O(MJDI_d)$. 最后, 分布式主题模型的复杂度跟计算单元和单个主题模型复杂度相关. 例如, 如果有 M 个计算单元, 复杂度将是 $O(\frac{JDI_d}{M})$ 加上必要的 M 个计算单元之间通信的复杂度. 考虑到计算单元之间通信的计算复杂度通常远远小于单个主题模型的复杂度, 分布式主题模型近似是单个主题模型的复杂度的 $1/M$.

4 网数据主题建模的几个研究方向

尽管对独立文本的主题学习算法的研究在过去的十年有了长足的发展, 但是对文本网数据的主题提取精度和稳定度仍然不理想, 离大规模文本分析中的各项实际应用还有相当大的距离. 其中一个主要原因是尚未有完备的主题建模框架可以同时考虑复杂文本网数据的动态、高阶、多通路和分布式等 4 个主要特点. 因此, 可以采用一个基于三维马尔可夫模型和二型模糊系统的框架^[22,41,47-50], 系统地刻画大规模文本网数据的 4 个特点, 期望改善主题模型的精度和稳定度.

4.1 切片式的三维马尔可夫模型对动态主题的建模

在实际应用中, 研究主题网关系的演化通常比研究单一独立主题的动态过程更加有意义. 从网结

构的角度, 可以描述主题簇 (topic clusters) 之间的转化、产生或者消亡的过程, 并且通过文本网数据的外在联系进行主题分布调控, 从而分析并检测出稳定的主题簇的动态演化过程. 目前, 时变 MRF^[3] 可以用来恢复残缺动态网数据的结构, 对研究动态文本网数据的主题建模有很重要的启发. 此外, 文本网数据外部的响应, 比如对每一篇文章的评价分数或者每一个网页被点击的次数, 可以用来开发监督式的主学习算法^[5], 对网数据的动态主题分布进行调节.

切片式的三维马尔可夫模型 (Three-Dimensional Markov Models, 3DMM) 可以描述动态文本网数据. 如图 10 所示, 我们主要用文本网数据的两种主要链接类型为例 (即相互引用和共同作者) 来说明如何用切片式的 3DMM 对动态文本网数据进行主题建模. 该模型具有坚实的理论基础^[29], 即马尔可夫随机场 MRF 和隐马尔可夫模型 HMM. 类似的 3DMM 框架曾被应用到人体运动追踪中^[51]. 如图 10 所示, 不失一般性, 用 3 个 MRFs 来表示 3 个时间切片 $t = \{1, 2, 3\}$, 对片内的静态文本网数据的链接进行主题建模. 如图所示, 每一片静态网数据有两种链接类型: 相互引用 (细线箭头链接) 和共同作者 (黑色人形图标和无向链接). 类似的, 切片之间也存在相互引用和共同作者这两种链接类型. 此外, 深黑色粗箭头表示切片之间的时间依存关系, 采用基于 HMM 的方法来描述主题的动态演化. 这样, 切片式的 3DMM 可以刻画复杂文本网数据的主题演化.

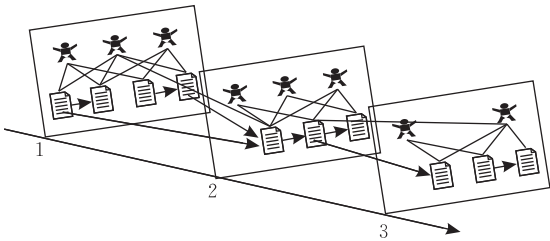


图 10 切片式 3DMM 对复杂文本网数据主题建模示意图

一方面, 用 MRF (或者 2DMM)^[49] 描述每一个切面的链接, 假设拥有链接的文本的主题之间具有很强的语义相干性, 这样文本主题之间会通过不同的链接类型相互影响. 这个 MRF “平滑” (smoothness) 先验知识被广泛使用在图像分割和机器视觉上^[29]. 具体而言, 如果两个文本之间存在链接, 它们就有很高的可能性分享类似的主题. 另一方面, 用 HMM (或者 1DMM) 刻画主题在切片之间的时间依赖关系. 同

时,主题还会通过切片之间的链接,例如相互引用或者共同作者,来相互影响.研究的目标是最小化以下的能量函数来搜索最优的主题标签配置:

$$E_{3\text{DMM}} = E_{2\text{DMM}} + E_{1\text{DMM}} \quad (1)$$

其中 $E_{3\text{DMM}}$ 是 3DMM 的能量,它由 2DMM 和 1DMM 的能量 $E_{2\text{DMM}}$ 和 $E_{1\text{DMM}}$ 共同构成.通过最小化 $E_{3\text{DMM}}$,不但鼓励了切片内的链接文本的主题相干性,并且鼓励了切片间文本的主题相干性.随着网数据在时间轴上的变化,主题在链接的关联下相互影响不断演化.利用传统的 LDA^[7] 来对文本内容进行主题聚类,然后同时利用链接和时间的依赖关系共同调节主题聚类结果.在 3DMM 的框架下,需要解决两个关键问题.首先,需要探索文本之间的链接如何影响主题随时间的演化.其次,需要研究主题的演化对未来的链接分布造成的影响.第一个是关于 3DMM 的推理和学习问题,即如何从动态网数据中推理和学习动态的主题分布.第二个是如何利用 3DMM 来预测链接的问题,即如何利用学习好的主题分布参数来预测未来网数据的链接分布.

为了开发 3DMM 的学习和预测算法,需要深入研究 3 种常用的图模型的推理和参数估计算法: Belief Propagation (BP)、Markov Chain Monte Carlo (MCMC) 和 Variational Methods (VM)^[29]. 给定动态网数据, BP 对每个主题变量求边际概率,然后利用主题变量的边际概率来估计模型参数. BP 算法的缺点是在有环图模型中无法保证收敛. MCMC 通过随机采样主题变量来更新模型参数,缺点是计算量庞大,并且不知道何时收敛. VM 通过转化原目标函数到一个近似目标来降低主题变量后验概率推理的难度,然后估计近似的参数,缺点是近似目标总是和原始目标有差距,导致学习的精度不高. 可以结合三种算法的特点,对基于 3DMM 的主题模型进行参数估计,开发出快速高效的学习算法. 在 3DMM 学习完毕后,需要进一步研究基于 3DMM 的预测算法,用来预测未知网数据的主题和链接分布.

通过外部响应来调整主题分布主要通过引入广义线性模型 (generalized linear model)^[5] 或者支持向量机 (support vector machines)^[52] 对文本数据的外部响应进行回归或者分类,进而调整主题分布以便能够更好地预测外部响应. 据此,可以采用一种不同的方案描述网数据的结构. 首先将网数据的链接作为一种外部响应特征,然后使用分类模型来更好

地预测链接,进而通过分类模型来影响网数据主题分布随时间的演化. 基于上述方案,可以类似地开发出对应的 3DMM 的监督式学习算法,不但可以学习主题分布,而且同时可以预测文本网数据的结构.

4.2 高阶 MRF 对高阶标签网数据建模

如图 2 所示,标签文本网数据是广泛存在的一类具有高阶结构的网数据. 研究的目标是利用高阶 MRF 对高阶标签网数据建模,从而推理和学习出更加稳定且具有语义关系的主题. 因为高阶关系是标签文本网数据的重要特征,因此可以采用高阶 MRF 来系统地表示标签文本网数据的高阶关系. 这里主要讨论如何构建快速有效的学习算法来降低优化 MRF 的复杂度. 在 3DMM 的框架下,进一步将 2DMM 的能量函数表示成为

$$E_{2\text{DMM}} = E_{2\text{-order}} + E_{\text{higher-order}} \quad (2)$$

其中, $E_{2\text{-order}}$ 表示传统的二阶关系产生的能量,而 $E_{\text{higher-order}}$ 表示高阶结构产生的能量. 可以采用两类有效策略来降低高阶 MRF 学习的复杂度. 首先,仅针对几类重要并且具有代表意义的高阶关系进行建模. 将所有高阶结构进行聚类,然后只鼓励几种典型的高阶结构. 这样,可以避免对所有的高阶结构进行学习,从而大大降低学习算法的复杂度. 其次,采用 BP 算法来学习高阶 MRF. 在信息传递的过程中,仅仅传递特定的主题组合信息,从而避免对所有的主題组合信息进行传递. 特别指出,这里在假设链接文本的主题间具有平滑特性 (smoothness) 的情况下,即链接文本具有很高的可能性共享相同主题. 这样,主题组合空间将大大减小,从而使得学习算法具有很好的扩展性. 通常,仅考虑三阶或者四阶的高阶关系就可以满足实际标签网数据的建模,因为绝大多数的实际文本关联的标签都少于 5 个. 尽管高阶结构对于抽取更加精确的主题有很大帮助,但是优化高阶 MRF 的参数通常需要庞大的计算成本^[53]. 如何降低优化或者学习高阶 MRF 的计算复杂度是未来的一个挑战^[54]. 当前,高阶 MRF 在机器视觉领域中的高层次图像先验建模方面有很多重要应用,例如图像恢复、视差估计和目标分割^[53-57]. 因此,可以对高阶的主题结构建立合理的能量函数,借鉴以上高阶 MRF 能量函数优化的方法,有效地搜索特定的高阶主题结构.

4.3 混合 MRF 的方法对多通路网数据的建模

现有的几种多通路主题模型并不能自动学习每

种通路对主题形成的影响,需要用户根据经验手动来调整每一种通路的权重. 因此,自动学习不同通路的权重是多通路主题模型迫切需要解决的一个重要问题. 可以借鉴基于 L_1 范数调整的 MRF 结构学习^[58]来自动地学习最优的权重参数,或者结合特征选择(feature selection)的方法进行最优通路的选择^[59]. 此外,如何设计有效的基于链接的输入特征并且学习输入特征的权重是一个有待探索的研究方向. 最近,多种链接特征被用于文本网数据聚类,可以用来平衡聚类重叠和层次之间的冲突^[2].

多通路的文本网数据的主题建模的关键是学习不同通路对主题分布影响的权重. 可以采用 MRF 方法结合基于广义投影的 Expectation-Maximization (EM)算法^[29]来学习多通路对主题形成的影响. 每一种链接对文本主题的形成影响不同. 尽管在时间轴上,这两种链接都同时存在,为了简便起见,这里仅考虑静态切片内部的多通路网数据. 在 3DMM 的框架下,可以进一步将 2DMM 的能量函数表示成为混合模型

$$E_{2DMM} = \sum_{n=1}^N \omega_n E_n \tag{3}$$

其中, ω_n 是第 n 种通路的混合权重,且 E_n 是每一种通路所产生的能量. 通过最小化混合能量函数(3),可以估计出每一种通路的混合权重. 在 3DMM 的框架下,将式(2)和(3)进一步合并成

$$E_{2DMM} = \sum_{n=1}^N \omega_n (E_{2\text{-order},n} + E_{\text{higher-order},n}) \tag{4}$$

因此,混合高阶 MRF(4)可以同时对外阶结构和多通路结构进行建模,并且学习稳定的主题分布. 类似的,也可以把时间轴上的高阶和多通路关系用以下目标函数表示

$$E_{1DMM} = \sum_{n=1}^N \omega_n (E_{2\text{-order},n} + E_{\text{higher-order},n}) \tag{5}$$

其中, E_{1DMM} 是 1DMM 的能量函数,可以写成混合和高阶 1DMM 的形式.

可以发展基于广义投影的 EM 算法^[44]对混合 MRF 进行优化. 在 E-step 中,可以依次估计出每一个通路 E_n 产生链接和节点的后验概率,称之为对 E_n 的广义投影. 然后,利用后验概率在 M-step 中最大化混合高阶 MRF 生成网数据的期望,同时估计出每一个通路的能量函数的参数. 其中,权重 ω_n 也可以用每一种通路的后验概率估计出来,并且受到

L_1 范数调整,即所有参数的 L_1 范数的和最小化. 通过多次广义投影之后,可以搜索出近似最优的主题文本来满足网数据的各个通路的约束. 通常,在推理后验概率中会遇到隐藏变量耦合的情况,所以很难得到精确的推理结果. 因此,可以借鉴近似推理的方法比如 variational Bayes^[7]来实现这一目标.

4.4 深浅学习对分布式网数据的主题建模

对于分布式的文本网数据,无论采用同步还是异步的分布计算技术,都需要在局部优化的过程中对全局的主题模型参数做出估计. 这对于大规模网数据的主题挖掘并不适用,因为每一个计算单元只能分配到局部的网数据的结构信息,如何在分布计算中融合全局网结构信息来调节动态、高阶和多通路主题模型参数仍然存在挑战. 二型模糊集合 (Type-2 Fuzzy Sets, T2FS)^[60] 近来被证明是一个能够处理单词计算 (Computing With Words, CWW)^[61] 中的不确定性的有效方法,因此可以作为分布式主题建模的信息融合的策略.

在 T2FS 框架下^[46-51,60-65] 利用深浅学习 (deep-shallow learning) 的方法能够较好的完成分布式主题模型的信息融合. 首先定义深度学习,即在每一个分布独立的计算单元使用 3DMM 学习动态高阶多通路的局部网数据的主题. 这个学习过程类似一个人在某个专业进行深入钻研,所以称之为深度学习. 在每个独立分布的计算单元学习完毕之后,对所有局部 3DMM 的参数进行融合,这个学习过程类似于一个人快速综合已有的知识,适应新的环境,因此称之为浅度学习. 可以采用 T2FS 来融合所有分布主题模型的参数,使用二级隶属函数来表示一级隶属函数的不确定性,实现深浅学习过程.

图 11 显示从分布式网数据提取主题分布的一个例子. 通常用一个多项分布来表示单词表的每个单词出现在一个主题的可能性,称之为主题分布. 从分布在 3 个计算单元上的网数据中学习出 3 个关于同一主题的分布如图 11(a) 所示. 因为每一个局部网数据的内容和结构不同,学习之后 3 个主题分布的参数有很大变化(图 11(b)). 这就要通过刻画主题分布参数的不确定性来融合 3 个主题分布的参数. 因此,可以使用 T2FS 的方法来刻画 3 个分布式主题的参数不确定性.

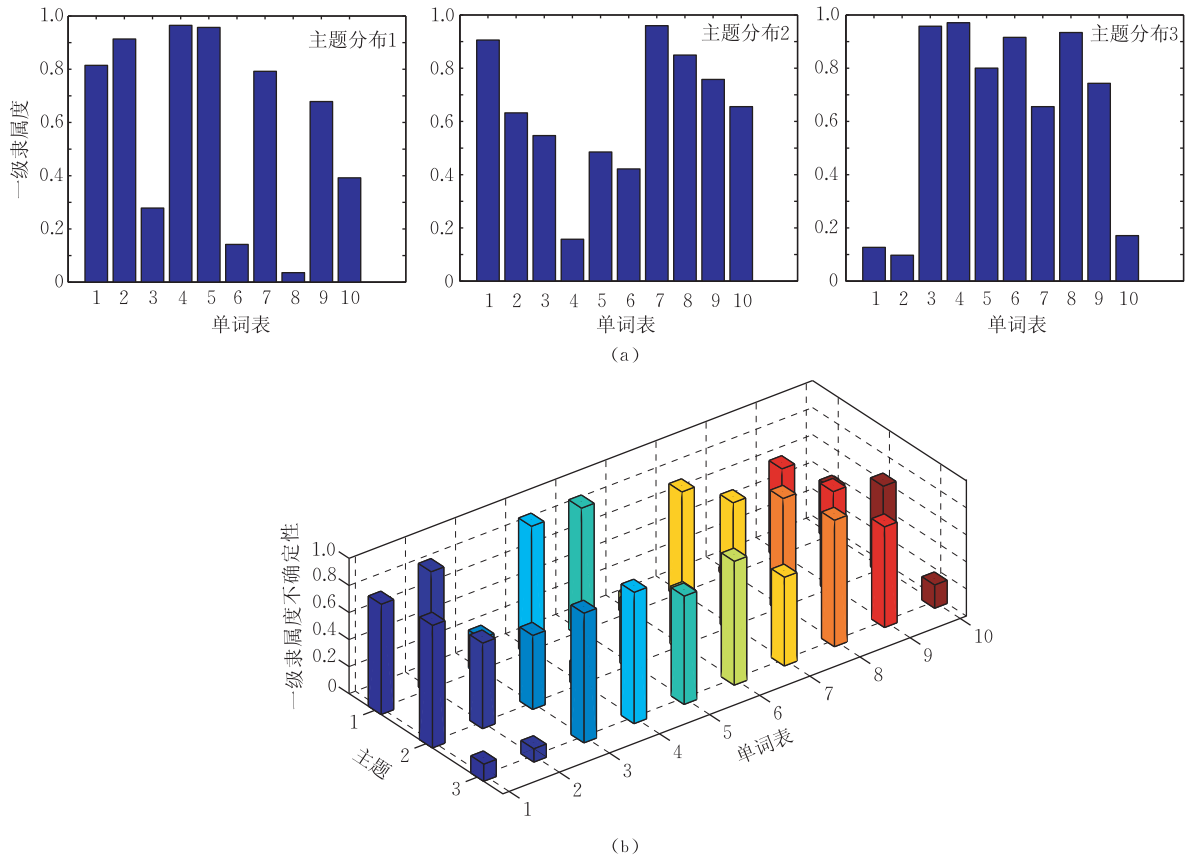


图 11 分布在 3 个计算单元上的网数据中学习出 3 个关于同一主题分布

使用 T2FS 可以刻画局部主题分布参数的不确定性. 如图 12(a)所示, 可以用一个 T2FS 来表示融合各个局部主题模型之后的主题模型. 因为局部网数据的内容和结构有很大变化, 融合之后的主题模型的参数有很大的不确定性. 在 T2FS 的框架下, 可以用一级隶属函数来表示单词表中的单词隶属于主题的可能性. 因为参数的不确定性, 一级隶属度本身可以用一个误差棒表示, 而在误差棒内的一级隶属度的不确定性可以进一步使用二级隶属函数来刻

画. 如图 12(b)所示, 假设在误差棒内的一级隶属度服从一个高斯形状的隶属函数, 意味着误差棒内的平均值对应的一级隶属度有很大的可能性来生成对应的主题单词. 如图 12(c)所示, 可以假设误差棒内的一级隶属度服从一个区间, 意味着误差棒内的一级隶属度具有等可能性来生成对应的主题单词. 这样, 就用二级隶属度函数(b)和(c)刻画了一级隶属度(主题分布参数)的不确定性.

从图 12 可以看出, 每一个基于 T2FS 的主题模型的输出都是一个模糊集合而不是一个精确的可能性. 这是由于主题模型参数不确定性造成的. 这样, 所有计算都可以采用二型模糊集合的运算法则^[64]. 深度学习的目标是保持各个局部主题分布的一致性. 提取分布在不同计算单元的网数据的主题过程中, 可以采用传统的异步通信来保持^[46]主题的一致性. 当深度学习完毕之后, 再采用 T2FS 来融合不同的局部主题模型的参数, 用二级隶属度表示参数的不确定性. 浅度学习的目标是估计二级隶属函数的参数, 反映每个局部主题模型的可能性. 通过浅度学习可以将局部的网信息用模糊集合来融合, 增强全局主题模型的精度和稳定性. 最后, 我们需要从基于 T2FS 的主题模型中推理出最好的主题分布来解释

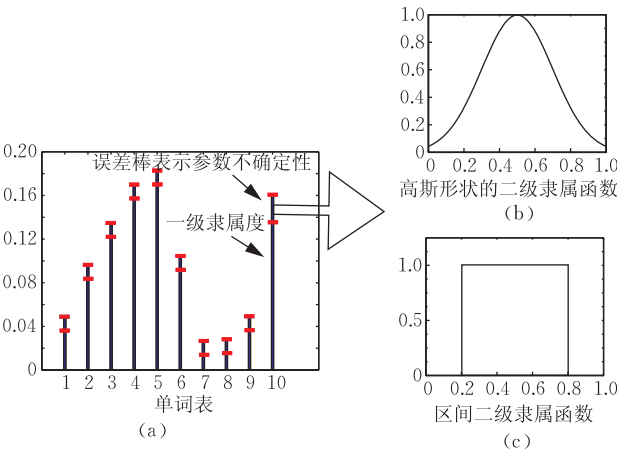


图 12 T2FS 对主题参数不确定性的刻画

未知的文本网数据. 通过该方法可以在二型模糊逻辑系统^[64]的推理框架下完成这一任务, 根据实际的网数据内容和结构, 推理得到较好的主题分布.

5 总结和展望

本文综述了近十年网数据主题建模研究的进展, 总结出了文本网数据的 4 个主要特点以及多个尚未解决的科学问题, 并讨论了几个可能在未来快速发展的研究方向. 由于主题建模在过去十年发展迅速, 本文没有详尽介绍大量的其它主题建模策略, 仅仅介绍了几个典型的主题模型范例. 未来网数据的主题建模将持续是机器学习和数据挖掘领域关注的热点, 我们寄希望于更多的研究者参与研究并解决本文提出的关键科学问题.

参 考 文 献

- [1] Mucha P J, Richardson T, Macon K et al. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 2010, 328(5980): 876-878
- [2] Ahn Y-Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*, 2010, 466(7307): 761-764
- [3] Xing E P, Ahmed A. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences of the United States of America*, 2009, 106(29): 11878-11883
- [4] Blei D M, Lafferty J. Correlated topic models. *Advances in Neural Information Processing Systems*, 2006, 18: 147-154
- [5] Blei D M, McAuliffe J D. Supervised topic models. *Advances in Neural Information Processing Systems*, 2008, 20: 121-128
- [6] Blei D M, Lafferty J. *Text Mining: Theory and Applications*, Chapter Topic Models. London: Chapman & Hall/CRC, 2009
- [7] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 2003, 3: 993-1022
- [8] Boyd-Graber J, Blei D M. Syntactic topic models. *Advances in Neural Information Processing Systems*, 2009, 21: 185-192
- [9] Chang J, Blei D M. Relational topic models for document networks//*Proceedings of the 12th International Conference on Artificial Intelligence and Statistics Clearwater Beach, Florida, USA*, 2009: 81-88
- [10] Griffiths T L, Steyvers M. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, 101: 5228-5235
- [11] Griffiths T L, Steyvers M, Blei D M et al. Integrating topics and syntax. *Advances in Neural Information Processing Systems*, 2005, 17: 537-544
- [12] Wang C, Blei D M, Heckerman D. Continuous time dynamic topic models//McAllester D A et al. *Uncertainty in Artificial Intelligence*, Helsinki, Finland 2008. Corvallis, Oregon: AUAI Press, 2008
- [13] Wang C, Thieson B, Meek C et al. Markov topic models//*Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics Clearwater Beach, Florida USA*, 2009: 583-590
- [14] Blei D M, Lafferty J D. Dynamic topic models//*Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, Pennsylvania, 2006: 113-120
- [15] Chang J, Boyd-Graber J, Blei D M. Connections between the lines: Augmenting social networks with text//*Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paris, France, 2009: 169-178
- [16] He Q, Chang K, Lim E P et al. Keep it simple with time: A reexamination of probabilistic topic detection models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(10): 1795-1808
- [17] Yang Xiao, Ma Jun, Yang Tong-Feng, Du Yan-Qi, Shao Hai-Min. Automatic multi-document summarization based on the Latent Dirichlet topic allocation model. *CAAI Transactions on Intelligent Systems*, 2010, 5(2): 169-176 (in Chinese)
(杨潇, 马军, 杨同峰, 杜言琦, 邵海敏. 主题模型 LDA 的多文档自动文摘. *智能系统学报*, 2010, 5(2): 169-176)
- [18] Wang Chao-Fei, Wang Kai. Application of topic model in digital library Web service. *Information Studies: Theory & Application*, 2010, 33(2): 118-123 (in Chinese)
(王朝飞, 王凯. 主题模型在数字图书馆 Web 服务中的应用. *情报理论与实践*, 2010, 33(2): 118-123)
- [19] Rzhetsky A, Seringhaus M, Gerstein M. Seeking a new biology through text mining. *Cell*, 2008, 134(1): 9-13
- [20] Gruber A, Rosen-Zvi M, Weiss Y. Latent topic models for hypertext//McAllester D A et al. *Uncertainty in Artificial Intelligence*, Helsinki, Finland 2008. AUAI Press, 2008: 230-239
- [21] Daum H. Markov random topic fields//*Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Suntec, Singapore, 2009: 293-296
- [22] Zeng J, Cheung W K, Li C-H et al. Multirelational topic models//*Proceedings of the Ninth IEEE International Conference on Data Mining*. Miami, Florida, USA, 2009: 1070-1075
- [23] Nallapati R M, Dittmore S, Lafferty J D et al. Multiscale topic tomography//*Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Jose, California, USA, 2007: 520-529

- [24] Liu Y, Niculescu-Mizil A, Gryc W. Topic-link LDA: Joint models of topic and author community//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada, 2009: 665-672
- [25] Dietz L, Bickel S, Scheffer T. Unsupervised prediction of citation influences//Proceedings of the 24th International Conference on Machine Learning. Corvallis, Oregon, 2007: 233-240
- [26] Kolchinsky A, Abi-Haidar A, Kaur J et al. Classification of protein-protein interaction full-text documents using text and citation network features. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010, 7(3): 400-411
- [27] Wang X, McCallum A. Topics over time: A non-Markov continuous-time model of topical trends//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. Philadelphia, PA, USA, 2006: 424-433
- [28] Pruteanu-Malinici I, Ren L, Paisley J et al. Hierarchical Bayesian modeling of topics in time-stamped documents. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(6): 996-1011
- [29] Bishop C M. Pattern Recognition and Machine Learning. New York: Springer, 2006
- [30] AlSumait L, Barbar D, Domeniconi C. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking//Proceedings of the Eighth IEEE International Conference on Data Mining. Pisa, Italy, 2008: 3-12
- [31] Wang Jin-Long, Xu Cong-Fu, Geng Xue-Yu. Study on research topic evolution based on probabilistic graphical models. Journal of the China Society for Scientific and Technical Information, 2009, 28(3): 347-355(in Chinese)
(王金龙, 徐从富, 耿雪玉. 基于概率图模型的科研文献主题演化研究. 情报学报, 2009, 28(3): 347-355)
- [32] Zhang Xiao-Ping, Zhou Xue-Zhong, Huang Hou-Kuan, Feng Qi, Chen Shi-Bo. A topic model based on CRP and word similarity. Pattern Recognition and Artificial Intelligence, 2010, 23(1): 72-76(in Chinese)
(张小平, 周雪忠, 黄厚宽, 冯奇, 陈世波. 基于词相似性与 CRP 的主题模型. 模式识别与人工智能, 2010, 23(1): 72-76)
- [33] Zhou D, Ji X, Zha H et al. Topic evolution and social interactions: How authors effect research//Proceedings of the 15th ACM International Conference on Information and Knowledge Management. Arlington, Virginia, USA, 2006: 248-257
- [34] Erosheva E, Fienberg S, Lafferty J. Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101: 5220-5227
- [35] Mei Q, Cai D, Zhang D et al. Topic modeling with network regularization//Proceeding of the 17th International Conference on World Wide Web. Beijing, China, 2008: 101-110
- [36] Li W, McCallum A. Pachinko allocation: DAG-structured mixture models of topic correlations//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, Pennsylvania, 2006: 577-584
- [37] Wallach H M. Topic modeling: Beyond bag-of-words//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, Pennsylvania, 2006: 977-984
- [38] Chen Jiang-Feng, Yu Jian-Jun. Topic model based structural Web services discovery. Journal of Beijing University of Aeronautics and Astronautics, 2008, 34(6): 734-738 (in Chinese)
(陈江锋, 于建军. 基于主题模型的结构化 Web 服务发现机制. 北京航空航天大学学报, 2008, 34(6): 734-738)
- [39] Rosen-Zvi M, Griffiths T, Steyvers M et al. The author-topic model for authors and documents//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Banff, Canada, 2004: 487-494
- [40] Ramage D, Hall D, Nallapati R et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, 2009: 248-256
- [41] Zeng J, Cheung W K, Li C-H et al. Coauthor network topic models with application to expert finding//Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010: 366-373
- [42] Zhu S, Zeng J, Mamitsuka H. Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. Bioinformatics, 2009, 25(15): 1944-1951
- [43] Zhu S, Takigawa I, Zeng J et al. Field independent probabilistic model for clustering multi-field documents. Information Processing and Management, 2009, 45(5): 555-570
- [44] Zeng J, Zhu S, Liew A et al. Multiconstrained gene clustering based on generalized projections. BMC Bioinformatics, 2010, 11(1): 164-177
- [45] Newman D, Asuncion A, Smyth P et al. Distributed inference for Latent Dirichlet allocation. Advances in Neural Information Processing Systems, 2007, 20: 17-24
- [46] Asuncion A, Smyth P, Welling M. Asynchronous distributed learning of topic models. Advances in Neural Information Processing Systems, 2008, 21: 81-88
- [47] Zeng J, Liu Z Q. Type-2 fuzzy hidden Markov models and their application to speech recognition. IEEE Transactions on Fuzzy Systems, 2006, 14(3): 454-467
- [48] Zeng J, Liu Z Q. Type-2 fuzzy sets for pattern recognition: The state-of-the-art. Journal of Uncertain Systems, 2007, 1(3): 163-177
- [49] Zeng J, Liu Z Q. Markov random field-based statistical character structure modeling for handwritten Chinese character recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 30(5): 767-780
- [50] Zeng J, Liu Z Q. Type-2 fuzzy Markov random fields and

- their application to handwritten Chinese character recognition. *IEEE Transactions on Fuzzy Systems*, 2008, 16(3): 747-760
- [51] Cao X, Zeng J, Liu Z. Human body parts tracking using sequential Markov random fields//*Proceedings of the 20th International Conference on Pattern Recognition*. Istanbul, Turkey, 2010; 1759-1762
- [52] Zhu J, Ahmed A, Xing E P. MedLDA: Maximum margin supervised topic models for regression and classification//*Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal, Quebec, Canada, 2009; 1257-1264
- [53] Kohli P, Kumar M P, Torr P H S. P3 & beyond: Move making algorithms for solving higher order functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 31(9): 1645-1656
- [54] Kohli P, Ladicky L U, Torr P H. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 2009, 82(3): 302-324
- [55] Lan X, Roth S, Huttenlocher D et al. Efficient belief propagation with learned higher-order Markov random fields//*Proceedings of the Vision-ECCV*. Graz, Austria, 2006; 269-282
- [56] Roth S, Black M J. Fields of experts: A framework for learning image priors//*Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Diego, CA, USA, 2005; 860-867
- [57] Agarwal S, Branson K, Belongie S. Higher order learning with graphs//*Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, Pennsylvania, 2006; 17-24
- [58] Lee S I, Lee H, Abbeel P et al. Efficient L1 Regularized Logistic Regression//*Proceedings of the 21st National Conference on Artificial Intelligence*. Boston, Massachusetts, USA, 2006; 401-409
- [59] Zhu J, Lao N, Xing E P. Grafting-light: Fast, incremental feature selection and structure learning of Markov random fields//*Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington, DC, USA, 2010; 303-312
- [60] Mendel J M, John R I B. Type-2 fuzzy sets made simple. *IEEE Transactions on Fuzzy Systems*, 2002, 10(2): 117-127
- [61] Wu D, Mendel J M. Aggregation using the linguistic weighted average and interval type-2 fuzzy sets. *IEEE Transactions on Fuzzy Systems*, 2007, 15(6): 1145-1161
- [62] Mendel J M. Advances in type-2 fuzzy sets and systems. *Information Sciences*, 2007, 177(1): 84-110
- [63] Zeng J, Xie L, Liu Z-Q. Type-2 fuzzy Gaussian mixture models. *Pattern Recognition*, 2008, 41(12): 3636-3643
- [64] Karnik N N, Mendel J M, Liang Q. Type-2 fuzzy logic systems. *IEEE Transactions on Fuzzy Systems*, 1999, 7(6): 643-658
- [65] Wu D, Mendel J M. Uncertainty measures for interval type-2 fuzzy sets. *Information Sciences*, 2007, 177(23): 5378-5393



ZENG Jia, born in 1980, Ph. D., professor, Ph.D. supervisor. His interests include pattern recognition, data mining, machine learning and computational biology.

YAN Jian-Feng, born in 1978, Ph. D., His research interests include pattern recognition and sensor networks.

GONG Sheng-Rong, born in 1966, Ph. D., professor. His research interests include pattern recognition, image processing.

Background

Topic models with high-precision not only summarize large-scale documents such as huge amount of scientific and technical literatures, government documents, community documents and news files, but also predict and track important and promising research topics from the scientific articles. Recently, variants of topic models have found many important applications in text mining, computer vision and computational biology, for example, topic detection and tracking of historical scientific articles with time stamps, unsupervised activity perception in crowded and complicated scenes, and large-scale microarray gene expression analysis and medicine

discovery.

This paper introduces topic modeling of large-scale document network data. The main focus of this review is on the four major characteristics of document network data and the current state-of-the-art topic models. Latent Dirichlet Allocation (LDA) is the classic topic model, based on which variants of improved topic models can be developed. However, LDA only uses content similarity to exchange topic information, but ignores other important information existing in external link structure of the document network data, such as citation, hyperlink and co-tagged relations. In this article,

we introduce the four major characteristics of document network data, such as dynamic, higher-order, multiplex and distributed properties, and discuss the state-of-the-art improvement of topic models for document network data. Furthermore, we propose the three-dimensional Markov models and type-2 fuzzy logic systems to handle these characteristics within a unified framework. We also discuss the inference and parameter estimation method to minimize the energy of the three-dimensional Markov models.

This project is supported by the National Nature Science Foudation of China (Nos. 61003154, 61003259, 61170124), Foundation of Shanghai Key Laboratory of Intelligent Information Processing Shanghai (No. IIPL-2010-009), and Major Program of National Nature Science Foudation of Jiangsu Province (No. 12KJA520004).

Our research group focuses on the development of pattern recognition and machine learning techniques, and some

key applications in computational biology and bioinformatics. We have accumulated several related work. For example, we propose variants of type-2 fuzzy graphical models published in Pattern Recognition and IEEE Transactions on Fuzzy Systems. We develop several variants of Markov models for handwritten Chinese character recognition published in IEEE Transactions on Pattern Analysis and Machine Intelligence and Information Sciences. We also develop graphical models for human promoter recognition and biomedical text mining published in Bioinformatics, Briefings in Bioinformatics, BMC Bioinformatics, Physical Biology, Physical Review E, Information Processing & Management, and IEEE Transactions on Computational Biology & Bioinformatics. This paper summarizes and extends our current research to hierarchical Bayesian models, which lays a solid foundation for the future applications in text/image mining and computational biology.