基于 LDA 的微博文本主题建模方法研究述评*

张培晶[1]　宋　蕾[2]

[1]中国人民公安大学办公室　北京 100038　[2]北京警察学院公安科技系　北京 102202

〔摘要〕在介绍概率主题模型发展过程以及概率主题模型的代表性模型 LDA 基本原理的基础上，分析 LDA 模型的特征及其用于微博类网络文本挖掘的优势；介绍和评述微博环境下现有的基于 LDA 模型的文本主题建模方法，并对其扩展方式和建模效果进行总结和比较；最后对微博文本主题建模的发展方向进行展望。

〔关键词〕LDA　概率主题模型　微博　主题建模

〔分类号〕G356

Overview on Topic Modeling Method of Microblogs Text Based on LDA

Zhang Peijing[1]　Song Lei[2]

[1]Office of Chinese People's Public Security University, Beijing 100038

[2]Department of Police Technology, Beijing Police College, Beijing 102202

〔Abstract〕Based on the development process of probability topic model and basic principle of its representing model – LDA, this paper analyzes the characteristics of LDA and advantage of microblogs network text mining. Then it introduces and comments the existing text topic modeling methods based on LDA in microblogs environment, and compares their expanded mode and modeling effect. Finally it prospects the development direction of microblogs text topic modeling.

〔Keywords〕Latent Dirichlet Allocation(LDA)　probability topic model　microblogs　topic modeling

1　引　言

以 LDA(latent dirichlet allocation)模型[1]为代表的主题模型是近年来文本挖掘领域的一个热门研究方向。主题模型具有优秀的降维能力、针对复杂系统的建模能力和良好的扩展性。利用主题建模挖掘出的主题可以帮助人们理解海量文本背后隐藏的语义，也可以作为其他文本分析方法的输入，完成文本分类、话题检测、文本自动摘要和关联判断等多方面的文本挖掘任务。LDA 主题模型已经在文本挖掘及其相关领域中得到了广泛应用，并且在以新闻类数据为主的传统网络文本挖掘方面获得了很大成功。

微博(microb log)是基于 Web2.0 实现的社会媒体(social media)的一种形式，近年来得到了爆发式的发展，已经成为人们表达思想、传播信息和交流学习的重要工具，被越来越多的用户和机构所关注。微博与传统网络文本有着显著不同，主要表现为：微博帖子的文本长度被限制在 140 个字符，文本多是只言片语，数据稀疏性问题严重；书写较为随意，使用语法不规范，网络用语、符号语言和新生词大量出现，数据噪声大；微博支持用户通过手机和网络等多种方式进行即时信息发布，具有更新速度快和文档数据规模庞大的特征。微博的上述特点给其主题建模带来了挑战。目前，有关主题模型在微博类社会媒体上的应用研究相对较少，尚处在探索阶段。但有关研究已经证明，不论是作为独一无二的使用特征，还是作为多重真实世界任务的补充特征，主题建模对于微博文本的挖掘都是非常有用的[2]。

2　LDA 模型概述

2.1　概率主题模型的提出

海量文本的复杂性导致了基于主题的分层次统计模型的研究，产生了以 LDA 为代表的概率主题模型。相对于可以观察到的文档和词，主题是一个抽象的概

* 本文系北京市社会科学规划项目“社区管理创新视角下北京地区虚拟社区综合治理机制研究”(项目编号：11SHC026)和国家社会科学基金项目“虚拟社区中的信息交流与导控机制”(项目编号：11CTQ026)研究成果之一。

收稿日期：2012－09－17　修回日期：2012－11－14　本文起止页码：120－126　本文责任编辑：杜杏叶

念，代表了一个潜在的语义主旨(subject) 。概率主题模型本质是通过对文本中词的分布规律的观察，实现对相似分布规律词集的聚类。主题相当于聚类中的簇，文档以不同的概率属于不同的主题。

概率主题模型的前身可以追溯到 LSA (latent semantic analysis) [3] 潜在语义分析。LSA 打破了人们以往基于“词典空间”进行文本表示的思维模式，创新地引入了语义维度，实现了文档在低维的隐含语义空间上的表示。但是，LSA 的方法论基础来源于线性代数，是基于奇异值分解 SVD(singular value decomposition) 的单词计数原始矩阵的近似分解。由 LSA 获取的概念层表示无法处理“一词多义”问题。同时，SVD 涉及到矩阵运算，计算复杂度较高，而且矩阵运算的结果在很多维度上为负数，没有对应的物理解释。霍夫曼针对 LSA 的缺陷提出了一个构建在可靠的概率统计学基础之上的新方法 PLSA(probabilistic latent semantic analysis) [4] 。此前，每个语义维度对应一个特征向量，而在概率模型中，每个语义维度对应一个词典上的概率分布。PLSA 可以明确地区分单词使用的不同意思和不同类型，解决了“一词多义”问题。但 PLSA 并没有在文档层提供概率模型，这导致了模型中待估参数的数量随着语料库的大小呈线性增长，容易出现过度拟合问题。2003 年，D. Blei 等人[1] 通过引入一个 Dirichlet 先验分布扩展了 PLSA 模型，克服了 PLSA 参数随着文档集增长而线性增长的不足，形成了当前被广泛应用的产生式概率主题模型 LDA。至此，主题模型被第一次正式提出。

2.2 LDA 模型的描述

LDA 模型是一个分层的贝叶斯模型，包含文档、主题和词三个层次。LDA 模型基本思想是每个文档都可以表示成若干潜在主题的混合分布，每个主题是词汇表中所有单词的概率分布。LDA 模型将主题混合权重 θ 视为 T 维参数的隐含随机变量，并对主题混合权重 θ 引进 Dirichlet 先验。Dirichlet 分布的公式为:

$$\mathrm{Dir}(\alpha_1 \cdots, \alpha_t) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} \theta_j^{\alpha_i - 1} \quad (1)$$

参数 $\alpha_1, \cdots, \alpha_T$ 均为多项式 $\theta = P(\theta_1, \cdots, \theta_T)$ 的超参数，每个超参数 α_j 可以被理解为从一个文本中抽样主题 j 的次数的一个先验。考虑不同的主题被利用的方式几乎没有变化，为方便计算，LDA 假定文本中的主题是可交换的，采用具有相同超参数 α 的对称 Dirichlet 分布，这里 $\alpha_1 = \alpha_2 = \cdots = \alpha_T = \alpha$。Dirichlet 分布和多项式分布是共轭分布。若 T 维随机向量 θ 服从 Drichlet 分布，则 θ 的 T 个分量 $\theta_1, \theta_2, \cdots, \theta_T$ 都是连续的非负值，且 $\sum_j \theta_j = 1$ [5] 。D. Blei 最早提出的原始 LDA 文献[1] 中只对文档 - 主题分布 θ 增加了 Dirichlet 先验，并没有对主题 - 词汇分布 φ 进行先验假设。为了充分利用共轭概率分布的特性，便于推理运算，Giffiths 等人在后续的研究中对 φ 加上了 Dirichlet 先验，并假定主题中的单词是可交换的，其超参数为 β。[6] 至此，LDA 中有了两组先验。一组是文档 - 主题的先验，来自一个对称的 Dirichlet(α) ; 另一组是主题 - 词汇的先验，来自一个对称的 Dirichlet(β) 。

2.3 LDA 模型的特征及其用于微博文本建模的优势

LDA 主题模型具有优秀的降维能力，可以将原来高维的单词空间降维到由一组主题构成的相对较小的主题空间上。降维技术不仅可以有效降低文本相似性计算的复杂度，还可以避免传统文本建模方法存在的数据稀疏性等问题。对于微博类的短文本而言，文本中单词非常有限，同一个词共现在两篇不同短文本中的概率较小，使用传统以词或短语为特征的向量表示方法，很难准确计算文本间的相似度。利用主题模型可以将每个短文本表示成低维的主题空间上的一个向量。低维的主题向量的空间则不会是稀疏的，有助于隐含的相关性被挖掘出来，从而减少数据稀疏性对度量文本之间相似度的影响，更充分地挖掘文本集合的内在信息。

LDA 主题模型具有扎实的概率理论基础。一方面，LDA 模型使用概率统计方法来分析文本，是建立在词袋(bag of words) 假设之上的。词袋假设认为一篇文章可以表示为一组单词的无序组合，而忽略了文章中的语法和单词的次序[5] 。与基于语法规则的语言学的文本分析方法相比，LDA 通过对文本中词汇共现次数的概率统计来挖掘潜在的语义结构，更适合用于存在着大量语法不规范问题的微博类网络文本的建模和分析; 另一方面，LDA 模型使用概率分布方法来表示文本，主题是单词表中词的随机分布概率，文本以不同的概率属于不同的主题，服从真实数据的概率分布。概率分布表达方法能够较好地表达和量化文本中存在的不确定性，减小噪声干扰。对于语言组织规范性差、新生词汇大量出现的微博类网络文本而言，以 LDA 为代表的概率主题模型更适合微博类网络文本挖掘过程中不确定性环境下相对准确的计算。

LDA 主题模型具有良好的扩展能力。LDA 模型属于贝叶斯网络模型，可以方便地把各种元数据、结构化信息甚至领域知识作为贝叶斯网络中的随机变量加

入到主题模型中;也可以将两个主题模型合并,以搭积木的方式形成一个新的主题模型[7]。微博具有多种非文本特征信息。例如,不同类型微博(按照不同的消息发布方式,微博可以分为广播类(broadcast)、对话类(conversion)和转发类(retweet)3种类型)的结构化特征,使用哈希标签(Hashtag)(微博文本中的Hashtag标签是一种用于简化搜索、索引和趋势发现的用户自定义标签,格式为"#话题名#"(中文微博)或"#话题名"(twitter)),具有用户特征和时间特征等。此外,微博也是一种社会关系网络,带有一些结构化的社会网络方面的信息。利用LDA模型良好的扩展性,可以将微博中的这些特征信息引入到LDA模型中,实现更准确的微博文本建模和主题挖掘。

综上所述,相对于其他文本建模方法,应用LDA主题模型进行微博类网络文本的建模和分析具有很大的优势。但是直接使用传统LDA模型对微博进行主题建模,一定程度上仍然受到篇幅过短、内容和格式散乱、数据噪声较大等方面的影响。目前,已经有一些基于LDA的改进方法用于微博环境下的主题建模。

3 微博文本主题建模方法介绍

3.1 直接使用LDA模型进行主题建模

3.1.1 直接使用LDA模型对微博帖子进行主题建模

将每个微博帖子对应为单个文档,可以直接使用LDA模型对其进行主题建模。由于LDA模型是无监督的,此方法实现过程最为简捷。但有关研究证明,LDA主题模型的效力很大程度上受文档长度的影响,一个短文本缺少足够的词出现次数,无法帮助判断是否这些词是相关的[8]。因此,这种方法的微博主题建模效果并不理想。

3.1.2 基于聚集的LDA主题建模方法

社会媒体中本身存在着一定的"聚集策略",按照这些策略可以将一系列存在某种关联关系的微博帖子聚集为一个长文档,然后使用传统的LDA模型在此聚集后的长文档上进行主题建模。常用的聚集策略是按照发布者进行聚集,即将同一微博用户的所有帖子聚集为一个文档。该方法得到的主题-单词分布φ和文本上的主题分布θ都是微博用户层面的,可以用于微博用户及其感兴趣主题的分析。Weng Jianshu等人使用这种方法有效实现了敏感主题有影响力微博博主的发现[9]。

3.1.3 基于训练的LDA分步建模方法

一些情况下,用户希望实现针对微博帖子的分析,例如进行帖子过滤、帖子分类等。如何在解决短文本信息不足的前提下,实现这一目标呢?Hong Liangjie等人提出了通过预先训练来分步完成的方法,本文将其总结为基于训练的LDA分步建模方法,其中具有代表性的实现模式有用户模式(user scheme)和术语模式(term scheme)[2]。基于训练的LDA分步建模方法的基本思想是先在聚集后的长文本数据集上进行LDA模型的训练,学习得到合理的主题,也就是挖掘出主题在词汇上的分布φ;然后再用训练好的模型去推导训练集上原始的短文本或等待测试的新文本的主题分布θ,在此推导过程中,并不再进行主题φ的重新训练,令$\varphi_{长文本,训练}=\varphi_{帖子,测试}=\varphi_{帖子,训练}=\varphi_{长文本,测试}=\varphi$。笔者认为,这种方法本质上是将LDA模型的$\varphi$和$\theta$的求解过程通过先期训练的方式进行人为的拆分,从而在应用中可以根据文本的实际情况,选择合理的解释性强的主题进行统一的文本主题分步推导,同时也简化了φ的求解。但这种方法也存在主题模型僵化的问题,当有新的文档来到时,若要更新模型仍需重新训练。另外,建模过程也需要大量的文本预处理工作和人工干预。

Hong Liangjie等人提到的用户模式是将训练数据集中相同用户的帖子聚集为一个长文本,然后进行LDA模型的训练;术语模式是将训练数据集中具有某个相同术语的帖子聚集为一个长文本,再进行LDA模型的训练。术语模式设计的出发点是考虑到微博用户经常使用自定义的Hashtag标签来标示主题或事件,构建术语聚集文本可以使我们直接获得与这些Hashtag标签相关的主题。

3.2 使用ATM模型进行主题建模

ATM(author-topic model)[10]是一种基于LDA的扩展模型,旨在通过引入文章的作者信息来指导LDA主题生成,最初用于对维基百科进行建模。实验表明,当测试文档中仅仅有小数量的单词被观察到时,该模型胜过LDA。不同于LDA的是,模型中"文本-主题"的分布被"作者-主题"分布所取代。由于每个微博帖子只有一名作者,微博上的ATM主题建模实际上是ATM模型的一个特例。ATM建模得到的是微博用户层面的主题混合分布,而不是微博帖子层面的主题混合。经过某些扩展或特殊的处理,扩展后的ATM也可以支持同时对微博帖子和用户进行主题建模[11]。

3.3 使用Twitter-LDA模型进行主题建模

ATM模型较好地解决了微博数据稀疏性问题,但同时也降低了针对单个帖子的文本分析能力。考虑到

一个较为普遍的观测规律: 一个单独的微博帖子通常只有一个单一主题,Zhao Wayne Xin 等人对 ATM 模型进行了扩展,提出了 Twitter-LDA 模型[12]。Twitter-LDA 模型与 ATM 模型的主要区别是、Twitter-LDA 不仅可以在微博用户层面进行主题建模,而且可以对单个的微博帖子进行主题建模。

Twitter-LDA 模型在 ATM 基础上有两处变化: 一是引入了背景模型 φ_B,可以有效地降低高频词汇的影响。类似"love"、"thanks"这些出现频率极高的词汇如果不被处理,会对文本挖掘造成严重干扰; 二是给每个微博帖子内部的所有单词赋予一个统一的主题,可以实现对用户和帖子两个层面的同时建模。此外,Twitter-LDA 还可以方便地开展帖子层面的数据统计。例如,可以计算语料集合中有多少帖子与某主题相关以及一对单词在某主题相关的帖子中共现的次数。

3.4 使用 Labeled LDA 模型进行主题建模

考虑微博作为一种网络文本,很多数据已经被读者贴上了标签(tags) ,利用这些已经存在的标签资源,有助于更好地进行主题挖掘,D. Ramage 等人提出了使用 Labeled LDA 对微博文本进行主题建模[13]。Labeled LDA 最早是 D. Ramage 等人[14]在 LDA 基础上提出的一个受监督的主题模型,可以描述一个标签化文档集的产生过程。与 LDA 不同的是,Labeled LDA 通过对主题模型的一个简单约束来引入监督,即仅仅使用那些与一个文档可以观察到标签集合相对应的主题。该模型学习得来的主题直接与每个标签对应关联。Labeled LDA 的优势在于通过提供一个标签集(Label Set) 到其学习过程中,使其学习得到的主题更具有解释性,也可以用于解决文本分类中的可信归属问题。

D. Ramage 等人将微博中的哈希标签对应为 Labeled LDA 中标签(Label) ,使用 Labeled LDA 模型在微博上进行主题建模,有效利用了微博上已有的用户自定义主旨信息,较为准确地找出了与每个哈希标签密切关联的词汇。另外,还可以使用主题模型对包含大量简易符号的微博文本进行建模,实现了符号标识的主题挖掘。例如,挖掘出符号标识 ":)" 与 thanks、thank、much、too、hi、following、love 等单词存在密切关联。还需指出,Labeled LDA 克服了 Supervised LDA 和 DiscLDA[15] 等早期引入监督的主题模型限制一个文档只可以与一个单一的标签关联的缺点,允许对多标签的语料集进行建模。因此,Labeled LDA 也支持在微博帖子的聚集文档之上的主题建模,聚集后的文档允许包含多个哈希标签。D. Quercia 等人[16]将基于微博帖子聚集文档上的 Labeled LDA 主题建模方法用于微博用户文档(Twitter profiles) 的分类,并通过实验证明,该方法能够在较少训练数据的情况下,实现较为准确的微博用户分类,其分类性能远远胜过支持向量机(support vector machines,SVM) 的线性分类方法。

3.5 使用 MB-LDA 模型进行主题建模

国内中文微博类似于国外 Twitter 的引入,但是二者除使用文字和语言结构的不同之外,在其他方面也存在差异: 一方面,中文微博在 Twitter 基础上进行了简单的扩展。例如,较早支持了对话评论,并且转发微博的同时也可以加入评论,而 Twitter 只能转发,不能同时加入评论; 另一方面,国内微博媒体的"转发"类微博所占比例相当大,频次要高出 Twitter 几个数量级[17]。考虑到中文微博的上述独有特征,结合微博文本内蕴含的社会网络方面的结构化信息,张晨逸等人在 LDA 基础上提出了专用于中文微博主题建模的 MB-LDA 模型[18]。

MB-LDA 在原有 LDA 的基础上作了两方面的扩展,分别将微博的文本关联关系和联系人关联关系引入到微博主题建模中。对于转发类微博(一般格式为"评论 RT@ 作者 原文") ,将微博的转发部分和原文部分关联起来进行主题建模; 对于对话类微博(以"@ 联系人"格式开头) ,将本次发布内容与联系人的微博内容关联起来进行主题建模。

张晨逸等人通过对衡量主题模型性能和推广性的 perplexity 指标的计算发现,MB-LDA 模型在中文微博上的主题挖掘功能明显优于传统的 LDA 模型。另外,通过 MB-LDA 模型不仅可以挖掘出每条微博可能所属的主题以及每个主题代表性的词,还可以推导出联系人的主题概率分布 θ_c,从而挖掘出每个联系人感兴趣的主题。

4 微博文本主题建模方法比较

上述微博文本主题建模方法主要表现为 LDA 模型两个方向的扩展: 一个方向是纵向的基于操作过程的扩展; 另一个方向是横向的基于模型的扩展。

4.1 基于操作过程的扩展

在基于操作过程的扩展方面,不论是基于聚集的 LDA 主题建模方法,还是基于训练的 LDA 分步建模方法,都需要预先进行短文本的聚集,与利用 ATM 模型在微博上进行主题建模的基本思想是类似的。但这几种主题建模方法产生的主题彼此间却存在很大的不

同。Hong Liangjie 等人对微博帖子上的 LDA 建模(文献中称为消息模式)、ATM 的扩展模型建模(同时获得用户和帖子两个层面)、基于训练的用户模式建模和术语模式建模共 4 种方式产生的主题进行了对比实验分析[2]。实验表明,在主题的一致性方面,不同建模方法学习得到的主题并不直接相一致。对于不同建模方法得到的相似主题,使用 JS(Jensen-Shannon divergence)算法计算相似主题构成单词的概率是各自不同的,并且这种差异性会随着主题数量的增加而略微增加;使用肯德尔(Kendall)等级相关系数去测量相似主题构成单词的排名也发现存在一定差别。通过上述两项指标的测量,可以推导出,通过不同模式或模型学习的主题,它们之间是存在明显区别的。相比较其他模式,消息模式和用户模式之间的分歧更加突出,而通过术语模式和 ATM 模型学习的主题更接近消息模式学习的主题。在建模性能方面,在聚集后的长文本上训练的模型能够产生更好的性能,其中基于训练的用户模式能够实现训练过程更快和训练质量更好。但当文本长度过于大时,主题模型反而变得低效。D. Quercia 等人在使用 Label-LDA 用于微博用户分类时也曾指出,中等活跃度用户的帖子聚集后形成的微博用户文档(Twitter Profile)长短适中,其分类性能和效果最理想。[16] 在主题数量方面,消息模式由于微博帖子数量巨大,与其他三种方式相比,会获得更大数量的主题。另外,通过 ATM 的扩展模型学习产生的针对帖子的主题,通常数量很少,并且比消息模式获得的结果还糟糕。

4.2 基于模型的扩展

在基于模型的扩展方面,ATM 和 Twitter-LDA 都是利用微博文本的用户特征来实现 LDA 模型的扩展。只是 Twitter-LDA 在 ATM 基础上作了进一步扩展,引入背景模型,并实现了用户层面和帖子层面的同时建模。本质上,MB-LDA 利用文本结构化信息也实现了微博短文本的扩展,只是这种扩展是依据微博的具体类型而区别对待的。Labeled-LDA 是通过引入微博中结构化的标签信息,将原先无监督学习的 LDA 模型变为监督型的模型,有效提高了主题建模效率和主题可解释性。在主题挖掘效果方面,从已有文献资料推断,相比直接在微博帖子之上的 LDA 主题建模,上述扩展后的主题模型更具优势;但多数扩展后的模型相互之间并无绝对的孰优孰劣,其建模效果与应用目标有着很大关系。例如,对于用户感兴趣主题的挖掘,ATM 模型较为简单实用;对于包含足够标签信息的微博主题挖掘和文本分类而言,Labeled-LDA 更有优势;对于转发频次很高的中文微博的主题建模,MB-LDA 适用性更强;而 Twitter-LDA 可以同时实现对用户和帖子两个层面的建模,并且 Zhao Xin 等人通过实验证明,与传统 LDA 模型和 ATM 模型相比较,该模型挖掘的某个主题的前几个单词具有更好的关联性和可解释性[19]。上述讨论的各种建模方法的比较情况,如表 1 所示:

表 1 微博上的主题建模方法比较

模型/模式	主题分布层面		扩展方式	实现方式	优势	局限性
	微博用户	微博帖子				
LDA(消息模式)[2]	否	是	无	直接使用	无需监督	主题挖掘不理想
基于用户聚集的 LDA[3]	是	否	过程扩展	文本聚集	解决短文本问题	只限微博用户层面建模,需要人工干预
基于训练的 USER 模式[2]	是	否	过程扩展	文本聚集、分步求解	解决短文本问题,简化推导	需要事先训练和人工干预,若要更新模型需重新训练
基于训练的 TERM 模式[2]	否	否	过程扩展	文本聚集、分步求解	解决短文本问题,简化推导,提高主题可解释性	需要事先训练和人工干预,若要更新模型需重新训练,要求文本具有标签信息
ATM[12] [19]	是	否	模型扩展	文本聚集	解决短文本问题	只限微博用户层面主题建模
ATM 扩展模型[2]	是	是	模型扩展	文本聚集	解决短文本问题	挖掘出的帖子层面主题少且不理想
Twitter-LDA [12] [19]	是	是	模型扩展	文本聚集,引入背景模型	解决短文本问题和高频词汇问题	一个帖子只能对应一个主题
Labeled-LDA [13] [16]	否	是	模型扩展	引入标签信息	提高主题可解释性	要求文本具有足够的标签信息
MB-LDA [18]	部分可以	是	模型扩展	引入结构化信息	解决短文本问题,提高主题可解释性	主要针对会话类和转发类中文微博

5 展 望

微博环境下的主题建模研究仍处于开始阶段,还有许多需要研究解决的问题和难点。本文通过对微博环境下主题建模方法进行梳理发现,现有的微博文本主题建模方法主要研究解决微博文本主题建模的适应性问题。例如,LDA 基于过程的扩展和 ATM、Twitter-

LDA 等模型解决了微博的短文本数据稀疏问题，Labeled-LDA 一定程度上降低了微博文本语法不规范的干扰。然而，面对海量微博文本信息的快速更新，如何实现主题模型在微博环境下的大规模部署和在线学习训练，有待进一步研究解决。为此，引入和探索更加高效的主题模型训练算法具有现实意义，有助于微博环境下主题建模由理论性向实用性的转变。例如，可以借鉴 R. Nallapati 等人提出的并行变分期望最大化算法[20]、A. Asuncion 等人提出的 LDA 模型分布式 Gibbs 采样算法等主题模型推导方法，将其应用于多处理器和分布式环境，加速微博环境下主题模型的训练[21]。

另外，从功能性角度来分析微博环境下主题建模能够发现，现有的微博文本主题建模主要用于微博的内容分类、信息过滤、用户推荐等方面的商业应用，以提高内容质量，改善用户体验；而对于社会媒体舆情监测方面的应用研究仍较少涉及。主题模型最主要的功能是抽取出文本的语义主题，可以直接用于文本话题挖掘。话题检测和话题跟踪紧密关联，是网络舆情监测的重要内容。路荣等人已经对主题模型应用于微博类新闻话题的自动识别做了一些研究，但尚未涉及微博文本的话题跟踪演化分析[22]。目前，已经有一些成熟的应用于话题演化分析的 LDA 扩展模型。例如，Wang 等人通过在 LDA 模型中引入了作为观测值的时间随机变量，得到了一个主题随时间变化的主题模型 TOT(topic over time)[23]；D. Blei 等人将当前时刻的模型参数后验作为下一时刻模型参数的条件分布引入主题模型，提出了动态主题模型 DTM(dynamic topic mode)[24]。将此类 LDA 模型的扩展应用于微博类新闻话题的演化分析，从而实现从海量微博文本信息中快速准确地追踪话题的演化，并根据演化及时做出相应的预测，为安全机构的领导决策提供支持，将会是微博环境下主题建模的一个重要研究方向。此外，在网络舆情监测领域的另一个研究热点－文本情感(倾向性)分析方面，也有主题模型的扩展研究。例如，Lin Chenghua 等人提出的联合情感话题模型 JST(joint sentiment/topic model)[25] 能够在无监督的情况下，抽取出文本主题，并同时实现文档级的情感检测；Mei Qiaozhu 等人提出的主题情感混合模型 TSM(topic-sentiment mixture)[26] 可以将抽取出的主题有关的单词分为中性、正面和负面三类，支持话题级的情感检测和情感随时间的动态演化分析。研究者也可以将此类主题模型应用到微博信息的舆情监测和分析中。总而言之，面对 LDA 模型良好的扩展能力和已有的丰硕研究成果，微博类社会媒体环境下的文本主题建模还有很多方面需要改进。

参考文献：

[1] Blei D, Ng A, Jordan M. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003(3): 993－1022.

[2] Hong Liangjie, Davison B. Empirical study of topic modeling in Twitter [C] // Proceedings of the First Workshop on Social Media Analytics(SOMA'10). New York: ACM Press, 2010: 80－88.

[3] Deerwester S, Dumais S, Landauer T, et al. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science, 1990, 41(6): 391－407.

[4] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning, 2001, 42(1): 177－196.

[5] Steyvers M, Griffiths T. Probabilistic topic models [M] //Landauer T, McNamara D, Dennis S, et al. Latent Semantic Analysis: A Road to Meaning. Mahwah: Lawrence Erlbaum Associates, 2007: 424－440.

[6] Griffiths T, Steyvers M. Finding scientific topics [C] //Proceedings of the National Academy of Sciences. Washington D. C.: United States National Academy of Sciences, 2004: 5228－5235.

[7] Tang Jie, Jin Ruoming, Zhang Jing. A topic modeling approach and its integration into the random walk framework for academic search [C] //Proceedings of the 2008 Eighth IEEE International Conference on Data Mining(ICDM '08). Washington: IEEE Computer Society, 2008: 1055－1060.

[8] Lu Yue, Zhai Chengxiang. Opinion integration through semi－supervised topic modeling [C] //Proceedings of the 17th International Conference on World Wide Web. (WWW '08). New York: ACM Press, 2008: 121－130.

[9] Weng Jianshu, Lim Ee-Peng, Jiang Jing, et al. Twitterrank: finding topic－sensitive influential Twitterers [C] // Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10). New York: ACM Press, 2010: 261－270.

[10] Zvi M, Griffiths T, Steyvers M, et al. The author－topic model for authors and documents [C] //Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence(UAI'04). Arlington: AUAI Press, 2004: 487－494.

[11] Zvi M, Chemudugunta C, Griffiths T, et al. Learning author－topic models from text corpora [J]. ACM Transactions on Information Systems, 2010, 28(1): 1－38.

[12] Zhao Wayne Xin, Jiang Jing, Weng Jianshu, et al. Comparing Twitter and traditional media using topic models [C] // Proceedings of the 33rd European Conference on Information Retrieval (ECIR '11). Berlin, Heidelberg: Springer－Verlag, 2011: 338－349.

[13] Ramage D, Dumais S, Liebling D. Characterizing microblogs with topic models [C] // Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM '10). Menlo Park, CA: AAAI, 2010: 130－137.

[14] Ramage D, Hall D, Nallapati R, et al. Labeled lda: A supervised topic model for credit attribution in multi - labeled corpora [C] // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing(EMNLP '09). Stroudsburg: Association for Computational Linguistics, 2009: 248 - 256.

[15] Lacoste - Julien S, Sha F, Jordan M. DiscLDA: Discriminative learning for dimensionality reduction and classification [C] // Proceedings of Neural Information Processing Systems Conference (NIPS'08). Vancouver: NIPS, 2008: 897 - 904.

[16] Quercia D, Askham H, Crowcroft J. TweetLDA: Supervised topic classification and link prediction in Twitter [C] // Proceedings of the 3rd Annual ACM Web Science Conference. New York: ACM Press, 2012: 247 - 250.

[17] Yu L, Asur S, Huberman B. What trends in chinese social . edia [C] // Proceedings of the Fifth International Workshop on Social Network Mining and Analysis (SNA - KDD' 11). San Diego: KDD Press, 2011: 37.

[18] 张晨逸,孙建伶,丁轶群. 基于 MB - LDA 模型的微博主题挖掘[J]. 计算机研究与发展, 2011(10): 1795 - 1802.

[19] Zhao Xin, Jiang Jing, He Jing, et al. Topical keyphrase extraction from Twitter [C] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL - HLT' 11), Portland: ACL Press, 2011: 379 - 388.

[20] Nallapati R, Cohen W, Lafferty J. Parallelized variational EM for latent dirichlet allocation: An experimental evaluation of speed and scalabiliy // Proceedings of the Seventh IEEE International Conference on Data Mining Workshops (ICDMW '07). Washington D. C.: IEEE Computer Society, 2007: 349 - 354.

[21] Asuncion A, Smyth P, Welling M. Asynchronous distributed learning of topic models [C] // Proceedings of the Twenty - Second Annual Conference on Neural Information Processing Systems. New York: Curran Associates, Inc, 2008: 81 - 88.

[22] 路荣,项亮,刘明荣,等. 基于隐主题分析和文本聚类的微博客新闻话题发现研究 [C] //中国中文信息处理学会. 第六届全国信息检索学术会议论文集. 北京: 中国中文信息处理学会, 2010: 291 - 298.

[23] Wang X, McCallum A. Topic over time: A non - markov continuous - time model of topical trends [C] //Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'06). New York: ACM Press, 2006: 424 - 433.

[24] Blei D, Lafferty J. Dynamic topic model [C] //Proceedings of the 23rd International Conference on Machine Learning. New York: ACM Press, 2006: 113 - 120.

[25] Lin Chenghua , He Yulan. Joint sentiment/topic model for sentiment analysis [C] // Proceeding of the 18th ACM conference on Information and Knowledge Management. New York: ACM Press, 2009: 375 - 384.

[26] Mei Qiaozhu, Xu Ling, Wondra M, et al. Topic sentiment mixture: Modeling facets and opinions in weblogs [C] // Proceedings of the 16th International Conference on World Wide Web. New York: ACM Press, 2007: 171 - 180.

〔作者简介〕张培晶,男,1979 年生,讲师,发表论文 10 余篇。
宋　蕾,女,1980 年生,讲师,发表论文 10 余篇。

(上接第 76 页)

[14] Xu W X. Zipf's law and mechanism of distribution of Chinese term frequency [C] //Proceedings of the 2nd International Conference on Bibliometrics, Scientometrics and Informetrics, London, 1989: 332 - 338.

[15] Le Quan Ha, Hanna Philip Ming Jik, Simth F J. Extending Zipf's law to n-grams for large corpora [J]. Artificial Intelligence Review, 2009(32): 101 - 113.

[16] 杨波,阎素兰. 齐普夫定律的汉语适用性研究及其在自动标引中的应用 [J]. 情报理论与实践, 2004(3): 252 - 255.

[17] Hill B. The rank-frequency form of Zipf's law [J]. Journal of the American Statistical Association, 1987, 69(348): 1017 - 1026.

[18] Cancho R F, Solé R V. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited [J]. Journal of Quantitative Linguistics, 2010, 3(8): 165 - 173.

[19] Cancho R F, Sole R V. Zipf's law and random texts [J]. Advances in Complex Systems, 2002, 5(1): 1 - 6.

[20] Adamic L A. Zipf, power-law, pareto-a ranking tutorial [EB/OL]. [2011 - 02 - 22]. http://www.hpl.hp.com/research/idl/papers/ranking, 2011.

[21] Rousseau R, Zhang Qiaoqiao. Zipf's data on the frequency of Chinese words revisited [J]. Scientometrics, 1992, 24(2): 201 - 220.

[22] 游荣彦. Zipf 定律与汉字字频分布 [J]. 中文信息学报, 2000, 14(3): 60 - 65.

〔作者简介〕路高飞,男,1988 年生,硕士研究生,发表论文 3 篇。
韩　普,男,1983 年生,博士研究生,发表论文 10 篇。
沈　思,女,1983 年生,博士研究生,发表论文 7 篇。