基于社区时空主题模型的微博社区发现方法

段 炼[1,2,3]，朱欣焰[3,4]

(1. 广西师范学院北部湾环境演变与资源利用教育部重点实验室 南宁 530001; 2. 广西师范学院资源环境科学学院 南宁 530001;
3. 武汉大学测绘遥感信息工程国家重点实验室 武汉 430079; 4. 武汉大学空天信息安全与可信计算教育部重点实验室 武汉 430079)

【摘要】提出了一种基于主题模型的微博社区发现方法。该方法采用狄利克雷过程(Dirichlet process)自适应生成多个潜在地理区域；利用多项式分布描述主题在连续时间中的强度；将用户对潜在地理区域和社区的选择偏好引入主题模型；最后通过EM方法和Gibbs采样，实现时空主题模型参数估算，以基于主题相似性进行社区发现。实验表明，该方法能更加准确地识别微博社区。

关 键 词 狄利克雷过程; 地理标识微博; 微博社区发现; 微博主题挖掘; 时空主题模型
中图分类号 P208 文献标志码 A doi:10.3969/j.issn.1001-0548.2014.03.025

Microblog Community Detection Method based on Community Spatio-Temporal Topic Model

DUAN Lian[1,2,3] and ZHU Xin-yan[3,4]

(1. Education Ministry Key Laboratory of Environment Evolution and Resources Utilization in Beibu Bay Guangxi Teachers Education University, Nanning 530001; 2. Resources and Environment Science Department, Guangxi Teachers Education University Nanning 530001;
3. State Key Laboratory of Information Engineering in Surveying, Mapping & Remote Sensing, Wuhan University Wuhan 430079;
4. Key Laboratory of Aerospace Information Security and Trusted Computing of Ministry of Education, Wuhan University Wuhan 430079)

Abstract This paper presents a novel micro-community detection method based on topic model. Multiple latent geographical regions by Dirichlet process are created adaptively. A multinomial distribution is then employed to depict topics evolutions within each time bin. User selection preferences of latent geographical region and community are introduced in topic model. Finally, the EM method and Gibbs sampling method are used to estimate spatio-temporal topic model parameters so that community detection can be realized by topics similarity. Experiment results show that this method can promote the performances of community identifying.

Key words dirichlet process; geo-tagged microblog; microblog community detection; microblog topic mining; spatio-temporal topic model

微博作为目前最具代表性的社交网络服务，逐渐成为一种重要的沟通工具和平台。由于其实时性高、流量大、内容覆盖面广，近年来，微博已经成为社会舆论、商业营销和城市功能的“传感器”。所谓社区，是指用户根据小世界特性聚集形成若干群体。发现微博中的社区，能更好地理解信息传播模式和用户交互模式群体演化规律，具有重大的学术和应用价值。微博社区主题表现为：属于该社区的用户所发表和转发微博的内容趋向于某(几)个特定的主题，如“体育”“科技”等。多个用户对某一主题的频繁讨论形成了针对该主题的用户社区。然而，大部分算法基于用户社会关系(如关注对象、好友)和微博消息转接应答(转帖，跟帖、评论)的疏密程度来发现社区，或基于聚类等模型进行网络分割以获取社区，忽略了社区的潜在主题特征。社区潜在主题表现为：属于该社区的用户所发表和转发微博的内容趋向于若干个特定的信息类别，如“体育”“科技”等，反映了用户在若干方面的兴趣倾向。而多个用户对某一主题的频繁讨论形成了针对该主题的用户社区。可见，社区主题与社区结构相互影响，特定主题的形成反映了某个社区的出现，而社区的出现促使了某些主题更加突出。在引入微博主题进行社区发现的研究中，文献[1]利用LDA[2]分析Twitter中的用户同质性，挖掘活跃的微博用户群组，文献[3]基于用户间的互访类型和微博主题相似性进行社区发现，文献[4]利用主题模型获取社区主题，

收稿日期：2013 – 10 – 29; 修回日期：2014 – 03 – 18
基金项目：国家863计划(2013AA12A203)；国家自然科学基金(41361022)
作者简介：段炼 (1981 –)，男，博士生，主要从事时空数据挖掘方面的研究.

计算出用户隶属于某个社区的概率，文献[5]通过伯努利分布表达用户主题在时间上的分布。

此外，用户发送的微博主题与周边地理环境特征紧密相连[6]。如在东湖周边人们发的微博以“游览”主题为主，在武汉广场发的微博则以“购物”主题为主。可见，局部地理区域具有特定的经济文化，环境对微博主题具有较大影响。近年来，微博对地理位置标识的功能为研究时空环境与微博内容之间的关系提供了支撑。一般来说，两个用户访问相同地理区域的次数越多，这些局部地理区域社会环境对他们的吸引程度就越相似，表明他们社会生活模式或兴趣偏好越相似，则这两个用户越有可能属于同一社区；另一方面，属于同一社区的用户，由于他们具有相似的生活模式或兴趣偏好，则他们越倾向于访问相同的地理区域。此外，由于社会生活的作息规律、社会习俗等原因，不同时间段内用户关注的对象是不同的，造成微博在不同时间下表达不同的主题，因此发现微博主题随时间变化的特点，同样能提高对社区的识别能力。

由于主题模型对文本的强大建模能力和灵活的扩展机制，本文将扩展主题模型引入时空要素提高微博主题识别能力。在时空相关的主题模型中，已有研究将全局空间区域划分为若干地理区域，再依据落在地理区域内的微博获取该区域的主题。地理区域的划分方式主要有如下4种：规则格网[7]、辖区(如省界、区界等)[8-9]、不规则格网(如泰森多边形网)和自适应区域划分[6]。前3种方法固化了区域边界，不利于描述相似微博主题在空间上的转移；最后一种方式依据微博主题相似性和空间邻近性，利用二维高斯分布较好的表达了相似主题微博的空间覆盖范围。然而，已有自适应区域划分方法没有给出潜在地理区域空间范围的限制条件，易造成某些潜在地理区域覆盖的空间范围过大。如文献[7]采用二维高斯模型表达相似博客主题的潜在地理区域中，一些潜在地理区域跨越大半个美国，部分潜在地理区域间还相互重叠。这种情况造成区域内主题分布趋于背景主题分布，无法突出区域“特色”主题分布，失去在主题模型中引入空间要素的意义。同时，已有方法还需预先设定潜在地理区域数量，无法利用数据自身的特征自适应调整潜在地理区域范围和数量。此外，已有方法没有顾及用户对地理区域的选择偏好。

为克服以上问题，本文构建社区时空主题模型(community spatio-temporal topic model, CS-TM)，在主题模型中引入狄利克雷过程混合模型(dirichlet process mixture model)[11]，以自动生成不同覆盖范围地理区域和微博地理位置，并通过地理区域和社区两者微博主题的相互影响，提高微博社区的发现能力。

1　微博社区挖掘

1.1　微博要素构成

每条微博d表示为6个要素：$d=(W,t,l,r,u,c)$。其中，W表示该微博“词袋”模型；t表示微博发布时间；l表示微博发布的地理位置；r表示微博潜在地理区域；u表示微博用户；c代表用户所在社区。

1.2　空间和用户对微博主题的制约

微博中常常体现如“娱乐”“交通”“饮食”等话题，这些话题表达了大众用户的一种基本社会见识，称为背景主题，其多项式分布参数用θ_0表示。将全局空间划分为多个潜在地理区域。由聚集在潜在地理区域的微博主题共同产生区域-主题分布参数为θ_r。

狄利克雷过程混合模型是具有狄利克雷过程先验假设的非参数混合概率模型，能同时实现聚类个数和分布参数估计。在潜在地理区域划分中，引入狄利克雷过程混合模型，依据用户偏好进行潜在地理区域和微博位置生成过程：首先，从均匀分布获取基础分布G_0，设定聚集参数α_r；再通过Dirichlet process—$\mathrm{DP}(\alpha_r, G_0)$获得区域多项式分布$G$，基于用户$u$和$G$从多项式分布$G(\theta_r,u)$生成区域$r$；最后基于二维高斯分布生成微博位置$l_i$：

$$P(l_i \mid r) \sim N(\mu_r, \Sigma_0) \tag{1}$$

$$P(r_i) \sim G(\theta_r,u) = G(\theta_r,\eta_u) \tag{2}$$

$$G \sim \mathrm{GP}(\alpha_r, G_0(\theta_r)) \tag{3}$$

$$G_0 = \mathrm{Uniform}(\,) \tag{4}$$

式中，η_u为用户u对潜在地理区域的选择概率；μ_r, Σ_r分别表示二维高斯分布均值和方差。G每次依据式(3)采样，因而其每次获取的参数个数可能不同，而每个参数代表了一个潜在地理区域的生成概率，从而使G具有自适应表达潜在地理区域数量的能力。

1.3　社区和时间对微博主题的制约

采用多项式分布表达用户属于各社区的概率：

$$P(c \mid u) = \mathrm{Multi}(c \mid \gamma_u) \tag{5}$$

其中，γ_u表示参数。基于稀疏增量式生成模型[12]，微博的主题z即可由背景主题、社区和区域的多项式分布参数获取：

$$P(z \mid c,r) = P(z \mid \theta_0,\theta_r,\theta_c) = \mathrm{Multi}(z \mid \theta_0+\theta_r+\theta_c) \tag{6}$$

由于普遍的社会行为规律，在不同时段，用户

具有不同的行为和意图。如08:00～11:00时的微博主题主要与“工作”“交通”相关，而12:00时和18:00～20:00时左右微博主题主要与“餐饮”“娱乐”有关。可见，微博主题会随着时间推移而产生变化。本文中，每个主题对应一个时间多项式分布 ξ_z，以表达每个主题下每个时间戳出现的概率：

$$P(t \mid z) = \text{Multi}(t \mid \xi_z) \tag{7}$$

1.4 空间、社区对微博词汇的制约

相同主题下用户在不同位置会使用不同的词汇。如同样是“交通”主题，在飞机场发出的词汇和在火车站发出的词汇就不同。因此，微博词汇受背景环境、潜在地理区域和社区影响，共同控制主题z下的词汇生成概率，基于稀疏增量式生成模型可得词汇w的多项式分布：

$$P(w \mid r,c,z) = \text{Multi}(w \mid \varphi_0 + \varphi_{r,z} + \varphi_{c,z}) \tag{8}$$

式中，φ_0、$\varphi_{r,z}$、$\varphi_{c,z}$分别表示主题为z时的背景词汇分布参数、潜在地理区域词汇分布参数和社区词汇分布参数

1.5 时空主题模型

下面给出微博文档的生成过程(如图1所示)：

1) 从作为超分布的Laplace分布 $\text{L}(0,O_u)$、$\text{L}(0,\rho_u)$和 $\text{Dirichlet}(\alpha_0)$、$\text{Dirichlet}(\psi)$ (相关超参数不在图1中表示)中分别抽样得到用户–潜在地理区域分布参数 η_u、用户–社区分布参数 γ_u、背景主题分布参数 θ_0、主题–时间分布参数 ξ_z。

2) 对于每个潜在地理区域 $r = 1,2,\cdots,R$ 有：

①基于潜在地理区域–主题分布参数 θ_r 和用户–潜在地理区域分布参数 η_u 构造多项式分布，并从中抽取潜在地理区域 r。

②从潜在地理区域 r 中采样获得微博地理位置 l_d。

3) 对于每个社区 $c = 1,2,\cdots,C$，基于用户–社区分布参数 γ_u 构造多项式分布，并从中抽取社区 c。

4) 对于每个主题 $z = 1,2,\cdots,K$，有：

①从超分布 $\text{Dirichlet}(\alpha_r)$、$\text{Dirichlet}(\alpha_c)$ 中分别获取潜在地理区域–主题分布参数 θ_r、社区–主题分布参数 θ_c。

②由 θ_0、θ_r、θ_c 构造主题多项式分布，并从中抽取出主题 z。

③从主题–时间多项式分布中采样获得时间 t。

④从 $\text{Dirichlet}(\beta_0)$ 中得到背景中的主题–词汇分布参数 φ_0，从 $\text{Dirichlet}(\beta_r)$ 中得到潜在地理区域 r 中的主题–词汇分布参数 φ_r，从 $\text{Dirichlet}(\beta_c)$ 中得到社区 c 中的主题–词汇分布参数 φ_c。

5) 对于微博 d 中的每个词汇 $w = 1,2,\cdots,N_d$，从 φ_0、φ_r、φ_c 为参数的主题–词汇多项式分布中抽取词汇 w。

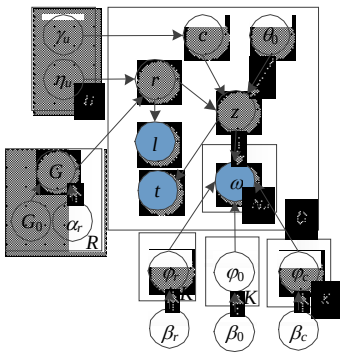

图1 社区时空主题模型

1.6 时空主题模型参数计算

基于EM方法和Gibbs采样[10]估计时空主题模型的参数。

1) E步骤，对潜在变量采样。这里的潜在变量分别是微博d所在的潜在地理区域r、社区c和主题z：

$$r_d \sim c_d z_d P(l_d \mid \mu_r, \Sigma_r) P(r \mid \eta_0, \eta_u) \tag{9}$$

$$c_d \sim r_d z_d P(c \mid \gamma_u) \tag{10}$$

$$z_d \sim P(z \mid \theta_0, \theta_r, \theta_c) P(t \mid \xi_z) \prod_{v \in N_d} P(v \mid \varphi_{r,z}, \varphi_{c,z}, \varphi_{0,z}) \tag{11}$$

通过狄利克雷过程表示微博d位于某个已存在潜在地理区域r_j或新潜在地理区域r'的概率。因此，修改式(9)，采用“Chinese restaurant”[11]的方式进行r_d采样：

$$r_d = \begin{cases} r^j \sim \dfrac{r_d^j}{\sum_{i \in R} r_d^i + G_0} \\ r' \sim \dfrac{G_0}{\sum_{i \in R} r_d^i + G_0} \end{cases} \tag{12}$$

$$r_d^i \sim c_d z_d P(l_d \mid \mu_i, \Sigma_i) P(r \mid \eta_0, \eta_u) \frac{1}{\ln(\Sigma_i) + 1} \tag{13}$$

式中，r^j 为潜在地理区域j；r_d^j 为微博d属于区域j的概率；G_0可看做为生成新潜在地理区域r'的概率。当微博d位于所有已存在潜在地理区域中的概率都比较低时，模型会倾向生成一个新的潜在地理区域r'。在r'的二维高斯分布中，均值为该微博的地理位置，方差为默认初始值 Σ_0。式(13)从潜在地理区域–主题分布、社区–主题分布、微博的地理位置和潜在地理区域空间范围等多个方面控制了微博与潜在地理区域的关系，较以往方法相比，其减少了潜在地理区域在空间上的相互重叠或嵌套，限制了巨型潜在地理区域的出现，进而提高了各潜在地理区域主题分

布的差异性，提高了微博主题获取准确性。

2) M步骤，固定各潜在因素，对模型的后验似然值最大化，获取模型参数。

更新代表潜在地理区域r的二维高斯分布参数：

$$\begin{cases}\mu_r=\dfrac{\sum\limits_{d\in r}l_d}{\#(r,d)}\\ \Sigma_r=\dfrac{\sum\limits_{d\in R}(l_d-\mu_d)^{\mathrm{T}}(l_d-\mu_d)}{\#(r,d)-1}\end{cases}\tag{14}$$

式中，$\#(r,d)$表示r中微博总数；l_d为微博地理坐标。

以下采用梯度下降法进行其他参数的迭代推理。

下式用于计算潜在地理区域分布参数的梯度值。其中，$d_{u,r}$表示用户u在区域r中所发微博数量；d_u表示用户u所发微博总数：

$$\partial\eta_{u,r}=d_{u,r}-d_uP(r\mid\eta_0,\eta_u)\tag{15}$$

$$\partial\eta_{0,r}=\sum_{u\in U}d_{u,r}-\sum_{u\in U}d_uP(r\mid\eta_0,\eta_u)\tag{16}$$

计算社区分布参数的梯度值：

$$\partial\gamma_{u,\mathrm{c}}=d_{u,\mathrm{c}}-d_u\times P(c\mid\gamma_u)\tag{17}$$

式中，$d_{u,c}$为用户u在社区c中所发微博数量。

计算主题分布参数的梯度值：

$$\partial\theta_{r,z}=\sum_{u\in U}d_{u,r,z}-\sum_{u\in U}d_{u,r}P(z\mid\theta_0,\theta_r,\theta_c)\tag{18}$$

$$\partial\theta_{\mathrm{c},z}=\sum_{u\in U}d_{u,\mathrm{c},z}-\sum_{u\in U}d_{u,c}P(z\mid\theta_0,\theta_r,\theta_c)\tag{19}$$

$$\partial\theta_{0,z}=\sum_{u\in U}d_{u,z}-\sum_{u\in U}\sum_{r\in R}d_{u,r}P(z\mid\theta_0,\theta_r,\theta_c)\tag{20}$$

式中，$d_{u,r,z}$为用户u在潜在地理区域r发表的主题为z的微博数量；$d_{u,c,z}$为用户u在社区c发表主题为z的微博数量；$d_{u,z}$为用户发表主题为z的微博数量。

计算词汇分布参数的梯度值：

$$\partial\varphi_{r,z,v}=\sum_{u\in U}n_{u,r,z,v}-\sum_{u\in U}n_{u,r,z}P(v\mid\varphi_r,\varphi_c,\varphi_0)\tag{21}$$

$$\partial\varphi_{\mathrm{c},z,v}=\sum_{u\in U}n_{u,\mathrm{c},z,v}-\sum_{u\in U}n_{u,\mathrm{c},z}P(v\mid\varphi_r,\varphi_c,\varphi_0)\tag{22}$$

$$\partial\varphi_{0,v}=n_v-\sum_{z\in Z}n_zP(v\mid\varphi_r,\varphi_c,\varphi_0)\tag{23}$$

式中，$n_{u,r,z,v}$为用户u在区域r中词汇v属于主题z的次数；$n_{u,r,z}$为用户在区域r中属于主题z的词汇次数；$n_{u,c,z,v}$为用户u在社区c中词汇v属于主题z的次数；$n_{u,c,z}$为用户u在社区c中属于主题z的词汇数量；n_v为词汇v出现次数，n_z为属于主题z的词汇数量。

计算时间采样概率分布参数的梯度值：

$$\partial\xi_{z,t}=d_{z,t}-\sum_{t\in T}d_{z,t}\times P(t\mid\xi)\tag{24}$$

式中，$d_{z,t}$为主题为z且在时间t发出的微博数量。

综上所述，在E步骤中，结合狄利克雷过程和模型中各参数，计算出微博主题、潜在地理区域和社区潜在因子；在M步骤中，通过梯度下降方法，得到模型中各个参数更新值。如此反复，直到模型各参数收敛。

最终按照用户u属于各社区的概率γ_u值的高低，即可将用户划分到不同社区中。社区间可相互重叠，即一个用户可属于多个社区，因此，取γ_u的top-k(一般k=3)个概率最大$\gamma_{u,c}$所对应的社区c作为候选集合，假设共有$|C|$个社区，设定阈值$\varepsilon=2/|C|$，将$\{\gamma_{u,c}\mid c\in k,\gamma_{u,c}>\varepsilon\}$的社区$c$作为用户$u$隶属的社区。

2 实　验

2.1 数据预处理和模型参数设置

实验使用的服务器配置为Intel(四核，3.1G)酷睿i53450，8 GB内存，装载Windows Server 2008操作系统。利用新浪微博API，基于用户好友和关注对象爬取微博后，经过去噪处理得到80 492条微博和9 264个用户。主题模型的超参数α_0、α_c、α_r统一设为50/Z，β_0、β_c、β_r统一设为0.005，O_u、ρ_u均设定为0.5。

2.2 社区主题

设微博数据集的主题数量$|Z|$为60，社区数量$|C|$为20，基于同一社区内各用户主题和词汇获得社区的主题-词汇分布。选择其中5个社区及与其相关度最高的前10个词汇，每个社区下词汇出现的概率列在该词汇右边，如表1所示，表中可明显发现，同一社区的词汇具有显著的语义相似性，不同社区的词汇含义相差显著，如社区1的词汇主要表达科技和体育主题，社区2主要表达社会和工作主题，社区3主要表达购物、饮食主题，社区4主要表达生活、娱乐等主题，社区5主要表达工作、学习等。

表1　5个社区及与其相关度最高的10个词汇

社区1 科技体育	社区2 社会政治	社区3 购物饮食	社区4 生活娱乐	社区5 工作学习
IT0.009 5	变化0.007 2	特色0.007 6	风景0.001	学校0.009 0
平台0.009 0	日本0.006 9	理想0.007 1	表情0.009 2	收入0.008 1
苹果0.008 4	中国0.006 4	消费0.007 0	累0.009 0	交通0.007 4
手机0.005 9	物价0.006 1	线路0.006 1	饭店0.009	堵车0.006 4
比分0.046	汽油0.006 0	馆子0.005 1	天气0.006 8	学习0.005 8
信息0.029	社会0.005 6	吃饭0.004 9	汽车0.003 6	社会0.004 2
现代0.002 2	改革0.005 4	商店0.004 5	房价0.002 9	超市0.002 8
NBA 0.001 6	投资0.005 1	物价0.003 2	热0.001 7	管理0.002 5
网络0.001 1	收入0.003 5	小吃0.003 0	美女0.001 2	节目0.002 1
季度0.000 7	危机0.000 1	交通0.001 1	美食0.000 3	休假0.001 1

2.3 实验结果和分析

本文提出的时空主题模型(CS-TM)与DCTM[1]和LDA[2]进行比较，其中，DCTM与CS-TM的社区确定方式相同；基于LDA的社区发现是通过主题分布进行*k*-means聚类。每个聚类簇即为社区。

本文利用社区内外链接比[5]反映社区−内用户交互程度，采用社区内用户间主题分布的KL距离(kullback-leibler)来衡量社区用户主题的相似性。社区内外链接比为：

$$f(c)=\frac{2I_c}{2I_c+O_c} \tag{25}$$

式中，I_c为社区c内用户间的链接数量，链接数量通过用户的好友、关注和跟帖数量获取；O_c为社区c内用户与社区c外用户的链接数量。$f(c)$越大，表明社区内用户较社区外用户的联系越密切；反之，表明社区内用户的联系并不紧密。图2显示了3种方法在不同社区数量情况下的平均社区内外链接比。随着社区数量的增加，各模型获取的平均社区内外链接比上升，社区涉及的微博内容范围逐渐缩小，用户间的联系频率提高。

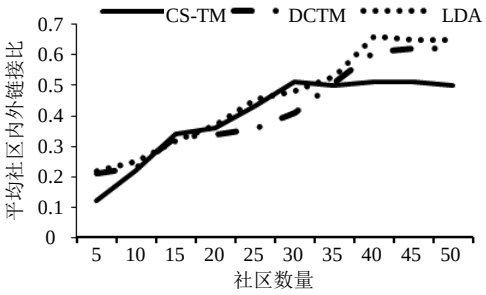

图2 3种方法的社区内外链接比

然而，在社区数量为5～35时，3种方法的平均社区内外链接比类似，但DCTM稍低；在社区数量超过35后，CS-TM的平均社区链接度较LDA和DCTM都要低。这表明本文方法获取的社区内用户的联系程度不如LDA方法和DCTM方法所获取的紧密。分析发现，CS-TM所获取的社区中包含的主题分布聚焦性强，即社区中的用户所发微博大都集中在若干特定主题，相对而言，LDA和DCTM划分出的社区，微博涉及的主题类型比较分散。这反映了主题聚焦的用户间，总体上相互交流较少；反之，一个人所发微博的主题类型越多，其对外交流越频繁。在社区数量超过40后，各方法得到的平均社区内外链接比趋于稳定。以上实验体现了微博中专业人员(或兴趣极少的用户)之间的交流不够紧密，而具有大众性和社会性主题的微博在用户间传播广泛。本文方法能发现那些兴趣类型少但兴趣类似的用户，如果应用于“用户推荐”，则能精准提高这一类用户之间的交流程度。

KL距离用以衡量相同事件空间里的两个概率分布的差异情况。KL距离越小，表明社区内用户主题相似度越高，反之，社区内用户的兴趣差异越大，则社区形成的可能性越低。由于，KL距离不具有对称性，因此基于KL距离的对称平滑版本——Jensen-Shannon(JS)距离表达社区c内用户间的平均KL距离：

$$\begin{cases}\mathrm{aKL}(c)=\dfrac{\sum\limits_{u,u'\in cm,u\neq u'}\mathrm{JS}(u,u')}{|c|(|c|-1)}\\ \mathrm{JS}(u,u')=\dfrac{\mathrm{KL}(u\,||\,e)+\mathrm{KL}(u'\,||\,e)}{2}\qquad e=\dfrac{z_u+z_{u'}}{2}\end{cases} \tag{26}$$

式中，u为用户，每个用户的主题由其所发微博主题表示；$|c|$表示社区c内的用户数量。3种方法得到的平均社区KL距离如图3所示。随着社区数量的增加，社区内用户兴趣相似性逐渐增强，3种方法的社区平均KL距离逐渐降低，但同样由于CS-TM引入的时空信息增强了微博主题获取的正确性，提高了社区内主题相似程度。LDA中，每个用户仅属于同一个社区，但由于LDA对微博主题获取准确率较低，类似语义的词汇被划分到不同主题中，造成处于同一社区的主题较为分散，同一社区内的用户主题差异性较大。DCTM对微博主题的识别能力较LDA要高。因此，其构建的社区中，用户间的主题较LDA更为相似。随着社区数量的增加，3种方法得到的社区平均KL距离的区域平稳。

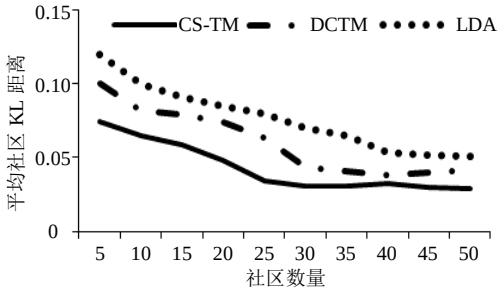

图3 3种方法的社区内平均KL距离

可见，由CS-TM模型生成的社区，其内部用户主题相似性高，社区间区分度良好。

3 结 语

本文将地理区域、社区和连续时态要素引入主题模型，综合考虑了用户对社区及潜在地理区域的偏好，利用Dirichlet process mixture model自适应划分潜在地理区域，弥补了以往方法中单个潜在地理

区域范围过大和主题分散的不足；同时，揭示了社区与具有特定社会功能地理区域间的互动关系。最终通过实验验证了该方法对社区发现的有效性。今后将引入用户间的链接信息，进一步提高微博主题提取和微博社区发现的性能。

参 考 文 献

[1] WENG Jian-shu, LIM E P, JIANG Jing, et al. Twitter rank: finding topic-sensitive in fluential twitterers[C]//Proc of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM, 2010.

[2] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[3] 闫光辉, 舒昕, 马志程, 等. 基于主题和链接分析的微博社区发现算法[J]. 计算机应用研究, 2013, 30(7): 1953-1957.
YAN Guang-hui, SHU Xin, MA Zhi-cheng, et al. Community discovery for microblog based on topic and link analysis[J]. Application Research of Computers, 2013, 30(7): 1953-1957.

[4] YIN Zhi-jun, CAO Liang-liang, GU Quan-quan, et al. Latent community topic analysis: integration of community discovery with topic modeling[J]. ACM Transactions on Intelligent Systems and Technology, 2012, 3(4): 63-84.

[5] LI Dai-feng, DING Ying, SHUAI Xin, et al. Adding community and dynamic to topic models[J]. Journal of Informetrics, 2012, 6(2): 237-253.

[6] YIN Zhi-jun, CAO Liang-liang, HAN Jia-wei, et al. Geographical topic discovery and comparison[C]//The 20th international conference on World Wide Web(WWW). New York, USA: [s.n.], 2011.

[7] EISENSTEIN J, O'Connor B, SMITH N A, et al. A latent variable model for geographic lexical variation[C]//The 20th Conference on Empirical Methods in Natural Language Processing. MIT, Massachusetts, USA: Association for Computational Linguistics, 2010.

[8] SIZOV S. GeoFolk. Latent spatial semantics in web 2.0 social media[C]//The 3rd International Conference on Web Search and Data Mining(WSDM). New York, USA: ACM, 2010.

[9] MEI Qiao-zhu, Liu Chao, SU Hang. A probabilistic approach to spatiotemporal theme pattern mining on weblogs[C]//The 15th international conference on World Wide Web(WWW). Edinburgh, Scotland: ACM, 2006.

[10] GRIFFITHS T L, STEYVERS M. Finding scientific topics[C]//Proceedings of the National Academy of Sciences (NAS), USA: [s.n.], 2004.

[11] BLEI D M, GRIFFITHS T L, JORDAN M I. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies[J]. Journal of the ACM, 2010, 57(2): 111-142.

[12] EISENSTEIN J, AHMED A, XING E P. Sparse additive generative models of text[C]//The 28th International Conference on Machine Learning(ICML). New York, USA: ACM, 2011.

编 辑 叶 芳