DOI:10.13203/j.whugis20120604 文章编号:1671-8860(2014)02-0210-04

基于时空主题模型的微博主题提取

段 炼[1,2,3] 吴 维[1] 朱欣焰[4] 胡宝清[2,3]

1 武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079
2 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,广西 南宁,530001
3 广西师范学院资源环境科学学院,广西 南宁,530001
4 武汉大学空天信息安全与可信计算教育部重点实验室,湖北 武汉,430079

摘 要:已有地理主题模型没有考虑不同区域对微博主题影响程度的差异性,同时他们将时间要素离散化,难以得到连续时间上的微博主题强度。提出了一种顾及连续时间及区域影响力因素的时空主题模型。该方法将城市划分为多个区域,依据各兴趣点类型及数量对区域赋予权重以表达区域社会功能对微博主题的影响程度,基于稀疏增量式生成模型表达微博主题分布,利用Beta分布描述主题在连续时间中的强度,最终通过Gibbs采样得到时空主题模型各参数。实验表明,本文方法能发现连续时间上微博主题的演变,与已有地理主题模型相比,能更加准确地提取微博主题。

关键词:地理主题模型;微博主题挖掘;时空分布;时空推理
中图法分类号:P208 文献标志码:A

提取带有地理标识的微博主题,用以获取人们社会移动模式、热点事件时空演变和城市区域功能变化等知识,能为商业服务推荐、舆情监控、灾害预警管理等提供有力支撑。

近年来,LDA(latent dirichlet allocation)[1]以其出色的降维能力成为文本挖掘领域的一个热门研究方向。但由于微博字数有限、噪音大等原因,直接利用LDA提取微博主题的效果并不理想[2]。然而,用户发送的微博主题与周边地理环境特征及其在时间上的行为规律紧密相连[3],例如,白天在黄鹤楼周边人们发的微博多以"游览"主题为主,而晚上则可能以"饮食"主题为主。因此,在主题模型中引入地理或时态因素可有效地提高微博主题获取的准确性。

当前,根据地理区域表达方式的不同,地理主题模型可以分为以下两类:① 将全局空间划分为若干边界固定的区域[4-5],再利用多项式分布表达每个区域的主题分布;② 利用二维高斯分布[6]或经纬度上的双重高斯分布[7]来表达主题的全局空间分布,同时,文献[8]还引入了用户对区域的选择偏好因素。为表达主题与时间的关系,文献[9]采用泊松分布来捕捉主题的时间强度,但无法在主题模型中直接引入时间因素来推断主题;文献[4]在主题模型中引入了离散时间因素,但难以表达连续时间上的微博主题分布。

可见,已有地理主题模型没有考虑不同区域对微博主题影响程度的差异性,同时他们将时间要素离散化,难以得到连续时间上的微博主题强度。因此,本文提出了一种基于地理区域和连续时间要素的时空主题模型,以提高城市内微博主题获取的准确性。

1 时空主题模型

1.1 区域划分和时间划分

城市中的社区或街区,由于其所包含各兴趣点的类型和数量,会形成独特的局部"社会功能主题分布",对区域内的微博主题产生影响。本文利用武汉市各社区的地理中心位置,构造了覆盖武汉市区的泰森多边形网,其中每个多边形网格代表一个区域,以此表达微博所受区域的主题分布影响。

收稿日期:2013-05-15
项目来源:国家863计划资助项目(2013AA12A203,2011AA010502);国家科技支撑计划资助项目(2012BAH35B03);广西北部湾重大基础研究专项基金资助项目(2011GXNSFE018003,2012GXNSFEA053001)。
第一作者:段炼,博士生,讲师,现从事智能空间信息服务研究。E-mail:wtusm@163.com
通讯作者:吴维,博士,讲师。E-mail:geosta@163.com

在时间划分上,本文利用Beta 分布[10] 来描述不同主题随时间连续变化过程,避免了时间离散化导致的时间槽跨度大小的选择问题,能描述和获取连续时刻主题的强度。

1.2　区域对微博的影响

不同区域对其范围内的微博主题影响程度是不同的。某些区域的社会功能影响力突出(如繁华商业区),来到这里的用户所发的微博主题大都与该区域社会功能相关(如购物),某些区域的社会功能影响力较弱,该区域的用户所关注的主题与区域社会功能关联不大(如住宅区)。有效判别区域主题对微博的影响程度,能显著提高微博主题获取的准确性。本文认为,一个区域内的被标注出的兴趣点数量越多,区域面积越小,兴趣点的类型越集中,表明该区域的社会功能越显著,该类型区域内的微博主题受其的影响就越强烈;反之,区域社会功能对微博主题的影响就越弱,这时,微博主题的分布就越倾向区域背景主题分布。由于熵能表达信息的确定程度,我们采用熵表达区域兴趣点类型的集中度;由于Sigmoid 函数连续且严格单调递增,能平滑地将实数控制在0 和1 之间,因此,利用该函数描述区域主题影响力权重,0 代表无影响,1 代表完全受控:

$$\sigma_r = \frac{1}{1+\exp\left(-\frac{p_r}{s_r}+H_r\right)},\quad H_r = P(s_i)\ln P(s_i) \tag{1}$$

式中,σ_r 为区域r 的影响力权重;p_r 为区域r 的兴趣点数量;s_r 为区域r 的面积;s_i 为区域r 中类型为i 的兴趣点出现概率;H_r 为兴趣点类型熵。

本文采用稀疏增量式生成模型(sparse additive generative model ,SAGM)[11] 表达微博主题受到区域主题分布和背景主题分布的共同影响。SAGM 是在表达变量受到多个因素(这些因素需可由指数家族概率分布表达)影响时,在一个指数分布中对代表不同影响因素的参数进行混合,近似代表多个因素的总体影响,提高计算效率。因此,基于稀疏增量式生成模型,微博主题z 受区域主题分布和背景主题分布的影响可表示为:

$$P(z \mid \theta_0,\theta_r) := P(z \mid (1-\sigma_r)\theta_0 + \sigma_r \times \theta_r) \tag{2}$$

1.3　时空主题模型形式化描述

微博的主要生成过程如下:通过背景主题分布和区域主题分布的采样获取微博主题z;基于主题z 下各词汇和时态的分布概率,采样获取微博词汇w 和时间t。反复这个过程,直至该微博所有词汇及其对应的时间生成完毕。微博生成过程的形式化描述如图1 所示。

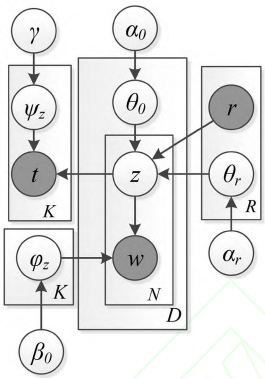

图1　时空主题模型

Fig.1　Spatio-Temporal

1) 从Dirichlet(α_0)中抽样得到背景主题参数θ_0。

2) 对于每个区域$r=1,\ldots,R$:① 从均匀分布中抽取一个区域r;② 从Dirichlet(α_r)中抽样得到区域r 的主题分布参数θ_r。

3) 对于每个主题$z=1,\ldots,K$:① 由σ_r、θ_0 和θ_r 构造主题多项式分布,并从中抽取出主题z,如式(2);② 从Dirichlet(β_r)中得到词汇分布参数ϕ_z;③ 从Beta(γ)中抽得到时间分布参数ψ_z。

4) 对于微博d 中的每个词汇$w=w_1,\ldots,w_{N_d}$ 和相应的时间$t=t_1,\ldots,t_T$:① 从ϕ_z 为参数的多项式分布中抽取词汇w;② 从ψ_z 为参数的Beta 分布中抽取时间t。

以上概率公式中,α_0、α_r、β_0、γ 为θ_0、θ_r、ϕ_z、ψ_z 这些多项式分布参数的超参数,需预先设定。

1.4　时空主题模型参数计算

利用Gibbs 采样[12] 估计时空主题模型的参数。模型需要求解的潜在变量为每条微博d 中词汇i 的主题分布z_{di},其后验概率可写成:

$$P(z_{di} \mid W,Z_{-di},R,T,\alpha_0,\alpha_r,\beta_0,\gamma) = \frac{P(W,Z,R,T,\alpha_0,\alpha_r,\beta_0,\gamma)}{P(W,Z_{-di},R,T,\alpha_0,\alpha_r,\beta_0,\gamma)} = \frac{P(Z,W,T,R \mid \alpha_0,\alpha_r,\beta_0,\gamma)}{P(Z_{-di},W_{-di},T_{-di},R_{-di} \mid \alpha_0,\alpha_r,\beta_0,\gamma)} \tag{3}$$

基于时空主题模型中各元素依赖关系并引入欧拉公式,可得:

$$P(z_{di} \mid W,Z_{-di},R,T,\alpha_0,\alpha_r,\beta_0,\gamma) \propto \frac{(1-t_z)^{\psi_z^1+\gamma^1-1} t_z^{\psi_z^2+\gamma^2-1}}{B(\psi_z^1+\gamma^1-1,\psi_z^2+\gamma^2-1)} \times \frac{n_{z_{di}}+\beta_i-1}{\sum_v^V (n_{z_{dv}}+\beta_v)-1} \times \frac{(\sigma_r)n_{z_r}+\alpha_{z_r}-1}{\sum_z^K ((\sigma_r)n_{z_r}+\alpha_{z_r})-1} \times$$

$$\frac{(1-\sigma_r)n_{z_d}+\alpha_{z_0}\quad 1}{\sum_z^K((1-\sigma_r)n_{z_d}+\alpha_{z_r})-1} \tag{4}$$

式中，n_{z_d} 表示微博d 的词汇i 属主题z 的次数；n_{z_r} 为区域r 属于主题z 的次数；n_{z_d} 为微博d 属于主题z 的次数，α_{z_r} 和α_{z_0} 分别为区域r 中主题z 的先验概率和z 的先验背景概率；t_z 表示时刻t 属于主题z 的次数；ψ_z^1、ψ_z^2 的计算方法参考文献[10]。

模型参数收敛或迭代达次数到阈值后，Gibbs 采样结束，即可获取词汇主题分布、区域主题分布和背景主题分布的参数：

$$\phi_{w,z}=P(w\mid z)=\frac{n_{z_{dw}}+\beta_w\quad 1}{\sum_v^V(n_{z_{dv}}+\beta_v)-1}$$

$$\theta_{r,z}=P(z\mid r)=\frac{n_{z_r}+\alpha_{z_r}\quad 1}{\sum_z^K(n_{z_r}+\alpha_{z_r})-1}$$

$$\theta_{d,z}=P(z\mid d)=\frac{n_{z_d}+\alpha_{0,z}-1}{\sum_z^K(n_{z_d}+\alpha_{0,z})-1}$$

式中，$\phi_{w,z}$ 表示w 属于主题z 的概率；$\theta_{r,z}$ 表示区域r 中主题z 的出现概率；$\theta_{d,z}$ 表示背景主题z 的概率。基于这些分布参数即可求出微博主题分布和时间主题分布[10]。

2 实 验

2.1 数据获取和模型参数设置

利用新浪微博API 下载微博数据后，对其进行去除助词、停用词和符号等预处理，最终得到128 023 条带有地理标识的微博数据。模型的超参数α_0、α_r、β_0、γ_1、γ_2 分别为50/Z、50/Z、0.01、0.1、0.1，Z 为主题数量。实验使用的服务器配置为Intel(四核，3.1 G)酷睿i5 3450，8GB 内存，Windows Server 2008 操作系统。

2.2 微博主题的时间分布

实验提取出最受关注的5 个主题(图2)：娱乐、社会、饮食、工作和生活，给出了在2012-09-30～10-09 期间，这5 个主题在武汉市的强度变化情况。黄金周期间，"娱乐"主题的频度非常高，"饮食"和"社会"主题次之，"工作"相关的主题受关注的程度最低；在黄金周之后，娱乐(旅游)主题的频度迅速降低，"生活"主题的频度变为最高，而"饮食"和社会主题次之，"工作"和"娱乐"主题的关注度交替为最低。从以上两类时间段的主题关注度可知，"饮食"和"社会"主题受到普遍较高的关注度。

2.3 方法比较

采用困惑度(Perplexity)来衡量语言模型对测试语料建模能力的强弱。困惑度越小，表示模型的泛化能力越强[12]，其表示为：

$$Perplexity(D_{test}\mid M)=exp\left\{-\frac{\sum_{d=1}^{M}lgp(w_d)}{\sum_{d=1}^{M}N_d}\right\}$$

将本文的时空主题模型与文献[1,3-4]分别提出的3 个主题模型进行困惑度比较。文献[1]为LDA；文献[3] 的地理主题模型引入了区域要素，但其区域影响权重是固定的，且缺少时间要素，称为RLTM(region-location topic model)；文献[4]的主题模型集成了区域和时间要素，但时间是离散表示的，称为RTTM(region-tempo topic todel)。在测试中，将80% 的数据用来训练相关的主题模型，20% 用来测试主题模型的困惑度。

由图3 发现，在相同主题数量的情况下，我们提出的时空主题模型的困惑度都为最小，几乎是LDA 的一半；而RTTM 次之，RLTM 的模型表达能力一般。此外，在所有主题数目的设置中，RLTM 与LDA 的困惑度较为接近，特别是在主题为180 个至240 个时，RLTM 与LDA 的困惑度几乎相当，这表明RLTM 中主题相互重叠的区域限制了微博主题的正确提取。可见，传统LDA 对微博主题的提取力不从心，但引入时空要素后的主题模型对微博主题的正确提取能力有较显著的提升。

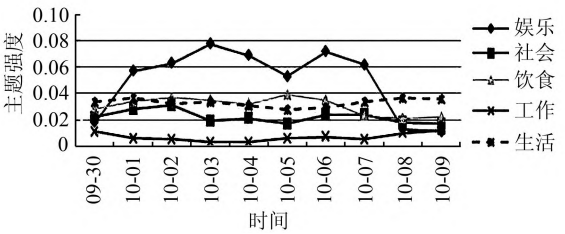

图2 武汉市2012-09-30～10-09 中5 类主题演变情况

Fig.2 Evolutions of Five Topics in Wuhan City from 2012-09-30 to 2012-10-09

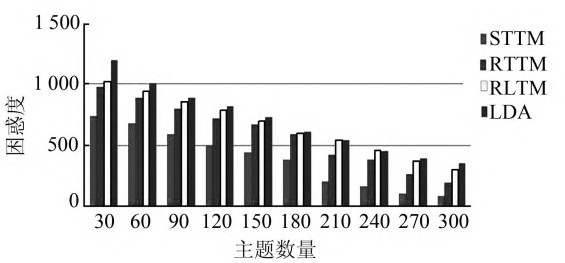

图3 4 个主题模型困惑度比较

Fig.3 Comparison of Perplexity Among Four Topic Models

3　结　语

本文基于城市内不同区域特征、区域影响程度和连续时态要素,阐述了时空主题模型构建和参数估计过程,从时空角度提高微博主题获取的准确性。该模型对预测微博主题时空分布、研究城市居民移动行为模式、发现城市功能分布和转换规律等应用具有重要意义。今后研究将在时空主题模型中集成用户偏好性,以进一步提高微博主题识别性能。

参　考　文　献

[1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1 022

[2] Lu Yue, Zhai Chengxiang. Opinion Integration Opinion Integration Through Semi-supervised Topic Modeling[C]. The 17th International Conference on World Wide Web(WWW), New York, USA, 2008

[3] Yin Zhijun, Cao Liangliang, Han Jiawei, et al. Geographical Topic Discovery and Comparison[C]. The 20th International Conference on World Wide Web(WWW), New York, USA, 2011

[4] Mei Qiaozhu, 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, 2006

[5] Wang Chong, Wang Jinggang, Xie Xing, et al. Mining Geographic Knowledge Using Location Aware Topic Model[C]. The 4th ACM Workshop on Geographical Information Retrieval (GIR), New York, USA, 2007

[6] 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,2010

[7] 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, 2010

[8] Hong Liangjie, Ahmed A, Gurumurthy S, et al. Discovering Geographical Topics In The Twitter Stream[C]. The 21th International Conference on World Wide Web(WWW), Lyon, France, 2012

[9] Pozdnoukhov A, Kaiser C. Space-Time Dynamics of Topics in Streaming Text [C]. International Workshop on Location-Based Social Networks (LBSN), Chicago, USA, 2011

[10] Wang Xuerui, Mccallum A. Topics over Time: A Non-Markov Continuous-time Model of Topical Trends[C]. The 12th International Conference on Knowledge Discovery and Data Mining (KDD), Philadelphia, USA, 2006

[11] 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, 2011

[12] Shan Bin, Li Fang. A Survey of Topic Evolution Based on LDA[J]. Journal of Chinese Information Processing, 2010, 24(6): 43-49(单斌, 李芳. 基于LDA 话题演化研究方法综述[J]. 中文信息学报, 2010, 24(6): 43-49)

Constructing Spatio-Temporal Topic Model for Microblog Topic Retrieving

DUAN Lian[1,2,3]　GUO Wei[1]　ZHU Xinyan[4]　HU Baoqing[2,3]

1 State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

2 Education Ministry Key Laboratory of Environment Evolution and Resources Utilization in Beibu Bay, Nanning 53001, China

3 Resources and Environment Science Department, Guangxi Teachers Education University, Nanning 530001, China

4 Key Laboratory of Aerospace Information Security and Trusted Computing of Minstry of Education, Wuhan University, Wuhan 430079, China

Abstract: Existing geography topic models do not consider the degree to which different regions influence microblog topics. Meanwhile, these models describe the topic evolutions in a discrete manner which prevents the acquisition of topic intensities over continuous time. This paper proposes a novel spatio-temporal topic model to discover microblog topics by introducing continuous time and region in-

(下转第242 页)

Optical and SAR Imagery [C]. EARSeL Joint Workshop: Remote Sensing-New Challenges of High Resolution, Bochum, 2008

[14] Liu Kang, Timo B, Liao Mingsheng. Investigation on Building Height Extraction via Radar Back Scattering Characteristics in High Resolution SAR Image[J]. Geomatics and Information Science of Wuhan University, 2012, 37(7):806-809 (刘康, Timo Balz, 廖明生. 利用后向散射特性从高分辨率SAR 影像中提取建筑物高度[J]. 武汉大学学报· 信息科学版, 2012, 37(7):806-809)

[15] Balz T, Stilla U. Hybrid GPU Based Single and Double-bounce SAR Simulation[J]. IEEE Trans on Geoscience and Remote Sensing, 2009, 47(10): 3 519-3 528

Change Detection for Low Buildings After Earthquake Using Very-high Resolution Pre-event Optical and Post-event SAR Images

ZHANG Bin[1,2]　MA Guorui[2]　YU Jian[3]　WANG Shaojun[1]

1 School of Public Administration, China University of Geosciences (Wuhan), 430074 Wuhan, China

2 State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

3 Wuhan Land Resource and Planning Information Center, Wuhan 430014, China

Abstract: A damage assessment method is proposed by using pre-event very-high resolution (VHR) optical and post-event synthetic aperture radar (SAR) images to detect buildings damaged in an earthquake. First, the length, width, height and other 3-D parameters of a rectangular building are extracted using a pre-event VHR optical image. Second, an image-based GPU ray-tracing approach is used to simulate SAR images. Third, the similarity between the simulated SAR images and post-event actual SAR images is analyzed to determine thief the building is damaged. We demonstrate the feasibility and effectiveness of the method by using remote sensing images of the Sichuan Wenchuan Earthquake of May 12, 2008.

Key words: very high resolution (VHR); synthetic aperture radar (SAR); damage assessment

First author: ZHANG Bin, PhD, specializes in image segmentation, classification, target detection and multiplicative noise removal. E-mail: bin.zhang.whu@gmail.com

Corresponding author: WANG Shaojun, PhD, associate professor. E-mail: 3slab@cug.edu.cn

Foundation support: The Central Universities, China University of Geosciences (Wuhan), No.CUGW140907; the National Natural Science Foundation of China, Nos. 61001187, 41001256; the National High Technology Research and Development Program of China (863 Program), No.2013AA122301.

(上接第212 页)

fluences. A city was divided into multiple geographic regions. Region weights, expressing the region function influence degree on microblog topics, were allocated to regions based on the number of different POI (Point of Interest) types. Then a sparse additive generative model was applied to generate microblog topic distributions. Beta distributions were employed to depict topic evolution over continuous time. Finally, we use a Gibbs sampling method to estimate model parameters. Experimental results showed that not only does our model track the temporal distribution of microblog topics but also enhances topic extraction accuracy when compared with other geography topic models.

Key words: geography topic model; microblog topic mining; temporal-spatio distribution; temporal-spatio inference

First author: DUAN Lian, PhD candidate, specializes in intelligent spatial information service. E-mail: wtusm@163.com

Corresponding author: GUO Wei, PhD. E-mail: geosta@163.com

Foundation support: The National Science and Technology Support Programs of China, No. 2012BAH35B03; the National High-tech R&D Program of China, No. 2011AA010502; the Guangxi Beibu Gulf Key Basic Research Program, No. 2011GXNSFE018003; the Guangxi Beibu Gulf Key Basic Research Program, No. 2012GXNSFEA053001.