·先进计算与数据处理· 文章编号:1000-3428(2018)02-0001-08 文献标志码:A 中图分类号:TP391

基于 HDP 模型的领域微博主题演化研究

高永兵[1],杨利莹[1],胡文江[1],马占飞[2]

(1. 内蒙古科技大学 信息工程学院,内蒙古 包头 014010; 2. 包头师范学院 信息工程系,内蒙古 包头 014010)

摘 要: 领域微博中包含较多的专业领域信息,并且随时间表现出较强的演化性。为分析领域的主题演化情况,构建一个基于分层 Dirichlet 过程(HDP)的 DM-HDP 模型。以用户为单位抽取领域相关的微博,利用微博的领域特征和时间特征,提取领域相关带有明显时间特征的微博并自动挖掘其主题分布,最终构建领域主题演化分析过程。实验结果表明,基于 DM-HDP 模型的分析方法能够表现领域微博主题的演化过程,与基于 LDA 和 HDP 模型的方法相比,在内容困惑度和模型复杂度等方面均具有明显优势。

关键词: 领域微博;主题挖掘;分层 Dirichlet 模型;DM-HDP 模型;Gibbs 采样;主题演化

中文引用格式: 高永兵,杨利莹,胡文江,等. 基于 HDP 模型的领域微博主题演化研究[J]. 计算机工程,2018,44(2):1-8.

英文引用格式: GAO Yongbing,YANG Liying,HU Wenjiang,et al. Research on Domanial Microblog Topic Evolution Based on HDP Model[J]. Computer Engineering,2018,44(2):1-8.

Research on Domanial Microblog Topic Evolution Based on HDP Model

GAO Yongbing[1],YANG Liying[1],HU Wenjiang[1],MA Zhanfei[2]

(1. College of Information Engineering,Inner Mongolia University of Science and Technology,Baotou,Inner Mongolia 014010,China;
2. Department of Information Engineering,Baotou Teachers' College,Baotou,Inner Mongolia 014010,China)

【Abstract】 Domanial microblog contains much professional information that show a strong evolution over time. In order to analyze the topics of professional microblog automatically, a domanial microblog topic evolution method based on Hierarchical Dirichlet Process(HDP) model is built. Firstly, domain-related microblog is extracted with the individual user as the unit. Then, accurate extraction of domain-related microblog with distinct temporal features and automatic mining of its topics using domain features and temporal features. At last, the process of domanial topics evolution analysis is constructed. Experimental results show that the method based on the DM-HDP model can show the evolution of the field of microblog, and compared with the methods that based on the LDA and HDP model, it has obvious advantages in terms of content confusion and model complexity.

【Key words】 domanial microblog; topic mining; hierarchical Dirichlet model; DM-HDP model; Gibbs sampling; topic evolution

DOI: 10.3969/j.issn.1000-3428.2018.02.001

0 概述

近年来,微博已经成为最具即时性的信息共享公共平台,不仅可以在第一时间分享社会热点、交流看法和观点,也能及时传播专业领域的发展信息。以 2016 年旅游业为例,相关博文数近 12.5 亿,旅游业用户达 9 623 万,这些微博反映出该领域的重要信息,具有较高的参考价值,以时间为轴线对领域的发展情况进行追踪具有重要意义。主题提取可以高度概括重要信息,提高阅读效率,目前,国内外学者都针对主题演化开展了相关研究。现存的主题演化研究方法可以归纳为 3 类:基于社会网络的方法,基于本体的方法和基于主题模型的方法。基于社会网络的方法[1-2]将主题表示为网络节点,主题间的演化关系用节点间的有向边表示,边的权重代表演化强度,演化强度与主题词和引文有关,但是这种基于链接的方法对于新出现的主题敏感度不高,此外由已经出现的主题可能会链接到下一个无关主题从而造成主题漂移;基于本体的方法[3-4]关注主题的语义信息和增量构建过程,借鉴词共现思想,结合本体构建主

基金项目: 国家自然科学基金(61163025);内蒙古自治区自然科学基金(2015MS0621)。

作者简介: 高永兵(1974—),男,副教授、硕士,主研方向为计算语言学、Web 数据管理;杨利莹,硕士研究生;胡文江,教授;马占飞,教授、博士。

收稿日期: 2017-02-23 **修回日期:** 2017-03-27 **E-mail:** 1689743336@qq.com

题与主题间的关系网络,然而该方法对训练集的依赖性很大,不适用于动态主题演化过程。基于主题模型的方法能够自动挖掘潜在的语义信息,且模型的灵活度高,能够根据不同的应用场景做出相应的调整,具有较强的适用性。

主题模型中应用最广泛的是由文献[5]提出的潜在 Dirichlet 分布(Latent Dirichlet Allocation,LDA)模型,以此模型为基础衍生出许多主题演化模型[6-8],解决了原始 LDA 模型忽略文本的时间信息而无法描述文本主题演化的问题。然而以 LDA 为基础的主题模型需要人为预先指定主题数目,在整个事件的演化过程中,主题的数目是不固定的,某个主题也并非贯穿于整个事件的始终。例如,"人民币加入 SDR"包含多个主题,随着时间推移,微博上讨论的主题从前期的一系列基础准备工作到如何适应市场发展,再到加入 SDR 后给中国经济带来的影响。在没有任何先验知识的前提下,很难准确把握主题数量。为克服以上问题,研究者提出了分层 Dirichlet 过程(Hierarchical Dirichlet Process,HDP)模型[9],利用狄利克雷过程(Dirichlet Process,DP)无限维度的特征实现主题数目的自动确定。

主题演化是传统主题挖掘技术的延伸与发展,指的是按照时间发展顺序对不同阶段的文本进行主题分析,要求既能概括文本信息,又能表现发展动态。对领域微博进行主题演化研究存在诸多难点。首先,传统的主题模型都是针对长文本提出的,如何调整主题模型以适应微博数据的短文本性和交互性是一大难点;此外,随着社会分工的细化,存在许多交叉行业与领域,如何准确地抽取指定行业领域的微博也是需要考虑的问题;最后,利用主题模型挖掘出领域微博的主题后,如何既能表现该领域的主题分布,又能增加对新主题的敏感性既而表达其演化过程,也是一个难点。

综合考虑以上问题,本文提出一种针对领域微博主题演化的分析方法。首先,基于用户标签和简介的领域微博提取出领域微博数据;然后,综合考虑领域特征和时间特征,构建适用于领域微博主题挖掘的 DM-HDP 模型,并设计相应的采样方法推导该模型;最后,在真实的数据集上进行实验,验证 DM-HDP 模型的有效性。

1 相关研究

关于主题演化最早的研究可以追溯到文献[10]中提出的动态主题挖掘思想,该思想将连续的文本按时间段划分,以 LDA 模型为基础对每个时段内的文本建模,综合考虑 α 和 β 参数随时间的变化,建立 LDA 模型链,最后得到随时间变化的主题分布。然而这种方法得到的主题数目是固定的,与实际情况不符。文献[11]中提出的时段性 DP 混合模型,假定每个时期内的主题数量不限定,随时间的推移可以产生新的主题,已存在的主题可以保留或者消亡。该方法解决了主题数量不确定的问题,但是运用了循环中国餐馆构造过程(Chinese Restaurant Construction Process,CRCP)采样推导,假设每个词仅属于一个主题,忽略了文本单词一词多义的特性,丢失单词部分主题信息,导致模型的精度降低。文献[12]提出基于 HDP 模型的主题演化方法,该方法构建了三层 DP 代表不同层次的主题分布,认为某时段的主题分布受上一时段的参数影响,结合折棒构造方法(Stick-breaking construction)和 CRCP 实现动态的文本主题挖掘。该方法通过建立动态主题模型系统性地挖掘文本流的主题分布,却没有对主题的演化过程进行分析,呈现的结果不直观。文献[13]中提出的一种面向多文档流的主题演化模型,允许每个文本流都有本地主题和共享主题,为每个主题的受欢迎程度建立时间变化函数,定量地分析主题变化规律。文献[14]在 LDA 模型的基础上增加 DP,不仅能获取模型的隐变量,还能完成超参数的动态更新和主题数的变动,然而该方法并没有对子主题的划分做详细介绍。文献[15]中提出的在线分层 DP 的非参数贝叶斯模型,主题的演化过程分为 2 个层次,对时间块内的文档建立在线 HDP 模型,对跨时间块的文档用时间衰减函数衡量其时间相关性,假设当前主题分布受之前几个时段主题分布的影响,就造成系统对新主题敏感度不高。

上述研究都基于长文本,对于微博这种短文本不一定适用。文献[16]提出用于挖掘微博主题的 MB-LDA 模型,综合考虑微博的联系人信息和文本信息,改进 LDA 模型以适应微博的特殊结构。文献[17]提出 MB-HDP 模型,利用微博的时间信息、用户兴趣和话题标签,聚合主题相关的信息。至此,目前还很少有关于领域微博主题演化的研究。

主题演化过程需考虑的两大因素是内容和时间,一方面要求在内容上按主题进行分类识别,另一方面又要保证时间上的延续和关联。本文首先按照用户标签提取出领域微博,然后以时间周期为界,将领域微博划分为多个独立单位,在此单位内部忽略时间信息,建立领域微博主题模型 DM-HDP,以挖掘该时段的主题分布,将主题划分到不同的大主题下,按时间顺序为相同的大主题建立关联,进而分析主题的演化过程。

2 准备工作

2.1 HDP 模型

DP 是关于分布的分布,其采样点本身就是一个随机概率分布。HDP 本质是 DP 的多层形式,可视为基于贝叶斯的传统主题模型 LDA 在无参方向的衍生[18]。以下介绍基于文档的两层 HDP 模型生成

过程。首先从基分布 H 和参数 γ 构成的 DP 中,抽样出分布 G_0;然后从基分布 G_0 和参数 α_0 构成的 DP 中,为每篇文档抽取主题分布 G_j,其中 DP 代表 DP 过程。

$$G_0 \mid \gamma, H \sim DP(\gamma, H)$$
$$G_j \mid \alpha_0, G_0 \sim DP(\alpha_0, G_0)$$
$$\theta_{ji} \mid G_j \sim G_j$$
$$W_{ji} \mid \theta_{ji} \sim Mult(\theta_{ji}) \quad (1)$$

式(1)中,θ_{ji}指示了词 W_{ji}的主题。HDP 的图模型生成过程如图 1 所示,其中圆形代表分布,圆角矩形代表参数,阴影部分表示可观测量,矩形表示该过程可循环。

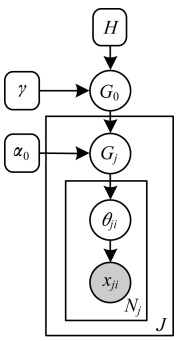

图 1　HDP 图模型生成过程

2.2　Stick-breaking 构造

以上关于 HDP 的定义并不能直接应用,可应用 2 次 Stick-breaking 方法对 HDP 的过程进行构造,详细的构造过程参见文献[19]。

1)第 1 层构造如下:

$$\beta'_k \sim Beta(1, \gamma)$$
$$\beta_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_1)$$
$$\phi_k \sim H$$
$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \quad (2)$$

其中,β_k 表示一组服从 $Beta$ 分布的随机数,ϕ_k 表示从基分布 H 中抽样的点,δ_{ϕ_k}表示抽样点的值。

从式(2)中可以看出,G_0 由一系列从基分布 H 中采样得到的离散点组成,保证了文档间的主题共享,采样点记为$\{\phi_k\}_{k=1}^{\infty}$,其权重为$\{\beta_k\}_{k=1}^{\infty}$,关于 β 的分布也可以记为 $\beta \sim GEM(\gamma)$。

2)第 2 层构造如下:

$$\pi_{jk} \sim GEM(\alpha_0)$$
$$\phi_k \sim H$$
$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \quad (3)$$

其中,π_{jk}表示取 δ_{ϕ_k}点的概率。

该构造方法并没有改变采样点,只是对采样点的权值做连续处理,HDP 模型的 Stick-breaking 构造过程如图 2 所示。

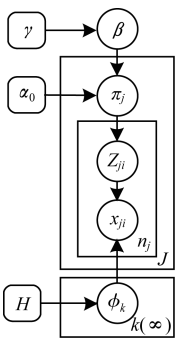

图 2　HDP 图模型构造过程

2.3　CRCP 构造

CRCP 将单词匹配主题的过程形象地比喻为顾客挑选餐桌并点菜的过程,具体如下:假设餐馆可容纳无数张餐桌,每张餐桌可容纳无数位顾客,每张餐桌上只供应一道菜。顾客进入餐馆选择餐桌并点菜,顾客可以就坐于已有餐桌也可以选择新餐桌。

1)若就坐于已有餐桌,便可以共享已点的菜肴,其概率与该餐桌上的顾客数量成正比,顾客数越多则被选中的概率越大;

2)顾客也可以以某参数概率选择新餐桌,作为该餐桌的第一位顾客应负责点菜,选择已有菜肴的概率受其被点次数的影响,菜肴被点到的次数越多则再次被点的可能性越大,顾客也可以以参数概率选择新菜肴。

3　领域微博提取

专业领域微博由长期从事具体业务的特定人群发布,带有学科性、技术性的微博数据集合一般带有明显的用户标签及简介信息。用户标签包含丰富的个性化描述信息以及用户本身的特性,能够代表用户的所属行业领域,本文利用用户标签进行领域标识。

首先构建领域关键词词典,通过抓取领域资讯网站的文档,利用现有的基于文本分类的网页爬虫技术[20]及领域模型[21]爬取领域相关信息,分析文档中的关键词,按照关键词词频统计结果,构建领域 $D = \{D_1, D_2, \cdots\}$。将领域 D_i 的专业词汇集表示为 $d_{ij} = \{d_{ij1}, d_{ij2}, \cdots\}$。

定义 1　基于用户标签的领域相似度。将用户 U_r 的标签词汇集 $d(U_r) = \{d_1, d_2, \cdots\}$ 与领域 D_i 的专业词汇集 d_{ij}进行相似度计算,若该领域的专业词汇集 d_{ij}中的第 k 个词汇与用户的标签词汇相同(不考虑用户标签词的先后顺序),定义领域相似度如下:

$$f_{\text{tag}} = \frac{\sum_{i=1}^{count} \frac{1}{k}}{Lenght_{\text{tag}}} \quad (4)$$

式(4)中分子表示所有匹配该领域的标签词位置权重的累加和,分母表示用户所有标签词汇的长

度。$\frac{1}{k}$表示相同词的位置越靠前,则值越大,实质是考虑专业词汇库的词频信息。

定义 2 领域特征指数。根据用户 U_r 的领域相似度 f_{tag},统计其所发布的总微博数 $T(U_r)$ 以及利用分类模型得到的专业领域微博数 $D(U_r)$,定义领域特征指数为:

$$F(U_r)=(1+f_{\text{tag}})\times\frac{D(U_r)}{T(U_r)} \tag{5}$$

分析用户发布的领域特征指数,如果其值大于某阈值,则认为此用户属于该领域,即表示为:

$$D_i(U)=\{U_r \mid U_r\in U, F(U_r)>thr_1\} \tag{6}$$

其中,集合 U 表示用户集合,U_r 是集合中的元素。仅当其领域特征指数条件 $F(U_r)>thr_1$ 成立时,此用户属于领域 D_i,由此得到属于该专业领域的用户集合 $D_i(U)$。

4 领域微博生成模型 DM-HDP

上一节中,得到了基于用户的领域微博数据,然而其中不可避免地掺杂了如个人生活领悟等非专业领域信息。如果直接进行主题提取,不仅会造成时间上的浪费,也会降低主题提取的精度和质量,所以,本文提出考虑微博类型的 DM-HDP 模型,对专业领域微博进行主题提取。

将属于该领域的所有用户的所有微博按时间排序,并分为 T 个时段,在此将 1 个时段设为 1 周。将用户 U_r 在第 t 时段内发布或者转发的微博表示为 $S(U_r)^t=\{\upsilon_j\}_{j=1}^{j=N}$,每条微博可以用一连串的词表示,$\upsilon_j=\{\omega_{ji}\}_{i=1}^{i=M}$,其中 M 代表此条微博的总词数,则用户 U_r 所有时段的微博可表示为 $S(U_r)=\sum_t S(U_r)^t$。DM-HDP模型中符号的具体含义详见表 1。

表 1 DM-HDP 模型中符号含义说明

符号	含义
η_x、η_y	θ_x、θ_y 关于 *Beta* 分布的参数
θ_x、θ_y	x_v、y_v 关于二项分布的参数
x_v、y_v	判断微博类型的参数
α、γ、d、κ	DP 中构造参数
H、ϕ_k	基分布和采样点
G_0、G_t、G_d、G_d^t	4 种不同类型的微博主题分布
r、ψ_t、β_d、π_{dt}	不同类型的微博采样点权重系数
z_v、w	微博中的词主题索引及词
M_k、T_i^k	不同类型微博分布的餐桌数

4.1 模型框架

在 DM-HDP 模型中,每条领域微博与 x_v 和 y_v 2 个参数相联系,代表该微博主题是否为特定时间、是否与专业领域相关,每条微博的主题分布表示为 z_v。由此将微博分为 4 类:领域特定时间主题,领域一般时间主题,公共特定时间主题以及公共一般时间主题。由 x 和 y 确定的 4 种微博类型如表 2 所示。每种微博均在不同的主题分布下,需要建立 4 种主题分布与之相对应,其中主要任务是领域特定时间主题的识别。

表 2 由 x 和 y 确定的 4 种微博类型

x 值	$y=1$	$y=0$
$x=1$	领域特定时间主题	公共特定时间主题
$x=0$	领域一般时间主题	公共一般时间主题

x_v 和 y_v 服从参数为 θ_x 和 θ_y 的二项分布,分别代表某条微博是特定时间还是任何时间均可讨论的内容,是专业领域相关的还是谈论生活等非领域相关的信息。且 θ_x 和 θ_y 服从参数为 η_x 和 η_y 的 *Beta* 分布,表示为:

$$\theta_x\sim Beta(\eta_x),\theta_y\sim Beta(\eta_y) \tag{7}$$

DM-HDP 模型要解决的关键问题是如何为 4 种分布建立主题模型,DM-HDP 模型框架图如图 3 所示。当 $x=0,y=0$ 时,有从基分布 H 中抽样得出的全局分布 G_0,G_0 代表在任何时间均可讨论的主题分布,公共一般时间主题可直接由 G_0 得到;当 $x=1$,$y=0$时,有特定时间的主题分布 G_t 与之对应,代表讨论的主题是公共特定时间主题;当 $x=0,y=1$ 时,领域相关的微博主题分布 G_d 从 G_0 中抽样得到,代表领域相关一般时间的主题分布;当 $x=1,y=1$ 时,领域特定时间主题分布 G_d^t 从 G_d 中抽样得到。

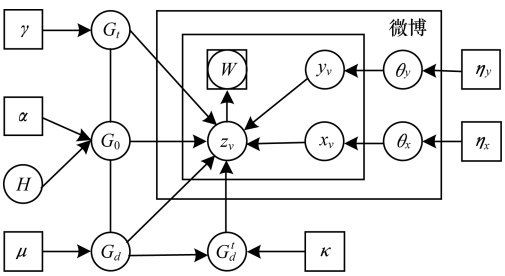

图 3 DM-HDP 模型的贝叶斯网络

由于所有的分布均从基分布 H 中得出,保证了微博间主题共享,不同的是 4 种分布 G_0、G_t、G_d 和 G_d^t 的采样点权重。图中 γ、α、μ 和 κ 代表 DP 的构造参数,DM-HDP 模型的生成过程算法描述如下:

1. Draw measure $G_0\sim DP(\alpha,H)$
2. For each time t
   -draw measure $G_t\sim DP(\gamma,G_0)$
3. For each domain parameter μ
   -draw $\theta_x\sim Beta(\eta_x)$
   -draw $\theta_y\sim Beta(\eta_y)$
   -draw measure $G_d\sim DP(\mu,G_0)$
   -for each time t:
   * draw measure $G_d^t\sim DP(\kappa,G_d)$
4. For each microblog υ_j, for domain d, at time t
   -draw $x_v\sim Multi(\theta_x)$
   $y_v\sim Multi(\theta_y)$

-if x = 0, y = 0: draw $z_v \sim G_0$
-if x = 1, y = 0: draw $z_v \sim G_t$
-if x = 0, y = 1: draw $z_v \sim G_d$
-if x = 1, y = 1: draw $z_v \sim G_d^t$
-for each word $\omega_{ji} \in \upsilon_j$
　* draw $w_{ji} \sim \text{Multi}(z_v)$

4.2　Stick-breaking 构造

根据式(3),可以将 G_0 构造如下:

$$G_0 = \sum_{k=1}^{\infty} r_k \delta_{\phi_k}, r \sim GEM(\alpha) \tag{8}$$

根据式(1),G_t、G_d 可以表示为:

$$G_t = \sum_{k=1}^{\infty} \psi_t^k \delta_{\phi_k}, \psi_t \sim DP(\gamma, r) \tag{9}$$

$$G_d = \sum_{k=1}^{\infty} \beta_d^k \delta_{\phi_k}, \beta_d \sim DP(\mu, r) \tag{10}$$

同样地,可将 G_d^t 构造为:

$$G_d^k = \sum_{k=1}^{\infty} \pi_{dt}^k \delta_{\phi_k}, \pi_{dt} \sim DP(\kappa, \beta_d) \tag{11}$$

利用 Stick-breaking 方法构造 DM-HDP 模型中的各种先验分布,并且采样点均为 $\{\phi_k\}_{k=1}^{\infty}$,保证不同分布间的主题共享,其构造的权重不同,体现了主题的差异性。

4.3　模型推理

MCMC 方法基于改进的 CRCP,根据可观测量 w 对所有的分布采样。由于篇幅有限,本文仅给出 Gibbs 采样的迭代式,详细推导可参考文献[9]。

1)r 采样:由图 2 可以看出,所有的分布均由全局分布 $G_0 = \sum_{k=1}^{\infty} r_k \delta_{\phi_k}$ 抽样得出。将所有时段中选择菜肴 k 的餐桌总数记为 $M_k = \sum_t M_k$,对于所有的菜肴,餐桌总数记为 $M_{\cdot} = \sum_k M_k$,由于 $G_0 \sim DP(\gamma, H)$,假设已知每道菜被点到的总数,即已知 $\{M_k\}_{k=1}^{K}$,则 G_0 可以表示为:

$$G_0 \mid \alpha, H, M_k \sim DP\left(\alpha + M_{\cdot}, \frac{\varepsilon H + \sum_{k=1}^{\infty} M_k \delta_{\phi_k}}{\varepsilon + M_{\cdot}}\right) \tag{12}$$

由文献[9]的 5.2 节,G_0 也可以被记为:

$$G_0 = \sum_{k=1}^{K} r_k \delta_{\phi_k} + r_u G_u \tag{13}$$

$$G_u \sim DP(\varepsilon, H) \tag{14}$$

$$r = (r_1, r_2, \cdots, r_k, r_u) \sim \text{Dir}(M_1, M_2, \cdots, M_k, \alpha) \tag{15}$$

其中,k 代表已有的菜肴数,将 r 扩展到 $k+1$ 空间,依据式(15)采样得到 r 的分布。

2)ψ_t、β_d、π_{dt} 采样:对于式(10)和式(11)中的权重系数 $\psi_t, \beta_d, \pi_{dt}$,按照 1)中对 r 的采样方法,得到 G_t、G_d 和 G_d^t 对应的主题统计量。由于 $\beta_d \sim DP(\mu, r)$,假设已知领域微博中关于主题 k 的餐桌总数,即已知 $\{T_k\}_{k=1}^{K}$,则 β_d 也可以表示为:

$$(\beta_d^1, \beta_d^2, \cdots, \beta_d^k, \beta_d^u) \sim \text{Dir}(T_d^1 + d \cdot r_1, T_d^2 + d \cdot r_2, \cdots, T_d^K + d \cdot r_K, d \cdot r_u) \tag{16}$$

同理得到 ψ_t、π_{dt}。

3)z_v 采样:给定 x_v 和 y_v 的值,可以得到相应的 z_v 值,表示如下:

$$P(z_v = k \mid x_v, y_v, w) \propto P(v \mid x_v, y_v, z_v = k, w) \times P(z = k \mid G_{x,y}) \tag{17}$$

$P(v \mid x_v, y_v, z_v = k, w)$ 代表微博 v 在主题 z 下产生的概率,$P(z = k \mid G_{x,y})$ 代表选择菜肴 z 的概率,表示如下:

$$P(z = k \mid G_{x,y}) = \begin{cases} r_k, x = 0, y = 0 \\ \psi_t^k, x = 0, y = 1 \\ \beta_d^k, x = 1, y = 0 \\ \pi_{dt}^k, x = 1, y = 1 \end{cases} \tag{18}$$

4)M_k、T_i^k 采样:在各层的餐桌总数 M_k,T_i^k 的分布中运用 CRCP,文献[9]中有详细的推导过程,在此不再赘述。

5)x 和 y 采样:

$$P(x_v = x) \mid X_{-v}, v, y) \propto \frac{E_d^{(x,y)} + \eta_x}{E_d^{(x,y)} + 2\eta_x} \times \sum_{z \in G_{x,y}} P(v \mid x_v, y_v, z_v = k) \cdot P(z = k \mid G_{x,y}) \tag{19}$$

$$P(y_v = y) \mid Y_{-v}, v, x) \propto \frac{E_d^{(x,y)} + \eta_y}{E_d^{(x,y)} + 2\eta_y} \times \sum_{z \in G_{x,y}} P(v \mid x_v, y_v, z_v = k) \cdot P(z = k \mid G_{x,y}) \tag{20}$$

其中,$E_d^{(x,y)}$ 代表用户发布的微博总量。在式(19)和式(20)中,右边第 1 部分代表用户发布不同种类的微博的概率,第 2 部分代表按照目前主题分布产生主题为 z 的微博的概率。参数 α、γ、d、κ 的设置方式与文献[9]中相同。

5　主题关联度计算与阈值确定

完成主题的挖掘后,如何从主题中选择既重要又新颖的主题词是一个关键。主题演化相比传统的静态主题挖掘要考虑时间因素,如果将静态主题挖掘视为空间中的一个点,主题演化过程就是一系列点以时间为轴连成的线,不仅要关注主题的分布情况,还要分析其变化过程。主题词的提取直接影响演化分析的准确度,要满足以下 2 个条件:1)覆盖率高,即主题词要尽可能覆盖领域微博的主要信息;2)冗余信息少,即主题词的重复信息尽可能少。因此,需要对提取出的主题做进一步的滤除。

1)子主题的相似度通常采用 KL 距离来衡量,在已知前一时段主题分布的情况下,对当前时段的主题词进行筛选,方法如下:设 t 时段中的领域微博被划分为 Kt 个子主题,子主题 $t_{m-1,n'}$ 与 $t_{m,n}$ 间的 KL 距离表示为:

$$KL(t_{m,n} \parallel t_{m-1,n'}) = \sum_{m=1}^{M} P(w_m \mid t_{m,n}) \ln \frac{P(w_m \mid t_{m,n})}{P(w_m \mid t_{m-1,n'})} \tag{21}$$

2)阈值的选取也是一个关键性问题。人工给定阈值的做法可能造成较大误差,阈值过高会导致主题太分散而无法覆盖领域的关键信息;阈值过低又会造成主题信息的冗余,不具有代表性。对此,本文依据主题平均距离,将本时段某一子主题与前一时段的所有子主题做距离计算,取其均值。

$$threshold_2 = \sum_{n'=1}^{M-1} KL(t_{m,n} \parallel t_{m-1,n'})/(M-1) \quad (22)$$

其中,$M-1$ 表示 t_{m-1} 时段的子主题总数目。

6 实验结果与分析

6.1 实验数据

以新浪微博作为数据来源,选取 2016 年 6 月 1 日—2016 年 12 月 3 日的数据进行实验。微博爬虫监测约 50 万活跃用户,其中,每条转发微博中包含如下信息:1)ID,表示该条消息的唯一 ID;2)Created-at,表示微博的时间戳;3)Text,表示微博的文本内容;4)User,表示微博的用户信息。按照定义的领域指数值筛选分类处理得到的部分数据如表 3 所示。

表 3 领域微博数据

领域	用户数	微博数
财经	15	319
医疗	17	193
教育	14	164
军事	22	159
旅游	19	198
体育	17	188
文化	12	187

由于时段划分由人为设定,可能会因为时段跨度小而导致数据量过小,且领域微博的发布数量随机,会由于某一热点消息而出现微博数量激增的情况,微博集的规模会在一定程度上影响系统性能。对此,采取下列相应措施:

1)当数据集过小时(少于 80 篇),数据较分散,得到的主题数可能高于微博条数,主题的代表性不明显。对此,将微博文本以相同的比例复制,即在不影响词语的词频及语境的情况下,提高词语的集中度,扩大词语的采样空间。

2)当数据集过大时(大于 400 篇),聚类过程中容易产生“超大类”。大类中包含的特征较多,待归类的数据容易归类到与其含有共同信息的数据中,从而导致“滚雪球”式的增长。对此,将微博数据进行二次划分后再进行主题的提取。

6.2 多领域的主题数分析

选取财经、教育、旅游和体育领域,截取 2016 年 7 月 4 日—2016 年 9 月 25 日之间 12 周的 523 条微博,设定实验的迭代次数为 200,统计每个领域主题数的分布及变化情况,实验结果如图 4 所示。

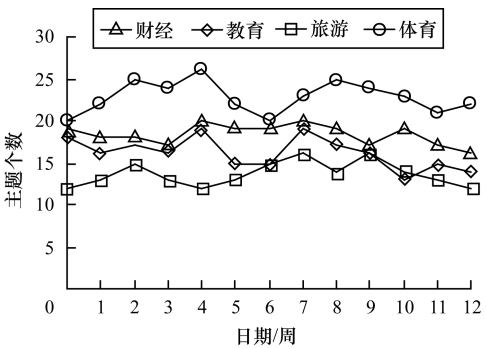

图 4 各领域主题数分布

分析发现,不同领域的主题数目在 12 ~ 26 之间变动,并且主题数的多少与微博数据集的大小没有正比例关系,决定因素在于主题的集中程度。例如,实验中第 8 周的财经相关微博为 58 条,其主题数为 19;体育相关的微博为 45 条,其主题数却为 25。

6.3 以医疗行业为例的主题演化分析

以医疗行业微博数据为对象,设定实验的迭代次数为 200,统计主题的分布情况,实验结果如表 4 所示。

表 4 医疗行业主题分布

主题编号	主题词分布
0	胆囊 胆固醇 患者 胆汁 胆囊炎 结石 形成 疾病 胆结石 女性
1	魏则西 孩子 儿子 执行 医院 竞价 法院 没有 赔偿 人民法院
2	禽流感 病毒 号贩子 行动 医托 打击 感染 医院 表示 部门
3	肿瘤 治疗 儿童 发生 中医 疫苗 结合 网友 原因 流动
4	女性 卵巢 月经 生育 年龄 避孕药 周期 正常 功能 问题
5	营养 临床 配制 医院 病人 进行 我国 混合液 合一 患者
6	大豆油 华瑞 包装 产品 质量 芦笋 成本 原材料 生产 国内
7	检查 早期 肺癌 没有 相对 CT 比较 医生 阴影 应该
8	男人 女人 问题 没有 夫妻 关系 芦笋 日子 是否 喜欢
9	创业 公司 企业 华瑞 成功 学术 员工 精神 发展 腐败

从表 4 中可以明显看出,在医疗行业下分为 10 个大主题,其中每个大主题又由 10 个子主题构成,表现出明显的层结构,有效地从微博文本中挖掘出了主题。

6.4 子主题相似性分析

为分析主题内容上的相似性,利用式(19)计算 2 个相邻时段的 *KL* 距离。选择表 4 中的前 5 个子主题,其演化结果如图 5 所示。

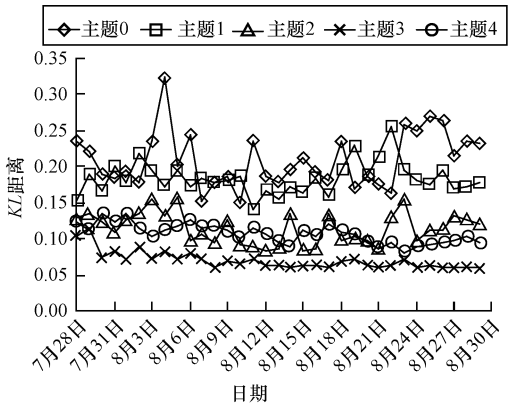

图 5　主题相似性变化图

从图 5 可以看出,主题 1 在 8 月 5 日主题间的差异较大,此时很有可能有突发事件;主题 4 和主题 5 的分布较平稳且 KL 值较低,说明该主题相似度较高,此段时间内没有突发性事件。

6.5　以魏则西事件为例的主题演化分析

以时间序列为轴线,分析医疗主题下的魏则西事件,具体演化过程分析如表 5 所示。

表 5　魏则西事件主题演化过程

时间	主题词分布
2016 年 1 月—2016 年 3 月	治疗　方案　手术　医院　患者　滑膜肉瘤　癌症　生日　疾病
2016 年 3 月—2016 年 5 月	魏则西　孩子　儿子　执行　医院　竞价　法院　没有　赔偿　人民法院
2016 年 5 月—2016 年 7 月	民营　搜索　质疑　医院　魏则西　管理　百度　体制　虚假　治疗
2016 年 7 月—2016 年 9 月	体制　监管　临床　医院　病人　监督　理性　广告　患者

演化过程的要点可简述为:2016 年 2 月,公众从知乎网站上得知魏则西的经历,讨论的内容集中在对治疗方案及病情的关注与分析上;2016 年 4 月,魏则西去世,公众关注点集中在魏则西之死事件存在的涉事医院外包诊所给民营机构、百度竞价排名等问题上;2016 年 6 月,针对网友对魏则西所选择的武警北京二院的治疗效果及其内部管理问题的质疑,相关部门立即展开调查;随后,公众的注意力放在了广告和医疗体制改革上。

6.6　对比实验

基于 LDA 模型的主题挖掘方法如图 6 所示,与 DM-HDP 方法相似,该模型以 LDA 为基础定义4 种不同的分布。参数 x 和 y 用来表征是否实时及是否领域相关,每种 x 和 y 的组合对应一组主题下的词分布。$x=0, y=0$ 对应 ϕ 分布;$x=1, y=0$ 对应 ϕ_t 分布;$x=0, y=1$ 对应 ϕ_d 分布;$x=1, y=1$ 对应 ϕ_d^t 分布。与 DM-HDP 方法的不同之处在于,对于不同的 x 和 y 值,主题索引 z 是不可共享的。

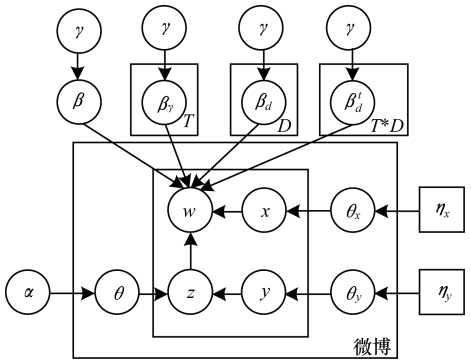

图 6　基于 LDA 模型的贝叶斯网络

6.6.1　内容困惑度

内容困惑度是广泛应用于主题模型效果评估的一项指标,其值越小,说明主题模型的效果越好。内容困惑度计算方法如下:

$$per(x)=\exp\left(-\sum_{j}^{J}\sum_{i}^{N_j}\ln(x_{ji})/\sum_{j}^{J}N_j\right) \tag{23}$$

设置迭代次数为 100,基于 LDA 模型的主题数设为 10,各模型的困惑度如图 7 所示。可以看到,在迭代次数达到 40 时,各模型的困惑度趋于平稳。DM-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 DM-HDP 模型略差,说明考虑时间信息及领域信息可以改善主题的挖掘效果。

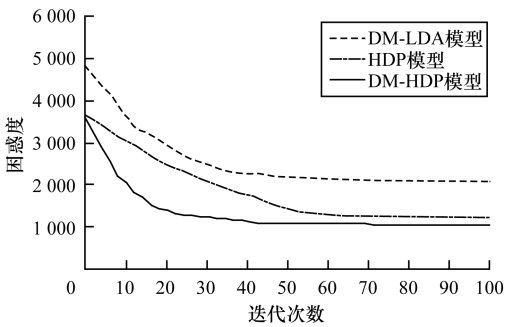

图 7　3 种模型的内容困惑度对比

6.6.2　模型复杂度

模型复杂度作为衡量模型的重要指标被广泛应用于主题模型的效果评估。复杂度越低表示模型用于描述数据集的主题越少。其他指标相同时,选择复杂度较低的模型。由于本次实验采用 MCMC 方法对模型中的各参数进行后验概率的推导,所以模型的复杂度为主题数目与所有主题的复杂度之和。模型复杂度的计算方法参考文献[22]。

$$com=K+\sum_{k}\sum_{j}\left\{\left(\sum_{i=1}^{N_j}(z_{ji}=k)\right)>0\right\} \tag{24}$$

设置迭代次数为 100,各模型的复杂度如图 8 所示。

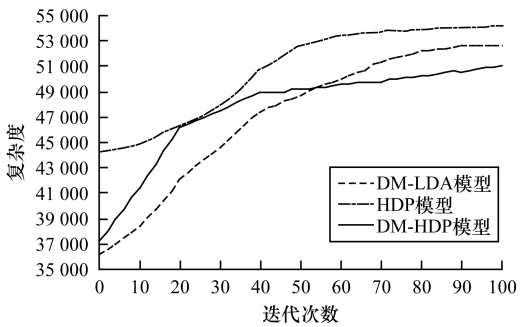

图 8 3 种模型的复杂度对比

可以看到,当迭代次数小于 30 时,DM-HDP 模型的复杂度增速明显高于其他模型,在迭代次数达到 80 时,各模型的复杂度趋于平稳,HDP 模型最高,DM-LDA 次之,DM-HDP 模型的复杂度最低。

7 结束语

本文主要研究对领域微博主题的提取方法。根据用户兴趣信息对爬取的混合微博进行领域筛选,判断其是否属于既定领域,并依据领域微博的特点,基于 HDP 模型建立领域微博主题挖掘 DM-HDP 模型,有效去除领域无关信息,使领域信息的专业性特点更加凸显。同时,为了推导 MB-HDP 模型的分布参数,应用基于改进 CRCP 的 MCMC 采样方法。对挖掘出的主题进行相关性计算,以捕捉主题的演化规律。实验数据表明,DM-HDP 模型性能优于基于 LDA 等现有模型。为了适应海量微博数据,后续研究将侧重于寻找更加高效的采样方法,以及设计有效方案进一步整合主题的层次结构。

参考文献

[1] JENSEN S,LIU X,YU Y,et al. Generation of Topic Evolution Trees from Heterogeneous Bibliographic Networks[J]. Journal of Informetrics,2016,10(2):606-621.

[2] 叶春蕾,冷伏海. 基于社会网络分析的技术主题演化方法研究[J]. 情报理论与实践,2014,37(1):126-130.

[3] 陈 千,桂志国,郭 鑫,等. 基于特征本体的文本流主题演化[J]. 计算机应用,2015,35(2):456-460.

[4] MA J,SUN M,LI C,et al. Ontology Evolution Algorithm for Topic Information Collection[J]. International Journal of Nonlinear Science,2014,18(1):86-91.

[5] BLEI D M,NG A Y,JORDAN M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research,2003,32(3):993-1022.

[6] WANG Y,AGICHTEIN E,BENZI M. TM-LDA:Efficient Online Modeling of Latent Topic Transitions in Social Media[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,USA:ACM Press,2012:123-131.

[7] 胡吉明,陈 果. 基于动态 LDA 主题模型的内容主题挖掘与演化[J]. 图书情报工作,2014,58(2):138-142.

[8] HOFFMAN M D,BLEI D M,BACH F R. Online Learning for Latent Dirichlet Allocation[J]. Advances in Neural Information Processing Systems,2010,23(5):856-864.

[9] TEH Y W,JORDAN M I,BEAL M J,et al. Hierarchical Dirichlet Processes[J]. Journal of the American Statistical Association,2006,101(476):1566-1581.

[10] BLEI D M,LAFFERTY J D. Dynamic Topic Models[C]//Proceedings of the 23rd International Conference on Machine Learning. Washington D. C.,USA:IEEE Press,2006:113-120.

[11] AHMED A,XING E P. Dynamic Non-parametric Mixture Models and the Recurrent Chinese Restaurant Process:with Applications to Evolutionary Clustering[C]//Proceedings of Siam International Conference on Data Mining. Atlanta,USA:SDM Press,2008:219-230.

[12] ZHANG J,SONG Y,ZHANG C,et al. Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,USA:ACM Press,2010:1079-1088.

[13] HONG L,DOM B,GURUMURTHY S,et al. A Time-dependent Topic Model for Multiple Text Streams[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,USA:ACM Press,2011:832-840.

[14] 方 莹,黄河燕,辛 欣,等. 面向动态主题数的话题演化分析[J]. 中文信息学报,2014,28(3):142-149.

[15] FU X,LI J,YANG K,et al. Dynamic Online HDP Model for Discovering Evolutionary Topics from Chinese Social Texts[J]. Neurocomputing,2015,171(C):412-424.

[16] 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,2011,48(10):1795-1802.

[17] 刘少鹏,印 鉴,欧阳佳,等. 基于 MB-HDP 模型的微博主题挖掘[J]. 计算机学报,2015,38(7):1408-1419.

[18] FERGUSON T S. A Bayesian Analysis of Some Nonparametric Problems[J]. Annals of Statistics,1973,1(2):209-230.

[19] 周建英,王飞跃,曾大军. 分层 Dirichlet 过程及其应用综述[J]. 自动化学报,2011,37(4):389-407.

[20] 陈晓伟. 基于主题爬虫与文本分类的微博资讯智能生成策略研究[D]. 武汉:华中科技大学,2013.

[21] 张力生,年 欢,宋 辉,等. 领域模型中关联语义的描述逻辑表示与应用[J]. 软件,2015(6):66-74.

[22] MENG X,WEI F,LIU X,et al. Entity-centric Topic-oriented Opinion Summarization in Twitter[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,USA:ACM Press,2012:379-387.

编辑 吴云芳