基于 MB-HDP 模型的微博主题挖掘

刘少鹏 印 鉴 欧阳佳 黄 云 杨晓颖
(中山大学信息科学与技术学院计算机科学系 广州 510006)

摘 要 主题模型是挖掘微博潜在主题的重要工具.然而,现有的主题模型多由 Latent Dirichlet Allocation(LDA)派生,它需要用户预先指定主题数目.为了自动挖掘微博主题,作者提出了一个基于分层 Dirichlet 过程(Hierarchical Dirichlet Process,HDP)的非参数贝叶斯模型 MB-HDP.首先,针对微博应用场景,假设消息是不可交换的;接着,利用微博的时间信息、用户兴趣以及话题标签,聚合主题相关的消息以解决微博短文本的数据稀疏问题;然后,扩展 Chinese Restaurant Franchise (CRF)对微博数据进行主题建模;最后,设计一个相应的 Markov Chain Monte Carlo (MCMC)采样方法,推导 MB-HDP 模型的分布参数.实验表明,在生成主题质量、内容困惑度和模型复杂度等指标上,MB-HDP 模型明显优于 LDA 和 HDP 两种模型.

关键词 主题挖掘;微博;分层 Dirichlet 过程;MB-HDP
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2015.01408

Topic Mining from Microblogs Based on MB-HDP Model

LIU Shao-Peng YIN Jian OUYANG Jia HUANG Yun YANG Xiao-Ying
(*Department of Computer Science, School of Information Science and Technology, Sun Yat-Sen University, Guangzhou* 510006)

Abstract Topic models have become important tools to mine latent topics from microblogs. However, most existing models are derived from Latent Dirichlet Allocation (LDA) and require a pre-determined number of topics. In order to mine topics from microblogs automatically, we propose a hierarchical Bayesian nonparametric model named MicroBlog-Hierarchical Dirichlet Process (MB-HDP). Firstly, our model assumes non-exchangeability of data which is suitable for the microblog application. Secondly, to tackle the sparsity problem caused by the short tweets, the temporal information, user's interests, and semantic #hashtags are integrated to aggregate topic-related tweets into lengthy pseudo-documents. Thirdly, the Chinese Restaurant Franchise (CRF) extension is adopted in modeling topics. Finally, we present a Markov Chain Monte Carlo (MCMC) sampling for posterior inference in the MB-HDP. Experimental results show that the MB-HDP clearly outperformed both LDA and HDP from three different perspectives: the quality of generated latent topics, the perplexity of held-out content and the model complexity.

Keywords topic mining; microblog; hierarchical Dirichlet process; MB-HDP

1 引 言

Twitter 是全球最流行的微博服务,它允许用户通过网页、WAP 页面、外部程序和手机短信等发布 140 字符以内的消息,实现信息分享.截止到 2012 年,Twitter 注册用户数已超过 5 亿,每天发布的消息达到 3.4 亿条.海量的微博数据蕴含了丰富

收稿日期:2013-01-30;最终修改稿收到日期:2014-12-25.本课题得到国家自然科学基金(61033010,61272065,61472453,U1401256)、广东省自然科学基金(S2011020001182,S2012010009311)、广东省科技计划项目(2011B040200007,2011B031700004,2012A010701013)资助.**刘少鹏**,男,1984 年生,博士,主要研究方向为数据挖掘、主题模型.E-mail:l-shaopeng@live.cn.**印 鉴**(通信作者),男,1968 年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为数据库、数据挖掘、人工智能.E-mail:issjyin@mail.sysu.edu.cn.**欧阳佳**,男,1986 年生,博士,主要研究方向为数据挖掘、隐私保护.**黄 云**,男,1976 年生,博士研究生,副教授,主要研究方向为数据挖掘、图挖掘.**杨晓颖**,男,1987 年生,硕士,主要研究方向为数据挖掘、主题模型.

的信息，为信息提取提供了机遇，同时也带来了巨大挑战. 微博潜在主题挖掘能够帮助用户及时获取信息，掌握社区动态，是当前的研究热点[1-6].

主题模型是挖掘微博潜在主题的重要工具，主要包括参数贝叶斯模型与非参数贝叶斯模型. Latent Dirichlet Allocation(LDA)[7]是典型的参数贝叶斯模型，它通过词项在文档集中的共现信息，抽取出语义相关的主题集合，有效地挖掘文档集的潜在主题. 然而，相比传统文本，微博文本长度较短，词项共现信息匮乏，导致数据十分稀疏，难以构建微博数据的主题结构. 针对该问题，研究者通过利用微博的时间信息和结构化数据，提出若干 LDA 派生模型[8-10]. 但是，这些模型需要预先指定主题数目，而主题数目的选取直接影响了主题挖掘效果. 在不具备任何先验知识的情况下，用户难以准确地估算出主题数目，因此人为设定主题数目加重了用户的负担，进而限制模型的应用推广.

非参数贝叶斯模型分层 Dirichlet 过程(Hierarchical Dirichlet Process，HDP)[11]常用于聚类问题，它能自动确定聚类的数目，并估计聚类的分布参数. 由 HDP 派生的各种模型，已广泛应用于文本挖掘[11-12]、音乐内容识别[13]、图像检索[14]以及视频监控数据处理[15]等. 在文本主题挖掘中，HDP 能够自动确定主题数目，准确估计文档集的分布参数，从而取得良好的主题挖掘效果. 但是，HDP 并不适用于微博数据，原因在于：HDP 假设数据具有可交换性，即在文档层次上，文档内的单词次序可交换且不影响模型的训练结果；在文档集的层次上，文档可交换性假设文档次序与模型的推导结果无关. 由于消息是严格按照时间先后顺序发布的，当前消息可能是针对已发布的消息进行评论或者转发，这与文档可交换性假设相矛盾. 因此，考虑时间信息是微博主题挖掘的关键. 比如，主题“Obama's health care Bill”探讨的医疗改革法案签署于 2010 年 3 月 23 日，当天发布的消息与该主题相关的可能性很大. 近年来，研究者逐渐意识到时间信息的重要性，并提出若干 HDP 改进模型[12,16-19]. 除此之外，用户兴趣和话题标签等信息也有助于微博的主题挖掘. 比如，用户“macworld”主要关注苹果公司的产品，发布的消息大多涉及“apple”、“iphone”和“mac”等主题，与其自身兴趣紧密相关. 而根据消息中的“#heat”话题标签，可推断该消息可能涉及“basketball”、“Miami”和“Wade”等主题. 因此，文献[20]引入话题标签作为附加的弱监督信息，以改进微博的潜在主题挖掘.

针对微博潜在主题的自动挖掘任务，本文提出了一个基于分层 Dirichlet 过程的非参数贝叶斯模型 MB-HDP. 首先，MB-HDP 模型根据微博应用场景的特点，假设消息是不可交换的；接着，利用微博的时间信息、用户兴趣以及话题标签，聚合主题相关的消息以解决数据稀疏问题；然后，扩展 Chinese Restaurant Franchise(CRF)[11]对微博数据进行主题建模；最后，设计一个相应的 Markov Chain Monte Carlo(MCMC)采样方法，推导 MB-HDP 模型的分布参数，挖掘微博的潜在主题. 实验表明，在生成主题质量、内容困惑度和模型复杂度 3 个指标上，MB-HDP 模型明显优于 LDA 和 HDP 等现有主题模型.

本文主要贡献如下：(1) 利用时间信息、用户兴趣和话题标签，有效地解决了微博潜在主题的自动挖掘问题；(2) 提出基于分层 Dirichlet 过程的非参数贝叶斯模型 MB-HDP，并设计相应的 MCMC 采样方法推导该模型；(3) 在真实数据集上进行大量实验，验证 MB-HDP 模型的有效性.

本文第 2 节回顾相关的研究工作；第 3 节定义微博潜在主题的自动挖掘问题，并详细阐述 MB-HDP 模型的框架；第 4 节是关于 MB-HDP 模型在真实数据集上的实验分析与讨论；最后第 5 节给出结论和未来的工作.

2　相关工作

主题模型是一种概率生成模型[21-22]，常用于挖掘大规模文档集的潜在主题. 词项作为自然语言的基本单元，往往隐含了潜在主题的语义信息. 字面上毫不相干的单词，却可能描述同一个抽象概念. 基于这种思想，主题模型假设主题遵循一定的规则生成单词. 在文档单词已知的情况下，可以通过概率反推文档的主题结构，进而得到整个文档集的主题分布. 主题模型通常划分为参数贝叶斯模型与非参数贝叶斯模型两大类.

2.1　参数贝叶斯模型

LDA[7]是典型的参数贝叶斯模型，它将文档视为主题的概率分布，主题则是单词的概率分布. 给定文档集的主题数目，每个文档的生成过程如下：首先，从先验 Dirichlet 分布中抽取该文档的主题分布；接着，从主题多项式分布中选择当前单词的主题；最后，从先验 Dirichlet 分布中抽取该主题的单词分布，并选择具体单词. 重复以上步骤，直到生成所有单词. LDA 常用的推导方法包括变分贝叶斯和

Gibbs 抽样等. 本质上,LDA 利用词项在文档层次上的共现信息揭示文档集所蕴含的主题结构. 由于微博文本长度较短,词项共现信息匮乏,导致了严重的数据稀疏问题,因而微博的主题结构难以直接估计. 所以,将 LDA 直接应用于微博数据无法取得令人满意的主题挖掘效果.

近年来,研究者通过利用微博的时间信息和结构化数据,开发了若干 LDA 派生模型,在一定程度上克服了数据稀疏的困难,提高了微博的主题挖掘效果. 文献[23]提出了一个半监督学习模型:Labeled LDA,它将消息内容映射到 substance、style、status 和 social 这 4 个维度,从而更好地表示文本内容,适用于用户特征描述与消息组织等任务. 文献[9]假设用户发布消息的行为受到突发性新闻、朋友消息和用户兴趣 3 个因素影响,设计了一个主题混合模型框架,实现消息发布行为的建模. 文献[10]介绍了一个微博主题挖掘的概率生成模型 MB-LDA,同时考虑微博的联系人关联关系和文本关联关系,以此辅助进行微博的主题挖掘. 实验表明该模型不仅能找出微博的潜在主题,还能发现联系人关注的主题. 文献[24]针对微博主题建模和主题转移问题,设计了一个时态感知的主题模型 TM-LDA. 该模型的核心思想是最小化消息序列主题分布的预测误差,从而求解主题转移的分布参数. 文献[8]基于"用户发布的消息与日常生活和个人兴趣紧密相关"以及"特定时间段的消息倾向于谈论热点新闻事件"两个现象,提出了主题模型 TimeUserLDA. 该模型假设消息只描述单一的主题,融合时间信息和用户兴趣,准确地发现微博突发性话题.

上述参数贝叶斯模型在训练阶段要求用户指定主题数目. 然而,假设用户对微博数据的潜在主题具备的先验知识并不合理. 由于主题数目随着时间动态变化,尝试搜索合适的主题数目往往耗时且低效. 考虑到非参数贝叶斯模型允许无限的主题数目,它们无疑是更适合微博主题自动挖掘的解决方案.

2.2 非参数贝叶斯模型

HDP[11]是典型的非参数贝叶斯模型,它能够揭示高维数据所隐藏的潜在语义结构. 在 HDP 的主题建模过程中,每个文档是一组可观察单词的集合. 单词按照主题聚类,最终确定文档的主题分布. 各个文档的主题服从同一个 Dirichlet 过程,从而保证文档集共享相同的主题;文档集的主题数目不限,具体数值在模型分布参数推导过程中自动确定. HDP 的采样常用 Stick-breaking 和 CRF 等构造方法实现,而模型参数后验分布推断方法主要包括 MCMC 和变分贝叶斯等. 由于 HDP 忽略数据集的时间等附加信息,简单地假设数据具有可交换性,因此,当数据集的主题表现出明显的时间模式时,HDP 的主题挖掘效果不佳.

近年来,研究者提出了若干 HDP 派生模型,通过使用数据集的时间信息辅助主题挖掘. 文献[16]针对时序数据随时间演变的统计性质,提出了一个动态分层 Dirichlet 过程 dHDP. 该模型为每个时间间隔建立混合分布,并保证初始混合分布与后续混合分布共享相同的主题. 文献[17]介绍了一个动态主题模型 dDTM,分析带有时间戳信息的文档集潜在主题. dDTM 模型按照特定的时间间隔划分整个文档集,每个子集的文档是由混合分布的某个主题生成,针对 NIPS 论文集的实验验证了该模型的有效性. 文献[12]设计了一个演化分层 Dirichlet 过程 EvoHDP,从多个相互联系的时变文档集中,发现有趣的主题演变模式,包括主题的形成、演变以及消亡. EvoHDP 模型由一系列携带时间信息的 HDP 构成,每个 HDP 依赖相邻的 HDP,并采用级联 Gibbs 采样模式推断. 文献[18]指出时序数据流的主题数目、分布和流行度等潜在语义结构是随时间演变的,为了描述这些演变规律,提出了一个非确定性动态主题模型 iDTM. 该模型根据时间间隔组织文档集,保持文档的时间顺序关系,消除了文档可交换的假设,最终在 NIPS 论文集上取得令人满意的结果. 文献[19]结合数据集的时间信息,提出了一个基于 HDP 的 distance dependent Chinese Restaurant Franchise(ddCRF)模型. 该模型由文档层次的 Chinese Restaurant Process(CRP)[25]和文档集层次的 distance dependent Chinese Restaurant Process(ddCRP)[26]构成. 其中,CRP 假设文档内的单词是可交换的,按照主题对单词进行聚类;而 ddCRP 假设文档之间具有不可交换性,首先计算各文档的时间间隔,再根据已定义的衰减函数求解文档之间的依赖关系,以便实现文档集的主题聚类. 实验表明 ddCRF 模型能够捕捉时序数据集潜在主题的时间模式,在学术论文集上的主题挖掘效果尤为突出.

尽管上述 HDP 派生模型利用了时间信息,且在学术论文潜在主题的挖掘上效果显著,但是对于微博数据,仅考虑主题的时间模式还不足以实现主题的准确建模,也无法有效地解决数据稀疏问题.

3　MB-HDP 模型

针对现有模型的不足，考虑到微博的时间信息、用户兴趣和话题标签反映了消息的主题分布情况，有助于聚合主题相关的消息以克服数据稀疏的困难.本文提出基于 HDP 的 MB-HDP 模型，充分利用时间信息、用户兴趣和话题标签，假设微博数据不可交换，并扩展 ddCRF 进行主题建模，实现微博主题自动挖掘.

3.1　问题描述

在微博主题自动挖掘问题中，消息集合表示为 $x=\{x_1,x_2,\cdots,x_J\}$，其中 J 是消息总数目，消息按照时间顺序排列.第 j 条消息 x_j 由 N_j 个单词组成，即 $x_j=\{x_{j1},x_{j2},\cdots,x_{jN_j}\}$，其中 $x_{ji}\in\{1,2,\cdots,V\}$，$V$ 表示词典的大小.x_j 由用户 u_j 生成，其中 $u_j\in\{1,2,\cdots,U\}$，U 表示用户集合的大小.x_j 的发布时间是 ts_j，其中 $ts_j\in\{1,2,\cdots,TS\}$，TS 表示时间戳集合的大小.x_j 与特定的话题标签 ht_j 关联，其中 $ht_j\in\{0,1,\cdots,HT\}$，HT 表示话题标签集合的大小.注意，如果 x_j 包含若干话题标签，则选取第一个标签；如果 x_j 不存在话题标签，则 ht_j 标记为 0.假设消息可由无限个主题的概率分布表示，而主题是词汇表中所有单词的概率分布，则微博主题自动挖掘指的是从消息集合 x 中自动发现 K 个潜在主题 $\Phi=(\phi_k)_{k=1}^{K}$.

3.2　HDP 主题模型

HDP 本质上是 Dirichlet 过程混合模型的多层形式，现已成为文本主题挖掘的重要工具.HDP 假定文档集共享相同的主题，且主题数目不限.相比参数贝叶斯模型，HDP 具有良好的鲁棒性和灵活性.下面详细介绍基于文本主题挖掘的两层 HDP 模型.首先，从基分布 H 和 Concentration 参数 γ 构成的 Dirichlet 过程中，抽样分布 G_0.其次，从基分布 G_0 和 Concentration 参数 α_0 构成的 Dirichlet 过程中，为每一个文档抽取主题分布 G_j.

$$\begin{aligned} G_0 &\sim DP(\gamma,H)\\ G_j\mid G_0 &\sim DP(\alpha_0,G_0)\end{aligned}\tag{1}$$

从式(1)可看出，各个文档的主题均服从基分布 H，保证了各个文档之间的主题共享.通常，基分布 H 选取参数为 η 的 Dirichlet 分布.每个主题 ϕ_k 是 H 的一个独立抽样，本质上是关于单词的概率分布.在文档 x_j 中，$(\theta_{ji})_{i=1}^{N_j}$ 是服从 G_j 的独立同分布的随机变量序列，θ_{ji} 指示了单词 x_{ji} 所分配的主题，隶属于 Φ.文档 x_j 的生成过程如下：

$$\begin{aligned} \theta_{ji} &\sim G_j\\ x_{ji}\mid\theta_{ji} &\sim F(\theta_{ji})\end{aligned}\tag{2}$$

其中，$F(\theta_{ji})$ 表示给定参数 θ_{ji} 下，单词 x_{ji} 的分布.为简化 HDP 采样过程的计算，F 常选取为多项式分布，与基分布 H 构成共轭分布.

为了实现 HDP 的采样，需设计相应的构造方法，以便推断模型参数后验分布.CRF 是其中一种经典的构造方法，已得到广泛应用.在 CRF 中，J 个餐厅共用一份相同的菜单：$\Phi=(\phi_k)_{k=1}^{K}$，K 表示菜肴数；第 j 个餐厅可容纳 m_j 张餐桌：$(\psi_{jt})_{t=1}^{m_j}$，每张餐桌至多可容纳 N_j 个顾客.顾客可自由选择餐桌，每张餐桌只供应一道菜.餐桌的第一位的顾客负责点菜，该餐桌的其他客人共同享用该菜肴.不同餐厅的不同餐桌，或者是同一餐厅的不同餐桌均可点用同一道菜.在文本主题挖掘应用中，餐厅、顾客和菜肴分别对应了文档、单词和主题.假设 δ 表示一个概率测度，将单词 x_{ji} 的分布参数 θ_{ji} 视为顾客，他以概率 $\frac{n_{jt}}{i-1+\alpha_0}$ 就座于餐桌 ψ_{jt}，并享用该餐桌提供的菜肴 ϕ_k，或者以概率 $\frac{\alpha_0}{i-1+\alpha_0}$ 就座于新餐桌 $\psi_{jt_{\text{new}}}$.其中，n_{jt} 表示第 j 个餐厅里第 t 张餐桌的顾客数.如果顾客选择一张新餐桌，则根据已选用菜肴的受欢迎程度，以概率 $\frac{m_k}{\sum_k m_k+\gamma}$ 为新餐桌指定菜肴 ϕ_k，或者以概率 $\frac{\gamma}{\sum_k m_k+\gamma}$ 选用新菜肴 $\phi_{k_{\text{new}}}$.其中，m_k 表示供应 ϕ_k 的餐桌数.

$$\theta_{ji}\mid\theta_{j1},\theta_{j2},\cdots,\theta_{j,i-1},\alpha_0,G_0\sim \sum_{t=1}^{m_j}\frac{n_{jt}}{i-1+\alpha_0}\delta_{\psi_{jt}}+\frac{\alpha_0}{i-1+\alpha_0}G_0\tag{3}$$

$$\psi_{jt}\mid\psi_{11},\psi_{12},\cdots,\psi_{21},\cdots,\psi_{j,t-1},\gamma,H\sim \sum_{k=1}^{K}\frac{m_k}{\sum_k m_k+\gamma}\delta_{\phi_k}+\frac{\gamma}{\sum_k m_k+\gamma}H\tag{4}$$

可见，CRF 的构造即是为顾客分配餐桌和菜肴的过程.首先为每个顾客分配餐桌，已有餐桌被选中的概率与其就座的顾客数成正比，而新餐桌也允许以一定的概率被选中.在完成餐桌指派后，为每张餐桌分配菜肴，已有菜肴被选中的概率与其供应的餐桌数成正比，而新菜肴也允许以一定的概率被选中.CRF 构造过程对应了文档内单词的主题指派以及文档集的主题聚类，一旦完成 CRF 构造，即可采用

模型参数后验分布推断方法(比如 MCMC 方法)求解 HDP 主题模型,进而获取整个文档集的主题分布.

3.3 MB-HDP 模型的框架

针对微博主题的自动挖掘问题,本文在 HDP 的基础上,融合时间信息、用户兴趣和话题标签,提出 MB-HDP 模型.相比现有的主题模型,MB-HDP 模型具有以下优点:首先,参数贝叶斯模型需要用户预先指定主题数目;而 MB-HDP 模型在主题建模过程中自动搜索合适的主题数目.其次,HDP 等模型简单地认为数据具有可交换性,影响了主题挖掘的效果;而 MB-HDP 模型根据微博应用场景的特点,假设消息是不可交换的.最后,现有的 HDP 派生模型处理的多是新闻和学术论文等传统文本;而 MB-HDP 模型针对的是微博数据.考虑消息的短文本特点,深入分析时间信息、用户兴趣和话题标签对主题建模的关键作用,本文提出如下假设:(1)如果消息的发布时间相同,则它们很可能谈论全局的新闻事件,并共享该事件涉及的主题分布;(2)如果消息的发布用户相同,则它们很可能反映了该用户的自身兴趣,并共享该用户兴趣的主题分布;(3)如果消息包含相同的话题标签,则它们很可能与该标签反映的抽象概念或者突发性新闻有关,并共享其内在的主题分布.MB-HDP 模型根据上述假设聚合主题相关的消息,从而丰富了词项的共现信息,有效地克服了数据稀疏的困难.而 ddCRF 等现有模型只利用时间信息,如果直接应用于微博数据,将面临严重的数据稀疏问题.

类似 HDP 的 CRF 构造,MB-HDP 模型采用改进的 CRF 构造.该构造方法由两层 CRP 构成:在文档层次上,使用 CRP 为顾客分配餐桌,而后将顾客按照所在餐桌进行划分,即完成文档内单词的聚类;在文档集层次上,使用 ddCRP 替换 CRP,餐桌供应的菜肴取决于相互关联的其他餐桌,以满足文档的不可交换性约束.而后将餐桌按照供应的菜肴进行聚类,即得到每个文档的主题分布.在改进的 CRF 中,同一个餐厅的顾客之间不存在依赖关系,但餐桌供应的菜肴受相互关联的其他餐桌的影响.具体来说,如果两个餐桌所在餐厅的位置彼此靠近(即消息的发布时间相同),或者拥有者是同一人(即消息的发布用户相同),或者招牌菜相同(即消息的话题标签相同),则认为它们是相互关联的.据此,MB-HDP 模型将时间信息、用户兴趣和话题标签融入到 CRF 构造过程,有效地聚合消息并辅助主题建模.基于改进 CRF 的 MB-HDP 模型的框架如图 1 所示.在文档层次的 CRP 中,长方形表示餐厅/文档,附加信息包括时间戳 ts、用户 u 和话题标签 ht;大圆表示餐桌,对应变量 ψ_{ji};小圆表示单词 x_{ji} 的分布参数 θ_{ji}.在文档集层次的 ddCRP 中,大圆表示菜肴/主题 ϕ_k;小圆、正方形和菱形均表示餐桌,分别表示在时间信息、用户兴趣和话题标签的影响下,餐桌按照其供应的菜肴进行聚类.

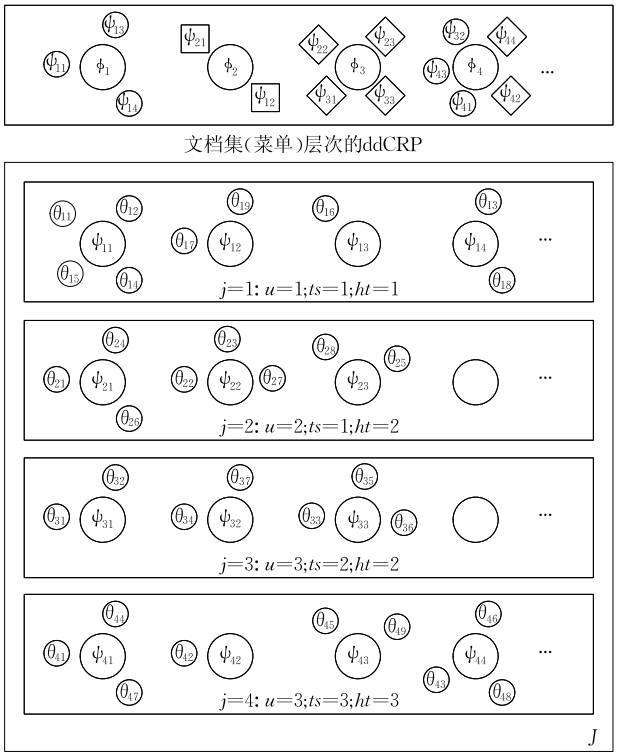

图 1 基于改进 CRF 的 MB-HDP 模型的框架

基于改进的 CRF 构造方法和 MB-HDP 模型框架图,对微博数据的生成过程进行解构.符号定义如下:t_{ji} 为单词 x_{ji} 所在餐桌的索引,$t=\{t_{ji}:\forall(j,i)\}$;$k_{jt}$ 是餐桌 ψ_{jt} 所供应菜肴的索引,$k=\{k_{jt}:\forall(j,t)\}$;$z_{ji}$ 表示 x_{ji} 的主题索引,$z=\{z_{ji}:\forall(j,i)\}$;$n_k$ 和 n_k^v 分别是主题 ϕ_k 中所有单词和单词索引为 v 的个数;分布参数 θ_{ji} 对应 x_{ji} 所在的餐桌,$\Theta=\{\theta_{ji}:\forall(j,i)\}$;餐桌 ψ_{jt} 供应某道菜肴 ϕ_k,$\Psi=\{\psi_{ji}:\forall(j,i)\}$;$x_{jt}$ 表示餐桌 ψ_{jt} 的所有顾客.约定某一变量的上角标有负号时,表示移除相应的变量或变量集.比如,t^{-ji} 表示除 t_{ji} 外的所有餐桌索引,n_{jt}^{-ji} 表示除 x_{ji} 外分配到餐桌 ψ_{jt} 的顾客总数,$n_k^{-x_{jt}}$ 表示除 x_{jt} 外享用菜肴 ϕ_k 的顾客总数.

MB-HDP 模型的生成过程包括 3 个主要步骤:第 1 步,生成全部单词所对应的分布参数 θ_{ji}.由于采用 CRP 对消息建模,θ_{ji} 的条件分布依赖 Θ 中剩余

的变量. 式(3)描述了详细的抽样过程：如果根据第 1 项抽取 θ_{ji}，则表示顾客就座于已有餐桌，令 $\theta_{ji}=\psi_{jt}$，$t_{ji}=t$；如果根据第 2 项抽取 θ_{ji}，则表示顾客就座于新餐桌，令 $m_j=m_j+1$，同时由分布 G_0 采样得到新餐桌 $\psi_{jm_j}\sim G_0$，令 $\theta_{ji}=\psi_{jm_j}$，$t_{ji}=m_j$. 第 2 步，生成每张餐桌 ψ_{jt} 供应的菜肴，实现消息集合主题划分. 由于采用 ddCRP 对消息集合建模，消息的主题指派受相互依赖的其他消息影响，ψ_{jt} 的条件分布取决于 Ψ 中相互关联的剩余变量：

$$\psi_{jt}\mid\psi_{11},\psi_{12},\cdots,\psi_{21},\cdots,\psi_{j,t-1},\gamma,H\sim\sum_{k=1}^{K}\frac{\sum\limits_{j't'\neq jt,k_{j't'}=k}S_{j't',jt}}{\sum\limits_{j't'\neq jt}S_{j't',jt}+\gamma}\delta_{\phi_k}+\frac{\gamma}{\sum\limits_{j't'\neq jt}S_{j't',jt}+\gamma}H \tag{5}$$

如果根据第 1 项抽取 ψ_{jt}，则表示顾客选择已有菜肴，令 $\psi_{jt}=\phi_k$，$k_{jt}=k$. 如果根据第 2 项抽取 ψ_{jt}，则表示顾客选择新菜肴，令 $K=K+1$，同时由分布 H 采样得到新菜肴 $\phi_K\sim H$，令 $\psi_{jt}=\phi_K$，$k_{jt}=K$. 式(5)中的 $S_{j't',jt}$ 反映了餐桌 $\psi_{j't'}$ 和 ψ_{jt} 之间的关联程度，其计算方法由消息的发布时间、发布用户和话题标签三者共同决定，该值大小与餐桌之间的关联程度成正比.

$$S_{j't',jt}=\mathbf{1}[ts_{j'}=ts_j]+\mathbf{1}[u_{j'}=u_j]+\mathbf{1}[ht_{j'}=ht_j] \tag{6}$$

第 3 步，生成每个可观察的单词 x_{ji}. 给定 x_{ji} 的主题索引为 k，则生成 x_{ji} 的条件概率为 $f_k(\{x_{ji}:z_{ji}=k\})$. 由于基分布 H 为 Dirichlet 分布，主题的单词分布 F 为多项式分布，它们互为共轭分布；再利用 Γ 函数的递推性质，可得

$$f_k(\{x_{ji}:z_{ji}=k\})=\frac{\Gamma(V\eta)}{\Gamma(n_k+V\eta)}\frac{\prod\limits_v\Gamma(n_k^v+\eta)}{\Gamma^V(\eta)} \tag{7}$$

一旦 $f_k(\{x_{ji}:z_{ji}=k\})$ 已知，给定 t 和 k，则消息集合 x 的条件分布为

$$p(x\mid t,k)=\prod_k f_k(\{x_{ji}:z_{ji}=k\}) \tag{8}$$

3.4　MB-HDP 模型的采样

本小节主要分析 MCMC 采样方法对 MB-HDP 模型实现采样. MCMC 方法基于改进的 CRF，根据可观测变量 x，对所有的分布参数 θ_{ji} 和 ψ_{jt} 采样. 为了方便起见，对餐桌索引 t 和菜肴索引 k 采样，而不是直接采样 Θ 和 Ψ. 一旦索引变量已知，容易重建相应的聚类，进而得到消息集合的主题分布以及主题的单词分布.

为了计算索引变量后验概率，需要先求解 x_{ji} 和 x_{jt} 的条件概率. 假设先验 Dirichlet 分布 H 以概率 $h(\phi_k\mid\eta)$ 抽样主题 ϕ_k，多项式分布 F 以概率 $f(x_{ji}\mid\phi_k)$ 从主题 ϕ_k 中抽样单词 x_{ji}. 给定主题 ϕ_k 中除 x_{ji} 以外的所有单词，x_{ji} 的条件概率为

$$f_k^{-x_{ji}}(x_{ji})=p(x_{ji}\mid x^{-ji},t,k)=\frac{p(x\mid t,k)}{p(x^{-ji}\mid t,k)}=\frac{\int f(x_{ji}\mid\phi_k)\prod\limits_{j'i'\neq ji,z_{j'i'}=k}f(x_{j'i'}\mid\phi_k)h(\phi_k\mid\eta)\mathrm{d}(\phi_k)}{\int\prod\limits_{j'i'\neq ji,z_{j'i'}=k}f(x_{j'i'}\mid\phi_k)h(\phi_k\mid\eta)\mathrm{d}(\phi_k)} \tag{9}$$

由于基分布 H 和主题的单词分布 F 是共轭分布，式(9)可进一步简化：

$$f_k^{-x_{ji}}(x_{ji}=v)=\frac{n_k^{-x_{ji},v}+\eta}{n_k^{-x_{ji}}+V\eta} \tag{10}$$

其中，$n_k^{-x_{ji},v}$ 表示除去 x_{ji}，主题 ϕ_k 中单词索引为 v 的总数. 给定主题 ϕ_k 中除 x_{jt} 以外的所有单词，x_{jt} 的条件概率 $f_k^{-x_{jt}}(x_{jt})$ 计算如下：

$$f_k^{-x_{jt}}(x_{jt})=\frac{\Gamma(n_k^{-x_{jt}}+V\eta)}{\Gamma(n_k^{-x_{jt}}+n^{-x_{jt}}+V\eta)}\frac{\prod\limits_v\Gamma(n_k^{-x_{jt},v}+n^{x_{jt},v}+\eta)}{\prod\limits_v\Gamma(n_k^{-x_{jt},v}+\eta)} \tag{11}$$

(1) t 采样. 考虑 t_{ji} 继承了 θ_{ji} 的可交换性，其后验概率正比于餐桌 ψ_{jt} 的顾客人数与 x_{ji} 条件概率之积：

$$p(t_{ji}=t\mid t^{-ji},k,x)\propto\begin{cases}n_{jt}^{-ji}\cdot f_{k_{jt}}^{-x_{ji}}(x_{ji}), & t_{ji}\in[1,m_j]\\ \alpha_0\cdot p(x_{ji}\mid t_{ji}=t_{\mathrm{new}},t^{-ji},k), & t_{ji}=t_{\mathrm{new}}\end{cases} \tag{12}$$

在式(12)中，$f_{k_{jt}}^{-x_{ji}}(x_{ji})$ 可由式(10)计算. 当 $t_{ji}=t_{\mathrm{new}}$ 时，表示顾客选择一张新餐桌，可观测单词 x_{ji} 的条件分布由餐桌之间的关联关系决定：

$$p(x_{ji}\mid t_{ji}=t_{\mathrm{new}},t^{-ji},k)=\sum_{k=1}^{K}\frac{\sum\limits_{j't'\neq jt,k_{j't'}=k}S_{j't',jt}}{\sum\limits_{j't'\neq jt}S_{j't',jt}+\gamma}\cdot f_{k_{jt}}^{-x_{ji}}(x_{ji})+\frac{\gamma}{\sum\limits_{j't'\neq jt}S_{j't',jt}+\gamma}\cdot f_{k_{\mathrm{new}}}^{-x_{ji}}(x_{ji}) \tag{13}$$

给定 $t_{ji}=t_{\mathrm{new}}$，x_{ji} 的先验条件概率：

$$f_{k_{\mathrm{new}}}^{-x_{ji}}(x_{ji})=\int f(x_{ji}\mid\phi_{k_{\mathrm{new}}})h(\phi_{k_{\mathrm{new}}}\mid\eta)\mathrm{d}\phi_{k_{\mathrm{new}}} \tag{14}$$

随后，对新餐桌分配菜肴 $\phi_{k_{jt_{\mathrm{new}}}}$：

$$p(k_{jt_{\mathrm{new}}}=k\mid t,k^{-jt_{\mathrm{new}}})\propto\begin{cases}\sum\limits_{j't'\neq jt,k_{j't'}=k}S_{j't',jt}\cdot f_k^{-x_{ji}}(x_{ji}), & k\in[1,K]\\ \gamma\cdot f_{k_{\mathrm{new}}}^{-x_{ji}}(x_{ji}), & k=k_{\mathrm{new}}\end{cases} \tag{15}$$

其中，$\sum_{j't'\neq jt,k_{j't'}=k} S_{j't',jt}$表示菜肴$\phi_k$的受欢迎程度，由供应$\phi_k$的所有餐桌与$\psi_{jt}$的关联程度累加得到.

(2)k采样. 一旦完成所有餐桌的分配，即可对餐桌分配菜肴. k_{jt}采样类似t_{ji}，每次修改$\phi_{k_{jt}}$等价于更新餐桌ψ_{jt}的所有单词的主题. 因此，k_{jt}的后验概率正比于菜肴受欢迎程度与x_{jt}条件概率之积：

$$p(k_{jt}=k\mid t,k^{-jt},x)\propto \begin{cases}\sum_{j't'\neq jt,k_{j't'}=k} S_{j't',jt}\cdot f_k^{-x_{jt}}(x_{jt}), & k\in[1,K]\\ \gamma\cdot f_{k_{\text{new}}}^{-x_{jt}}(x_{jt}), & k=k_{\text{new}}\end{cases} \tag{16}$$

(3)Φ采样. 给定t,k和可观察文档集x，每个主题ϕ_k后验分布概率只依赖该主题下的所有单词：

$$p(\phi_k\mid t,k,x,\Phi^{-k})\propto h(\phi_k\mid\eta)\prod_{ji:k_{jt_{ji}}=k} f(x_{ji}\mid\phi_k) \tag{17}$$

算法 1. 微博主题自动挖掘算法.

输入：消息集合x，分布参数η,α_0和γ

输出：主题数目K，以及主题集合$(\phi_k)_{k=1}^K$

步骤：

1. FOR 所有文档 $x_j\in x$ DO
2. FOR 所有单词 $x_{ji}\in x_j$ DO
3. 根据式(12)，为 x_{ji} 指派餐桌，索引为 t_{ji}.
4. IF $t_{ji}=t_{\text{new}}$ THEN
5. 根据式(15)，为新餐桌分配菜肴，索引为 $k_{jt_{\text{new}}}$.
6. END IF
7. END FOR
8. FOR 所有餐桌 $t_{ji}\in x_j$ DO
9. 根据式(16)，为 t_{ji} 分配菜肴，索引为 k_{jt}.
10. END FOR
11. END FOR
12. 根据式(17)，计算 K 个主题的单词分布$(\phi_k)_{k=1}^K$.
13. RETURN $K,(\phi_k)_{k=1}^K$.

3.5 微博主题自动挖掘算法

MB-HDP 模型通过 MCMC 采样方法，间接求解出消息集合中微博在主题上的概率分布参数Θ，以及主题在单词上的概率分布参数Ψ. 利用这些分布参数，可进一步挖掘整个消息集合的潜在主题，以及每个主题最具有代表性的单词. 在此基础上，设计一个基于 MB-HDP 模型的微博主题自动挖掘算法，具体流程如算法 1 所示.

4 实 验

本小节通过实验评估 MB-HDP 模型的有效性. 实验数据集为真实 Twitter 消息，实验指标是生成主题的质量、内容困惑度和模型复杂度. 为了验证时间信息、用户兴趣和话题标签对主题挖掘效果的重要作用，本文设计了 3 个简化的 MB-HDP 模型与 HDP 作对比.

4.1 数据集

实验数据集包含了 2009 年 9 月到 2010 年 1 月之间的 3845622 条消息[27]. 由于消息是非正式的短文本，内容质量差异大，因此需要采用数据预处理技术过滤低质量的消息. 首先，预先准备一个停用词列表，删除消息中频繁出现而没有实际意义的停用词. 接着，采用 Snowball 算法①提取词干. 将单词转化为词干有利于识别同根词，从而改善主题挖掘效果. 比如“stemmer”、“stemming”和“stemmed”的词干均为“stem”. 最后，删除文档频率低于 20 的单词，保留单词数不小于 8 的消息. 预处理后，得到一个包含 100000 条消息的数据集，详细描述如表 1 所示.

表 1 数据集描述

描述	消息数量	词典大小	用户数量	标签数量	单词数量	消息平均长度	消息最小时间戳/s	消息最大时间戳/s
数据	100000	41051	1476	884	1010351	10	1251734400	1260280443

本文选取 LDA[7]和 HDP[11]作为对比. LDA 的参数包括主题数目K、Dirichlet 分布超参数α和β，经过实验调优分别设置为 50、0.5 和 0.02. 在 HDP 及派生模型中，基分布H的超参数η设为 0.5，concentration 参数由 gamma 先验分布决定：$\gamma\propto\Gamma(1,0.1)$，$\alpha_0\propto\Gamma(1,1)$. 实验平台是 Windows XP 操作系统，Intel Core quad-core (3.10GHz)处理器，内存容量为 4GB.

4.2 主题质量

主题质量评估包括有效性、独特性和多样性等.

(1)主题有效性. 微博主题挖掘的目标是从海量消息中找出有趣的主题. 输出最能代表主题的单词，人工判断它们与描述主题的相关程度，是评价主题模型有效性的典型方法. 限于篇幅，表 2 只给出各个主题模型所找到的 9 个相同主题，每个主题由概率最大的前 10 个单词表示.

图 2 是人工评判主题有效性的投票结果. MB-HDP 模型抽取的主题，52.6%被认为是最准确的；LDA 和 HDP 分别得到 18.4%和 28.9%的投票. 说明MB-HDP模型抽取的单词能够准确表达主题的

① http://snowball.tartarus.org/

表 2　潜在主题的前 10 个单词

Obama Health Care			Real Estate			Free Download Music		
MB-HDP	HDP	LDA	MB-HDP	HDP	LDA	MB-HDP	HDP	LDA
obama	health	health	real	real	blog	free	music	free
#tcot	care	care	estate	home	post	download	itune	hour
health	#tcot	bill	home	estate	real	music	song	deal
#p2	bill	vote	credit	press	estate	song	record	gift
care	senate	senate	bank	release	job	video	album	code
bill	#p2	plan	loan	sale	wed	mp3	free	download
senate	house	image	mortgage	foreclosure	computer	full	smoke	promo
president	reform	insurance	sale	blog	home	promo	download	san
#tlot	#tlot	entertainment	foreclosure	com	today	itune	#ad	certificate
house	vote	cost	rate	www	archive	access	aol	full
Black Friday			**Swine Flu**			**Football Game**		
MB-HDP	HDP	LDA	MB-HDP	HDP	LDA	MB-HDP	HDP	LDA
friday	friday	friday	flu	flu	school	game	game	week
deal	black	black	school	school	county	nfl	nfl	search
black	deal	monday	swine	police	flu	week	football	top
shop	holiday	air	health	swine	student	play	week	pick
com	shop	thanksgiv	h1n1	student	official	football	pick	nfl
monday	sale	energy	student	county	death	pick	sport	football
holiday	monday	show	vaccine	say	local	team	play	sunday
free	store	power	children	kill	high	bear	bear	engine
online	com	deal	death	h1n1	update	coach	team	rank
save	online	clean	county	death	swine	say	college	sec
Job			**Stock Trade Market**			**Tiger Wood**		
MB-HDP	HDP	LDA	MB-HDP	HDP	LDA	MB-HDP	HDP	LDA
job	design	sale	stock	rate	michael	tiger	tiger	tiger
sale	engineer	job	trade	report	trade	wood	wood	wood
service	web	detail	forex	sale	jackson	golf	golf	car
manager	full	system	market	bank	stock	celebrity	file	sport
detail	boston	price	usd	job	daily	file	car	travel
engineer	#job	store	rate	market	forex	justice	tour	fight
health	list	manager	daily	york	dollar	tour	sport	golf
system	sale	engineer	report	financial	global	wife	pga	talk
part	manager	senior	dollar	stock	gold	pga	wife	latest
medical	click	#job	reuter	reuter	million	car	jon	buy

内容. LDA 和 HDP 由于未考虑微博数据的特点，抽取的主题容易相互混淆，不利于用户理解. 另外，LDA 预先设置主题的数目，与实际情况相符的机会较小，导致主题的混淆程度加剧.

(2) 主题独特性. 如图 3 所示，MB-HDP 模型挖掘出最多的潜在主题，HDP 次之，LDA 排在最后. 原因在于 MB-HDP 模型能够区分相似度很高的主题. 比如，主题“World Cup”和“Football Game”被正确区分开来. 再者，MB-HDP 模型能够发现粒度更细的主题. 比如，“Movie”的子主题“Hollywood Movie Star”被正确抽取出来. 表 3 展示了 MB-HDP 模型检测到的特有主题.

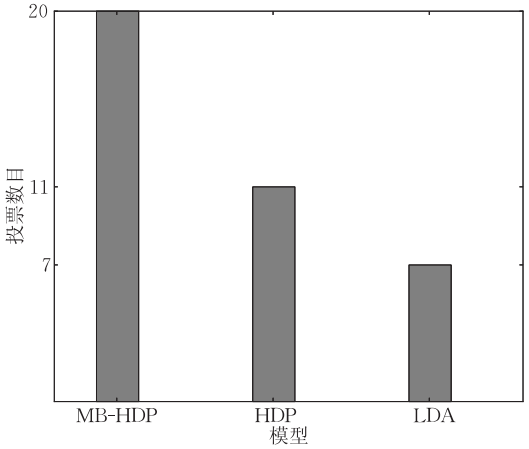

图 2　主题有效性的投票数

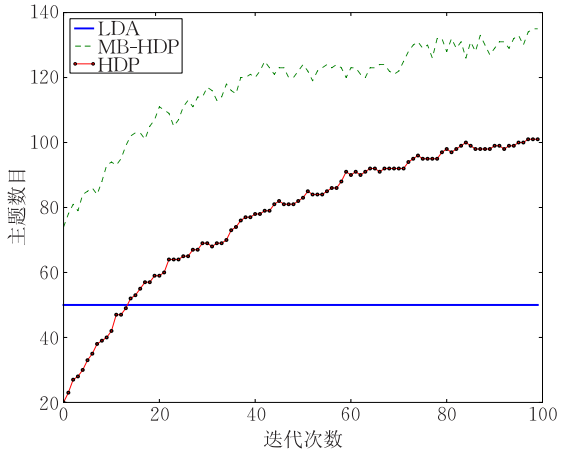

图 3　模型的主题数目

表 3 MB-HDP 模型的特有主题

主题描述	主题内容
Electric Car	care, file, ford, auto, fuel, vehicle, toyota, electric, motor, race
Energy Power	energy, power, solar, data, center, mur, wind, houston, plant, green
Google vs. Microsoft	google, microsoft, window, chrome, search, bing, bet, office, sirgold, nbc
Social Media Business	social, twitter, media, business, market, network, google, facebook, web, site
World Cup	cup, south, africa, france, ireland, soccer, cuba, italy, league, game
Lose Weight	weight, loss, health, slim, fat, bio, body, diet, video, lose
Campus Shooting Case	police, kill, say, arrest, shoot, suspect, hood, woman, charg, charge
Gay Marriage Ban	gay, marriage, ban, blog, post, council, court, church, vote, city
Hollywood Movie Star	movie, star, hollywood, jenifer, film, actor, aniston, adult, katie, ryan
Michael Jackson's Death	michael, jackson, janet, com, pop, death, music, jack, murray, simon

(3) 主题多样性. 主题的多样性评估是衡量主题模型优劣的另一个有效手段. 多样性可用主题之间的 Kullback-Leibler Divergence(KL 距离)表示:

$$KL(\phi_1,\phi_2)=\sum_{x_{ji}} p(x_{ji}\mid\phi_1)\log\frac{p(x_{ji}\mid\phi_1)}{p(x_{ji}\mid\phi_2)} \quad (18)$$

主题之间的差异性与 KL 距离成正比. 当 KL 距离为 0 时,表示两个主题完全一致. 图 4 展示了不同模型的 KL 距离,颜色越深,KL 距离越小. 显然,MB-HDP 模型发现的潜在主题差异性较大. 在对比模型中,不同主题出现相同主题词可能性更大,这种冗余现象导致 KL 距离偏小.

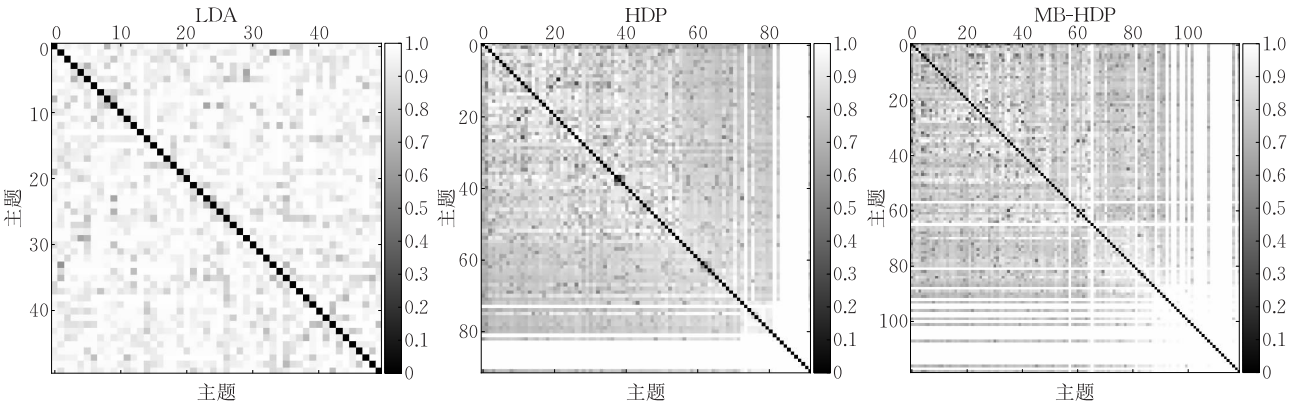

图 4 模型的主题 KL 距离

4.3 内容困惑度

内容困惑度指标被广泛应用于主题模型效果评估. 困惑度越低,表明主题模型效果越好,其计算方法如下:

$$perplexity(x)=\exp\left(-\sum_{j}^{J}\sum_{i}^{N_j}\log p(x_{ji})\Big/\sum_{j}^{J}N_j\right) \quad (19)$$

设置迭代次数为 100,各个模型的困惑度如图 5 所示. MB-HDP 模型明显优于 LDA,说明自动确定主题数目是提升挖掘效果的关键因素. HDP 比 MB-HDP 模型略差,说明考虑时间信息、用户兴趣和话题标签,在某种程度上提升了主题挖掘效果. 一旦迭代次数超过 30,各个模型的困惑度变化不明显.

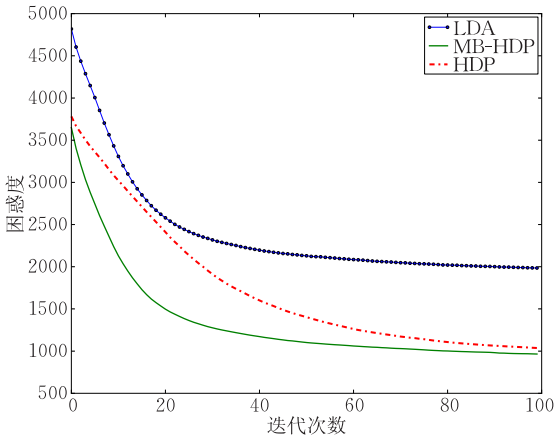

图 5 模型的内容困惑度

4.4 模型复杂度

模型复杂度是衡量非参数贝叶斯模型的另一个有效方法. 复杂度越低,说明模型用于描述数据集的主题越少. 当模型的困惑度差异不明显时,选择复杂度较小的模型. 如果模型基于 MCMC 采样方法推导后验概率,则其复杂度为主题数目与所有主题的复杂度之和. 其中,主题 ϕ_k 复杂度指的是至少存在一个单词的主题标号为 k 的微博总数. 模型复杂度计算方法参考文献[19]:

$$complexity = K + \sum_{k}\sum_{j}\mathbf{1}\left[\left(\sum_{i=1}^{N_j}\mathbf{1}[z_{ji}=k]\right)>0\right] \tag{20}$$

设置迭代次数为 100，各个模型的复杂度如图 6 所示. 两个模型的复杂度增长速率随着迭代次数递增而变化. 当迭代次数小于 30 时，MB-HDP 模型的复杂度增长速率明显高于 HDP；当迭代次数超过 30 以后，情况相反. 总的来说，MB-HDP 模型的复杂度低于 HDP. 因此，尽管 MB-HDP 模型的困惑度对比 HDP 优势不明显，但是综合考虑模型复杂度，MB-HDP 模型的总体效果更优.

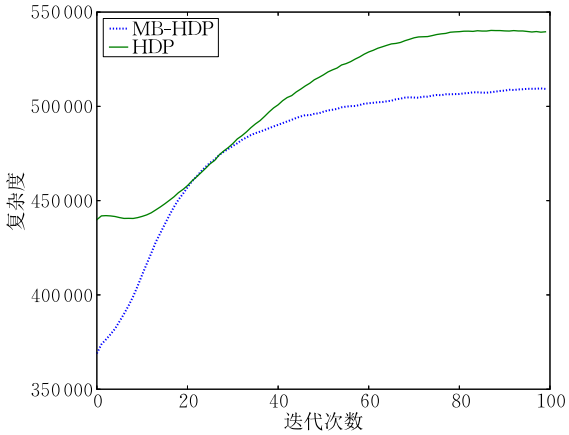

图 6　模型的复杂度

4.5　时间信息、用户兴趣和话题标签的重要性

为了说明时间信息、用户兴趣和话题标签在微博主题挖掘中的重要性，本文设计了 3 个 MB-HDP 模型的简化版本与 HDP 作对比. 简化的 MB-HDP 模型 TimeHDP、UserHDP 和 HashtagHDP 分别利用了时间信息、用户兴趣和话题标签. 设置迭代次数为 30，各个模型的困惑度和复杂度分别如图 7 和图 8 所示. 各个模型的困惑度很接近，TimeHDP 略

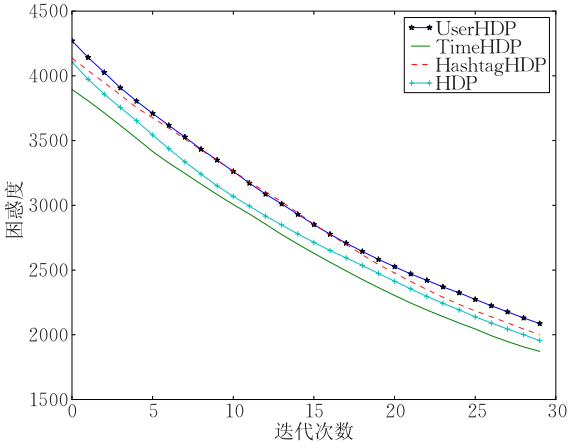

图 7　简化 MB-HDP 模型的内容困惑度

优于其他模型，说明时间信息是挖掘潜在主题的关键因素. 由于 HDP 抽取的主题包含更多的冗余信息，而简化的 MB-HDP 模型找到的主题区分度更高，因此，HDP 的复杂度在所有模型中是最差的. 综上所述，利用时间信息、用户兴趣和话题标签能够改善微博主题挖掘的效果.

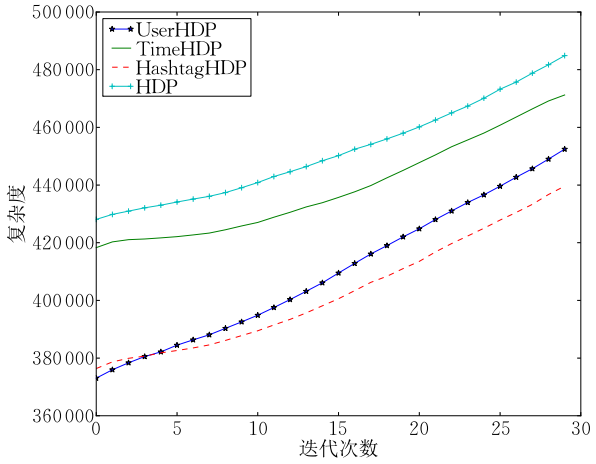

图 8　简化 MB-HDP 模型的复杂度

5　结　论

本文针对微博潜在主题的自动挖掘问题，结合时间信息、用户兴趣和话题标签，提出一个基于分层 Dirichlet 过程的非参数贝叶斯模型 MB-HDP，有效地克服了数据稀疏的困难，取得了良好的主题挖掘效果. 为了推导 MB-HDP 模型的分布参数，设计了一个基于改进 CRF 的 MCMC 采样方法. 大量真实数据集的实验表明，MB-HDP 模型明显优于 LDA 和 HDP 等现有模型.

后续的研究工作将侧重开发更加高效的 MCMC 采样算法，使得 MB-HDP 模型适合海量微博数据应用. 设计有效的方案进一步整合时间信息、用户兴趣和话题标签，以及深入分析主题之间的层次结构，也是今后的研究重点.

致　谢　感谢帮助本文写作的毕志升老师和洪佳明老师. 评审专家对本文提出了宝贵的修改意见，在此也一并表示感谢！

参　考　文　献

[1] Goorha S, Ungar L. Discovery of significant emerging trends//Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining. Washington, USA,

2010: 57-64

[2] Mathioudakis M, Koudas N. TwitterMonitor: Trend detection over the twitter stream//Proceedings of the 29th International Conference on Management of Data. Indianapolis, USA, 2010: 1155-1158

[3] Lin C X, Zhao B, Mei Q Z, Han J W. PET: A statistical model for popular events tracking in social communities//Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 929-938

[4] Budak C, Agrawal D, El Abbadi A. Structural trend analysis for online social networks. Proceedings of the VLDB Endowment, 2011, 4(10): 646-656

[5] Meng X, Wei F, Liu X, et al. Entity-centric topic-oriented opinion summarization in twitter//Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 379-387

[6] Angel A, Koudas N, Sarkas N, Srivastava D. Dense subgraph maintenance under streaming edge weight updates for real-time story identification. Proceedings of the VLDB Endowment, 2012, 5(6): 574-585

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

[8] Diao Q, Jiang J, Zhu F, Lim E. Finding bursty topics from microblogs//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea, 2012: 536-544

[9] Xu Z, Zhang Y, Wu Y, Yang Q. Modeling user posting behavior on social media//Proceedings of the 35th International Conference on Research and Development in Information Retrieval. Portland, USA, 2012: 545-554

[10] Zhang C Y, Sun J L. Large scale microblog mining using distributed MB-LDA//Proceedings of the 21st International Conference Companion on World Wide Web. Lyon, France, 2012: 1035-1042

[11] Teh Y, Jordan M, Beal M, Blei D. Hierarchical Dirichlet process. Journal of the American Statistical Association, 2006, 101(476): 1566-1581

[12] Zhang J, Song Y, Zhang C, Liu S. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora //Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1079-1088

[13] Hoffman M, Blei D, Cook P. Content-based musical similarity computation using the hierarchical Dirichlet process//Proceedings of the 9th International Society for Music Information Retrieval Conference. Utrecht, Netherlands, 2008: 349-354

[14] Li L, Fei-Fei L. Optimol: Automatic online picture collection via incremental model learning. International Journal of Computer Visio, 2010, 88(2): 147-168

[15] Emonet R, Varadarajan J, Odobez J. Extracting and locating temporal motifs in video scenes using a hierarchical non parametric Bayesian model//Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR) 2011. Colorado, USA, 2011: 3233-3240

[16] Ren L, Dunson D, Carin L. The dynamic hierarchical Dirichlet process//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 824-831

[17] 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

[18] Ahmed A, Xing E. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream//Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. California, USA, 2010: 20-29

[19] Kim D, Oh A. Accounting for Data Dependencies within a Hierarchical Dirichlet Process Mixture Model//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Scotland, UK, 2011: 873-878

[20] Meng X, Wei F, Liu X, et al. Entity-centric topic-oriented opinion summarization in twitter//Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 379-387

[21] Xu Ge, Wang Hou-Feng. The development of topic models in natural language processing. Chinese Journal of Computers, 2011, 34(8): 1423-1436(in Chinese)
(徐戈, 王厚峰. 自然语言处理中主题模型的发展. 计算机学报, 2011, 34(8): 1423-1436)

[22] Zhou Jian-Ying, Wang Fei-Yue, Zeng Da-Jun. Hierarchical Dirichlet processes and their applications: A survey. Acta Automatica Sinica, 2011, 37(4): 389-407(in Chinese)
(周建英, 王飞跃, 曾大军. 分层 Dirichlet 过程及其应用综述. 自动化学报, 2011, 37(4): 389-407)

[23] Ramage D, Dumais S, Liebling D. Characterizing microblogs with topic models//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Washington, USA, 2010: 10-17

[24] Wang Y, Agichtein E, Benzi M. TM-LDA: Efficient online modeling of latent topic transitions in social media//Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 123-131

[25] Pitman J. Combinatorial Stochastic Processes. Berlin: Springer, 2006

[26] Blei D, Frazier P. Distance dependent Chinese restaurant processes. The Journal of Machine Learning Research, 2011, 12: 2461-2488

[27] Cheng Z, Caverlee J, Lee K. You are where you tweet: A content-based approach to geo-locating twitter users//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada, 2010: 759-768

LIU Shao-Peng, born in 1984, Ph. D. His research interests include data mining and topic model.

YIN Jian, born in 1968, Ph. D., professor, Ph. D. supervisor. His research interests include database, data mining and artificial intelligence.

OUYANG Jia, born in 1986, Ph. D. His research interests include data mining and privacy preserving.

HUANG Yun, born in 1976, Ph. D. candidate, assistant professor. His research interests include data mining and graph mining.

YANG Xiao-Ying, born in 1987, M. S. His research interests include data mining and topic model.

Background

This paper studies the problem of topic mining from microblogs. Recently, Microblogging services have grown at an unprecedented rate and become a valuable source of extremely up-to-date information. Topic models can effectively explore such huge volume of information and improve the user experience by offering a list of interesting topics. The advantages of this technique are threefold. Firstly, it provides a reflection of the current main interests of the community and thus reduces the risk of information overload. Secondly, information seekers can collect valuable news easily. Finally, it is useful for government to monitor the Internet public sentiment.

In last decade, several topic models have been developed for topic mining from microblogs, including the parametric models and the nonparametric models. Although the parametric models derived from the LDA are widely used in practical applications, they require a pre-determined number of topics, which is difficult to set appropriately without any prior knowledge of the microblog corpus. The nonparametric models based on the HDP are more suitable for topic mining because they can determine the number of topics automatically. For simplicity, the nonparametric models make an exchangeability assumption that any permutation of the data would result in the same outcome. This assumption poses a problem for the microblog domain where there are clear temporal patterns of corpus and topics. Later on, several temporal-aware HDP models are devised into use. However, these models ignore other useful dependencies among tweets including user's interests and semantic #hashtags.

In this paper, we present a hierarchical Bayesian nonparametric model, MB-HDP, to automatically mining topics from microblogs. Firstly, our model assumes non-exchangeability of data which is suitable for the microblog application. Secondly, to tackle the sparsity problem caused by the short tweets, the temporal information, user's interests and semantic #hashtags are integrated to aggregate topic-related tweets into lengthy pseudo-documents. Thirdly, the Chinese Restaurant Franchise (CRF) extension is adopted in modeling topics. Finally, we present a Markov Chain Monte Carlo (MCMC) sampling for posterior inference in the MB-HDP. We demonstrate the superiority of the MB-HDP on a real dataset from three different perspectives: the quality of generated latent topics, the perplexity of held-out content and the model complexity. The results are satisfactory and the MB-HDP clearly performed better than its competitors.

This work is supported by the National Natural Science Foundation of China (Nos. 61033010, 61272065, 61472453, U1401256), Natural Science Foundation of Guangdong Province (Nos. S2011020001182, S2012010009311), Research Foundation of Science and Technology Plan Project in Guangdong Province (Nos. 2011B040200007, 2011B031700004, 2012A010701013).