在线社交网络中地域性话题发现

曹玖新 胥 帅 陈高君 赵力阳 周 涛 刘 波

(东南大学计算机科学与工程学院 南京 211189)

(东南大学计算机网络和信息集成教育部重点实验室(93K-9) 南京 211189)

摘 要 在线社交网络中日益丰富的地理位置信息为传统舆情感知、信息检索技术带来了新的思考.文中以在线社交平台 Twitter 为研究对象,以社交网络中地域性话题(Geographical Topic)发现为研究目标,工作主要分为社交网络话题性和地域性分析、地域性话题发现两个部分.首先,文中基于用户、位置和话题间的相互关系,阐述了社交网络用户具有地域性和话题性特征,分析了地理位置和话题对词项使用的影响,抽象出地域和话题之间的关联.其次,根据地域性话题的空间关联特征,综合考虑用户发布的文本内容和地理位置信息,按照主题模型思想构建地域性话题发现模型 GTTD(Geographical Textual Topic Discovering model),将用户、话题和地理位置间存在的紧密关系同时引入话题发现框架中.最后利用吉布斯采样算法进行模型的参数估计.基于 Twitter 真实数据集的实验表明:文中提出的 GTTD 模型能有效地发现社交网络中的地域性话题,并且与 LGTA、Geofolk 模型对比,在困惑度(perplexity)指标上体现出优势.

关键词 社交网络;地域性话题;话题发现;主题模型;参数估计

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2017.01530

Discovering Geographical Topics in Online Social Networks

CAO Jiu-Xin XU Shuai CHEN Gao-Jun ZHAO Li-Yang ZHOU Tao LIU Bo

(*School of Computer Science and Engineering, Southeast University, Nanjing* 211189)

(*Key Laboratory of Computer Network and Information Integration of Ministry of Education under Grants No.* 93K-9, *Southeast University, Nanjing* 211189)

Abstract The increasingly rich geographical location information in online social networks has brought new thought to traditional public opinion perception and information retrieval technology. Based on the online social networking platform Twitter, the research in this paper is carried out for the purpose of discovering geographical topics and this work consists of two parts: (1) analysis of the topical and spatial properties of online social networks; (2) discovery of geographical topics. Firstly, based on the relation among users, locations and topics, users' regional and topical features are presented; then the impact of locations and topics on the use of terms is analyzed after which the correlation between regions and topics is abstracted. Secondly, according to the spatial characteristics of geographical topics, we take into consideration both the user generated contents and the geographical location information, and construct the Geographical Textual Topic

收稿日期:2015-12-23;在线出版日期:2016-09-29.本课题得到国家"八六三"高技术研究发展计划项目基金(2013AA013503)、国家"九七三"重点基础研究发展规划项目基金(2010CB328104)、国家自然科学基金(61272531,61202449,61272054,61370207,61370208,61300024,61320106007,61472081)、江苏省网络与信息安全重点实验室(BM2003201)、江苏省科技计划项目(SBY2014021039-10)、无线通信技术协同创新中心、社会公共安全科技协同创新中心资助.**曹玖新**,男,1967 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为复杂网络与社交网络、移动互联网、云环境资源优化与安全技术.E-mail:jx.cao@seu.edu.cn.**胥 帅**,男,1991 年生,博士研究生,主要研究方向为社会计算.E-mail:xushuai7@seu.edu.cn.**陈高君**,男,1988 年,硕士,主要研究方向为社会计算.**赵力阳**,男,1990 年生,硕士研究生,主要研究方向为社会计算.**周 涛**,男,1989 年生,博士研究生,主要研究方向为社会计算.**刘 波**,女,1975 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为普适计算、社会计算.

Discovering model (GTTD) on the basis of theme modeling. The proposed GTTD model is able to introduce the close relationship among user, topic and region into one unified topic discovering framework at the same time. In the end, the Gibbs Sampling algorithm is applied for hidden variable parameter inference for the GTTD model geographical topics effectively, meanwhile it shows better performance in the criteria of perplexity than LGTA model and Geofolk model.

Keywords social networks; geographical topic; topic discovery; topic model; parameter estimation

1 引　言

在线社交网络是当今互联网最为活跃的交流平台，其变革性的力量已经深入社会的各个层面，由其引发的“社会化媒体”现象得到了广泛关注. 随着移动设备成本的降低，社交网络中的位置信息愈发普及[1]，用户在发布信息、上传图片时可以将精确的经纬度信息一同发布，也可以选择粗粒度的城市标签作为地理位置. 国内的新浪微博、腾讯微博，国外的Twitter、Foursquare、Flickr 等社交网络平台的移动端应用都支持类似功能，携带地理位置标记的社交媒体数据(Geo-tagged Documents)正在大量涌现. 当前在线社交网络中普遍存在一种具有空间关联性的用户生成数据(User Generated Data)，即基于地域维度的文本内容信息. 传统意义上，用户发布的文档内容仅由隐含话题决定，不具备区分文档发布地域的能力. 而随着可定位移动设备的普及，含有地理位置标记的用户文本同时受地域和隐含话题影响，相比于传统的用户文档不仅具有地域属性，亦具备区分文档发布地域的性质. 此外，那些在地理位置上接近、个人偏好比较一致的用户所发布的文档有更大的概率拥有一致的话题，并且这些话题有更大的可能出现在同一地域. 我们将这种兼具内容表达能力和地理位置特征的话题称为地域性话题(Geographical Topic). 分析在线社交网络中的地域性话题能够为区域风俗文化差异研究[2]、产品投放策略研究[3]以及网络舆情监控[3]提供重要的借鉴.

如何发现在线社交网络中的地域性话题并比较不同地域的话题分布情况成为了新的研究方向. 本文以在线社交平台 Twitter 为研究对象，基于社交网络用户的地域性和话题性分析，根据地域性话题的空间关联特性，综合考虑社交网络用户、地域、位置、话题和词项 5 种因素，提出地域性话题发现模型 GTTD(Geographical Textual Topic Discovering model)，并利用吉布斯采样算法进行模型的参数估计.

本文第 2 节介绍相关工作与当前研究进展；第 3 节描述 Tweet 文本语料的获取方法和必要的数据预处理操作；第 4 节详细分析社交网络典型的地域性和话题性特征，为模型构建提供理论支撑；第 5 节阐述 GTTD 模型的相关定义与符号表示，并给出参数估计的吉布斯采样算法；第 6 节进行地域性话题发现实验并分析实验结果；第 7 节是对全文工作的总结以及未来研究内容的展望.

2 相关工作

近年来，话题发现研究受到国内外学者的广泛关注，根据研究过程是否考虑地理位置因素，可以分为普通文档话题发现和地域性话题发现两部分. 文献[2]和文献[4-11]从文档生成过程出发，未考虑位置信息，研究了基于概率主题模型的非地域性话题发现. 文献[12-18]针对带有地理位置标记的文档，综合考虑社交媒体的文本特征和空间特征，研究了地域性话题发现.

(1) 非地域性话题发现

Deerwester 等人[4]提出的隐含语义分析模型 LSA 开创了文本主题发现的先河，完成了文档词项可观测空间到隐含语义空间的变换. 基于 LSA 模型的思想，后续的 PLSA 模型[5]和 LDA 模型[6]使用概率理论进行模型构建，在话题发现领域得到了大量应用，其中又以 LDA 模型应用最为广泛. Hong 等人[7]较早将基于 LDA 的 Author-Topic Model 应用在社交网络话题发现中，该模型在原生 LDA 中加入作者变量，将同一作者发表的 Tweet 合并为一篇文档，降低了社交网络短文本的稀疏性. Huang 等人[8]针对稀疏的微博客短文本提出了基于 LDA 模型和 Single-Pass 聚类的方法，实现在微博客上的

话题发现. AlSumait 等人[9]提出 Online-LDA 模型,引入时间因素,将已建立模型的先验参数作为新模型的输入,在实时更新的在线文本流中识别主题模式. Fang 等人[2]提出一种多维特征话题发现模型,融合用户属性、社交关系和时序因素 3 个方面特征设计话题发现算法. Vosecky 等人[10]认为 Twitter 平台上的每一篇推文都是由用户、组织、位置、词项和时间 5 种因素构成,5 种因素相互独立,以多项分布的形式共同构成推文中的隐含话题. 其提出的 MfTM 模型采用随机变分推断方法(stochastic variational inference)提升了参数估计速度,适用于大规模数据流环境下的推特话题发现和文本聚类. Guo 等人[11]针对在线论坛环境下的话题发现,首先利用传统 LDA 算法获取每个主题下的词项分布以及文档下的主题分布;接着,依据主题间的相关度构建主题邻接矩阵、融入时间序列构造三阶主题张量;最后,采用典范分解(canonical polyadic decomposition)识别时间序列上的主题.

(2)地域性话题发现

以上针对在线社交网络的话题发现仅从用户文本内容出发,并没有考虑日益丰富的地理位置信息,忽略了地域因素对话题产生的影响. Eisenstein 等人[12]将地理区域作为隐变量引入主题模型中,最先从单纯的文本处理角度提出了地理位置的应用方法. Sizov 提出的 Geofolk 模型[13]是对 LDA 模型的扩展,认为带有经纬度信息和文字标签的 Flickr 图片同时由话题和位置生成,话题生成词项的同时还生成图片的经度和纬度位置,也就是说 Geofolk 模型实质上是在原始 LDA 模型中增加了主题生成位置的过程. Yin 等人[14]亦研究了图片社交平台 Flickr 上的地域性话题发现,提出了 LGTA 模型. 其同时从地域聚类和话题建模两方面入手,不仅能发现地域性话题,还可以比较同一话题在不同地域的分布. 给定一个话题,借助 LGTA 模型,可以比较其在不同地域的流行程度. 本文提出的 GTTD 模型将主要与上述 Geofolk 模型和 LGTA 模型对比实验效果,原因主要体现在:① 本文模型与上述模型的基本假设相同,即文档发布者的空间距离越接近,其属于同一地域的概率越大,微博拥有一致话题的概率也越大;② 本文提出的 GTTD 概率图模型的生成过程与上述模型类似,但将用户的位置偏好(地域分布)也视为变量,与地域-话题多项分布和话题-词项多项分布类似,也需要通过 Dirichlet 先验分布产生.

Yuan 等人[15]研究了 Twitter 平台上的用户行为模式,将用户、时间、地域和用户活动统一起来,以天为时间单位划分签到数据,构建了时空主题发现模型 W^4. Liu 等人[16]认为用户在社交网络上发布的内容随着时间演化,每个用户在不同时间都对应着不同的主题分布和地域分布. 其采用连续的时间对用户在 LBSN 中的签到建模,避免了 W^4 模型中以天为单位划分用户文档带来的信息缺失,美中不足的是,其并未考虑用户生成的文本数据信息. Zhang 等人[17]提出了先按照地理位置聚类再识别文档主题的方法,其采用 DBSCAN 算法将标有地理位置的文档聚类,每个类都由处于该地理位置内的文档组成;然后,将每个地域内的文档拼接成一篇文档,构成跟地域数目等量的文档集合,最后将传统 LDA 算法应用于该文档集合进行话题识别. Hu 等人[18]综合考虑社交网络用户偏好信息和用户时空行为模式,研究了基于位置的社交网络中用户签到的生产过程,认为一条用户签到不仅受用户的主题历史分布、地域历史分布以及当前时间的影响,同时亦受到地域和签到位置背景分布(background distribution)的影响. 其提出的 STT 模型能够发现用户签到位置的主题.

现有关于地域性话题发现的研究工作主要存在如下两方面的不足:(1)现有工作充分研究了用户签到数据的时空模式,却忽视了社交媒体中用户发表的文本信息;(2)现有研究多认为用户的地域分布是不变的常量,即所有用户具有相同的地域分布,或者对给定用户,其地域分布保持不变. 但实际情况下,不同用户往往对应着不同的地域分布,表现为其出现在特定地域的概率不同. 针对上述研究存在的问题,本文首先阐述了用户的地域性和话题性特征,引出地域和话题对词项使用的影响;然后,综合考虑用户发布的文本内容和地理位置信息,将用户地域分布也视为变量,按照主题模型思想构建地域性话题发现模型 GTTD,并利用吉布斯采样算法进行模型的参数估计,实现在线社交网络中的地域性话题发现.

3 数据获取与处理

在线社交网络的话题发现依赖于文本挖掘技术,需要使用大量的文本数据进行模型训练,因此在 Twitter 平台上进行数据抓取是本研究的首要任

务.本文首先基于 Twitter API 进行数据抓取,为构建 Twitter 平台地域性话题发现模型提供基础.然后,根据研究需要对数据集进行必要的预处理.

3.1　数据获取

Streaming APIs① 提供的数据是全球范围内的采样数据,并且可以根据条件进行数据过滤,数据抓取程序使用到的 filter 端点过滤条件如表 1 所示.

表 1　数据过滤条件

过滤条件	说明
地理位置	逗号分隔的经纬度对
语言	推文的语言种类
地理信息	是否携带经纬信息
关键词	关键字序列,以英文逗号隔开

不施加任何过滤条件的情况下,filter 端点提供全球范围的 Tweet 采样数据,这种数据地域分布广阔、语言种类繁杂,而且并非所有的 Tweet 都携带地理位置信息.本文地域性话题发现模型的建立需要使用地理位置信息,因此从可行性角度出发,考虑到使用英文语料可以减少分词错误,将数据抓取范围设定在美国本土,并只抓取携带地理位置信息的英文 Tweet.

为了更好地支持本研究的实验工作,研究抓取了四个数据集:全美本土范围数据集(USA)、纽约数据集(NYC)、旧金山数据集(SFO)和篮球数据集(Basketball).USA、NYC 和 SFO 数据集限定了语言种类为英文且携带地理位置信息,经纬度范围分别为美国本土范围、纽约城范围和旧金山范围.Basketball 数据集则根据 USA 数据集进一步筛选获得,如果 Tweet 的文本中包含篮球专业词语、NBA 球队名称、NBA 球星的名字,则计入 Basketball 数据集.需要注意的是,由于本文的研究对象是社交网络中的地域性话题,而社交网络中的话题往往不会长时间存在,并且典型话题仅在一到两天内便可引起网民的热议,因此抓取数据集的时间范围并不需要很大跨度.四个数据集的统计情况见表 2.

表 2　数据集相关情况

名称	时间范围	Tweet 数量	用户数
USA	2015-03-05 至 2015-03-06	4291464	641906
NYC	2015-05-14 至 2015-05-18	584746	77684
SFO	2015-05-04 至 2015-05-09	463969	44151
Basketball	2015-03-05 至 2015-03-06	72586	17304

3.2　数据预处理

首先,数据抓取程序获得的数据包括每条 Tweet 的完整信息,为了便于后续实验,研究提取每条 Tweet 的 ID、产生时间、经纬度、文本内容和用户 ID 字段.

其次,由于获得的 Tweet 数据质量参差不齐,不宜直接作为模型输入,因此进行如下数据净化工作:

(1)去除 Tweet 文本中的停用词(如 is、the、of)以及表情符号、“RT”、“retweet”、URL 等标记;同时,去除标点符号和数字;

(2)有研究表明[19],用户发表的 Twitter 英文文档的平均长度为 14 个英文单词,我们可以认为长度短于 10 个单词的 Twitter 文档实际含义不大,对于话题发现工作无益,因此在数据预处理环节删除长度小于 10 个单词的 Tweet;

(3)Tweet 文本英文单词统一转换成小写形式,并使用开源工具 NLTK② 对所有单词进行词根还原.

4　社交网络地域性与话题性分析

本节对在线社交网络用户的地域性和话题性进行分析,旨在说明用户所处地域一定程度上决定了其要发表内容的话题,且地域和话题共同决定词项的使用,进而影响文档的生成.

基于上述目的,本文结合理论分析与实验验证阐述 3 个重要假设,为构建地域性话题发现模型提供支撑.

(1)社交网络中的用户具有地域性和话题性.这里使用一个 Foursquare 数据集来进行实验,该数据集是本项目组从 Foursquare 社交网络平台抓取的真实数据集,其中包含 6291 个用户的 788208 个签到,用户的主要生活城市在纽约.实验将一个用户所有签到的经纬度均值作为用户签到的中心点,计算该用户的每个签到和中心点之间的距离,并按距离范围统计频次,计算对应频次占整个签到次数的比例,图 1 展示了实验结果.可见,签到位置与签到中心的距离呈现出幂律分布的规律:距离签到中心点在 1 km 到 10 km 之内的签到比例较大,其中在中心点 1 km 以内的签到超过 20%;距离在 100 km 到 10000 km 的签到所占比例极小.这说明社交网络用户的活动范围有限,呈现出明显的地域性分布.

① https://dev.twitter.com
② http://www.nltk.org/api/nltk.stem.html

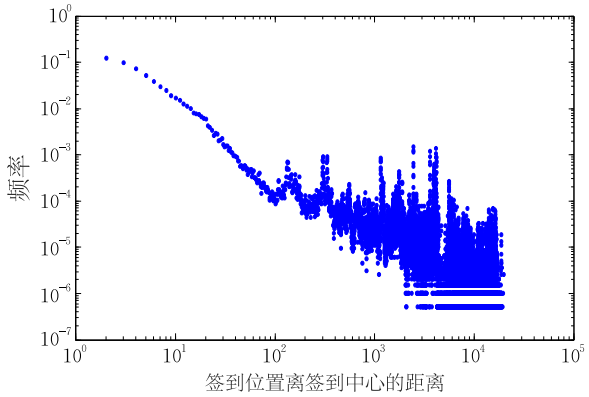

图 1 用户签到距其签到中心点的距离频率分布

此外,话题标签(hashtag)在社交平台广泛存在,用户发布的博客内容往往与特定话题紧密相关[2].对 USA 数据集的统计发现,超过 10%的原创 Tweet 和 20%的转发 Tweet 含有话题标签.不同用户倾向于讨论不同类型的话题,反映了 Twitter 用户兴趣的差异性.在线社交平台中丰富的话题类别和参与人群表明社交网络用户具有话题性.

(2)在线社交网络中地域会对话题产生影响.同样,采取验证性实验为分析提供支持.在实验中选定 wontons、suey、noodle、dumplings、sesame chicken 等词语作为抓取关键字,在纽约地区抓取 10 万条包含这些单词的 Tweet,将这些 Tweet 按照坐标在纽约地图上显示出来(如图 2),可以看出 Tweet 在纽约的各个地区都有分布,尤其在曼哈顿(Manhattan)的中南部最为密集,表明 Tweet 的位置点呈现出簇状分布的特征.实际上,这个区域商业发达,华人的数量、中国餐馆的数量都远高于纽约的其他街区,讨论中国饮食的话题也相应较多.

图 2 纽约地区包含中国饮食相关单词的 Tweet 地理分布

(3)在线社交网络中用户使用词项同时受地域和话题的影响.具体到 Twitter 平台,用户将日常的所见所闻使用文字表达,其使用的词语是在地理区域和话题的共同作用下产生.为了支持这一假设,本文通过数据抓取与分析做出以下验证性的实验.在同一时间段,分别在美国佛罗里达州的迈阿密(Miami)、堪萨斯州的堪萨斯城(Kansas City)、密苏里州的圣路易斯(Saint Louis)以及蒙大拿州的海伦娜(Helena)抓取 Tweet 数据,统计 love、day、good、time、people、life 等热度较大的词和单词 beach 的出现频次,为避免各城市采样 Tweet 数目不均带来的样本失衡问题,我们计算各城市各单词的"词频/采样数"比值并绘制结果如图 3 所示,图中纵轴数值表示词频与样本数的比值,即平均每篇文档包含的对应单词数.

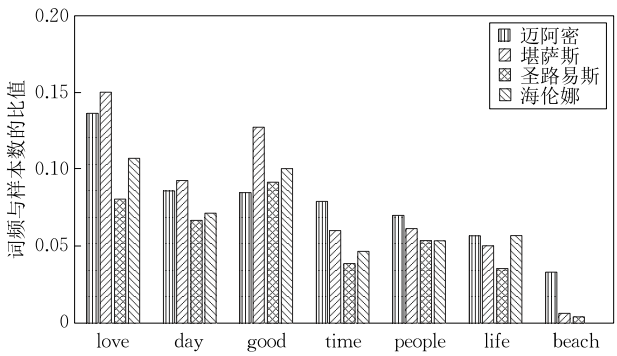

图 3 目标单词在不同城市的词频采样比

实验中除 beach 之外的单词在英语中较为常见,可以代表某地区正常的词频现象.由图可见,迈阿密数据中 beach 出现的平均频次远多于其它三个地区,而对于 love、day、good 等热度较大的单词各地区 Tweet 中出现的平均词频并无显著差距.这种现象与迈阿密的亚热带气候和发达的海滩文化有很大关系,相比之下地处美国中部的堪萨斯城和圣路易斯则没有更加突出的词频现象,而美国本土最北部的海伦娜地区 beach 一词出现的平均频次更加低.这些单词在上述城市 Tweet 数据中的出现频次支持了地域因素影响社交网络单词使用的假设.

5 地域性话题发现模型

本节在分析社交网络用户具有地域性和话题性特征的基础上,介绍地域性话题概念,综合考虑用户发布的文本内容和地理位置信息,阐述了模型的构建依据:地理位置接近、个人偏好比较一致的用户所发布的文档有更大的概率拥有一致的话题,并且这些话题有更大的可能出现在同一地域.根据主题模型的建模思想[6],本节构建了地域性话题发现模型(Geographical Textual Topic Discovering model,

GTTD)并给出模型参数的求解算法.

5.1 模型概述及符号定义

主题模型思想体现在文档的生成过程之中，本文将用户发表的每篇 Tweet 视为文档，给出如下定义.

定义 1. 文档是携带 GPS 信息的 Tweet，形式化地表示为 $d=\{u,w_d,l_d\}$，其中 u 表示发表该 Tweet 的用户，w_d 是一组单词的集合，$l_d=(x_d,y_d)$ 表示 Tweet 携带的 GPS 坐标，x_d 和 y_d 分别表示经度和纬度.

除了可观测的用户、位置和词项外，模型继续引入地域和话题的概念，使模型成为用户、地域、位置、话题和词项 5 种对象组成的整体. 首先，用户无论处于何地都属于一个特定的地域，这个地域代表了文化背景和环境因素. 然后，根据用户和地域确定一篇文章反映的主旨思想，即生成具体的话题. 基于地域和话题，每篇文档中使用的词项被确定下来. 依据这样的过程，一篇文档的生成与多种因素产生关联. 那些在地理位置上接近，兴趣爱好又比较一致的用户所发布的文档有更大的概率拥有一致的话题，这些话题有更大的可能出现在同一地域. 因此模型发现的话题不但具有内容表达能力，同时具有地理特征，可称之为地域性话题，这里给出地域性话题的具体说明：

地域性话题是一种具有空间关联性的话题，具备以下两方面性质：(1) 表达能力. 话题的语义表达能力是话题的基本特征，表现为语言含义上的关联性；(2) 地域性. 传统意义上话题没有空间概念，地域性话题在发现过程中添加了空间维度，相关话题的文档集合不但内容含义相似，而且在地理区域上较为接近.

为方便模型描述和公式推导，表 3 给出了本文使用到的符号及其对应的含义.

表 3　模型中的符号定义

符号	含义	符号	含义
u	用户	$\boldsymbol{U}$	用户集合
r	地域	$\boldsymbol{R}$	地域集合
z	话题	$\boldsymbol{Z}$	话题集合
w	单词	$\boldsymbol{V}$	单词表
d	文档	$\boldsymbol{D}$	文档集合
l	文档的地理位置	$mergeDist$	地域合并距离
$\boldsymbol{\theta}$	用户地域对多项分布	α	$\boldsymbol{\theta}$ 的 Dirichlet 先验参数
$\boldsymbol{\Phi}$	地域话题对多项分布	β	$\boldsymbol{\Phi}$ 的 Dirichlet 先验参数
$\boldsymbol{\pi}$	用户的地域多项分布	η	$\boldsymbol{\pi}$ 的 Dirichlet 先验参数
τ	模型起始变量用户的分布参数	σ	混合高斯分布的标准差

5.2 模型构建与表示

首先，由于混合高斯分布能够以任意精度逼近任何分布[20]，而文档的话题由地域生成，因此可以认为所有文档的位置服从混合高斯分布，混合分布中的每一维高斯分布即是由地理位置组成的一个地域，地域 $r\in\boldsymbol{R}$ 有中心点 loc_r. 生成地域时，先选取发布文档数量超过 5 次的用户，将其所有文档的位置中心作为一个地域的中心位置，得到初始地域的中心点集合. 然后，当两个初始地域中心点之间的距离小于 $mergeDist$ 时，撤销两个地域再合并为一个地域，并用两个点的均值作为新地域的中心点. 数据集中的文档与各个地域的关系按照式(1)判别：

$$P(d,r)=P(r)P(d|r) \tag{1}$$

其中 $P(d|r)=\boldsymbol{N}(l_d|l_r,\sigma)$，因此一篇文档 d 归属地域 r 的概率取决于文档的位置与地域中心点的距离，距离越近，文档 d 属于地域 r 的概率越大. 基于这种地域的概念，在空间上比较接近的用户所发的 Tweet 更可能被划分到一个地域中. 通过以上过程，模型可以完整得到数据集中的地域集合 $\boldsymbol{R}=\{r_0,r_1,\cdots,r_{|\boldsymbol{R}|}\}$.

在 GTTD 模型中，生成文档时可先为用户采样得到一个具体的地域，接着根据泊松分布得到文档的词项数量. 在生成每个词项时，为每个词采样一个话题，根据话题的词项分布得到具体词项. 因此，模型中的地域可以被认为是 K 个话题的混合，特定话题在某些地域的出现概率高，而其它话题在该地域出现概率低. 如果两篇文档已经归属同一个地域，则被分配同一个话题的概率就相应变大.

基于以上模型构建分析，可以给出文档语料的生成过程如下：

过程 1. 生成文档语料.

1. 采样 $\boldsymbol{\pi}=\{\boldsymbol{\pi}_u\sim Dir(\eta)\}$，$u\in\boldsymbol{U}$，$\boldsymbol{\pi}_u$ 为用户 u 的地域分布向量；

2. 采样 $\boldsymbol{\theta}=\{\boldsymbol{\theta}_{u,r}\sim Dir(\alpha)\}$，$u\in\boldsymbol{U}$，$r\in\boldsymbol{R}$，$\boldsymbol{\theta}_{u,r}$ 为用户地域对$\langle u,r\rangle$的话题分布向量；

3. 采样 $\boldsymbol{\Phi}=\{\boldsymbol{\Phi}_{r,k}\sim Dir(\beta)\}$，$r\in\boldsymbol{R}$，$k\in\boldsymbol{K}$，$\boldsymbol{\Phi}_{r,k}$ 为地域话题对$\langle r,k\rangle$的词项分布向量；

4. 对每个用户 $u\in\boldsymbol{U}$：

4.1 采样一个地域 $r_u\sim Mult(\boldsymbol{\pi}_u)$；

4.2 对用户 u 所有文档 $d\in\boldsymbol{D}_u$：

4.2.1 采样文档的位置：

$$l_d\sim\exp\left(-\frac{dist(l_d,l_r)}{\sigma^2}\right);$$

4.2.2 对文档 d 中所有词项 $w\in N_{u,d}$：

(1) 采样一个主题：

$$z_{u,d,w}\sim Mult(\boldsymbol{\theta}_{u,r});$$

(2)采样一个词项:

$$w_{u,d}\sim Mult(\boldsymbol{\Phi}_{z_{u,d,w}}).$$

本文提出的地域性话题发现模型 GTTD 如图 4 所示,由于用户、位置和词项都是可观测变量,用填充节点 u,l 和 w 表示,其他变量和参数使用空节点表示. 节点之间的边表示变量或参数之间存在依赖关系,例如:话题变量 z 和地域变量 r 指向词项变量 w,表示词项的生成受到话题和地域的影响. 给定用户 u,地域 r 和话题 z,$\boldsymbol{\theta}$ 为话题 z 在用户地域对 $\langle u,r\rangle$ 上的概率分布,即 $p(z|u,r)$;$\boldsymbol{\Phi}$ 为词项 w 在地域话题对 $\langle r,z\rangle$ 上的概率分布,即 $p(w|r,z)$;$\boldsymbol{\pi}$ 为用户 u 的地域分布向量,即 $p(r|u)$;α,β,η 分别是参数 $\boldsymbol{\theta},\boldsymbol{\Phi},\boldsymbol{\pi}$ 的 Dirichlet 先验分布参数. 图中的矩形框表示抽样过程,矩形框右下角标明了抽样维度,例如:图中最大的矩形框表示对用户集合 $\boldsymbol{U}$ 执行内部抽样的过程,而参数 $\boldsymbol{\theta}$ 外的矩形框表示 $\boldsymbol{\theta}$ 是一个 $\boldsymbol{U}\times\boldsymbol{R}$ 维的矩阵,话题由用户集合 $\boldsymbol{U}$ 和地域集合 $\boldsymbol{R}$ 共同确定,故需进行 $\boldsymbol{U}\times\boldsymbol{R}$ 次抽样.

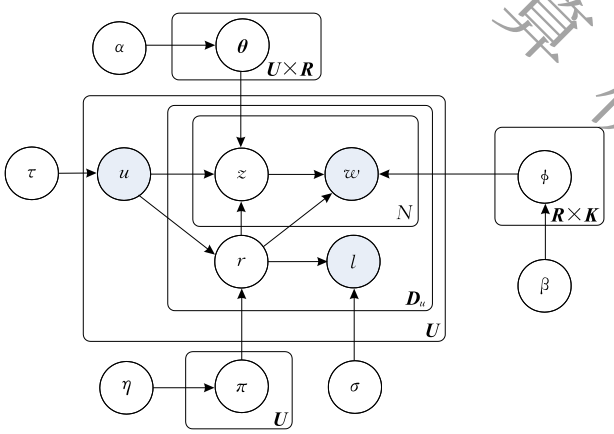

图 4 GTTD 模型的贝叶斯图表示

5.3 GTTD 模型参数估计

在 GTTD 模型中,用户的地域分布向量 $\boldsymbol{\pi}$、用户地域对 $\langle u,r\rangle$ 上的话题分布向量 $\boldsymbol{\theta}$、地域话题对 $\langle r,z\rangle$ 上的词项分布向量 $\boldsymbol{\Phi}$ 都是隐含变量,模型中包含连续的高斯分布和离散的多项分布,直接进行精确的参数推导较为困难. 一般地,概率主题模型采用近似推导的参数估计方法获得模型隐含变量的分布,主要方法包括变分 EM 算法[6]和基于蒙特卡洛马尔可夫链理论(MCMC)的吉布斯采样算法(Gibbs Sampling)[21],前者的思想是最大后验估计(MAP),后者的思想是贝叶斯估计(Bayesian Estimation). 文献[13]利用吉布斯采样获得图片的主题分布. 文献[14-15]采用变分 EM 算法求得地域上的话题分布及话题上的词项分布. 由于本文提出的地域性话题发现模型 GTTD 将待估计参数看做服从 Dirichlet 先验分布的随机变量,因此选择吉布斯采样算法进行隐变量推导.

根据 GTTD 模型对文档语料生成过程的描述,可以得到文档集合的联合概率分布如式(2):

$$\begin{aligned}P(\boldsymbol{\Phi},\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{z},\boldsymbol{r},SN|\tau,\alpha,\beta,\eta,\sigma)=&\prod_{u\in\boldsymbol{U}}P(u|\tau)\times \quad \text{(I)}\\ &P(\boldsymbol{\pi}|\eta)\prod_{u\in\boldsymbol{U}}\prod_{d\in\boldsymbol{D}_u}P(r|\boldsymbol{\pi}_u)\times \quad \text{(II)}\\ &\prod_{u\in\boldsymbol{U}}\prod_{d\in\boldsymbol{D}_u}P(l|\sigma,l_r)\times \quad \text{(III)}\\ &P(\boldsymbol{\theta}|\alpha)\prod_{u\in\boldsymbol{U}}\prod_{d\in\boldsymbol{D}_u}\prod_{w\in d}P(z|\boldsymbol{\theta}_{u,r})\times \quad \text{(IV)}\\ &P(\boldsymbol{\Phi}|\beta)\prod_{u\in\boldsymbol{U}}\prod_{d\in\boldsymbol{D}_u}\prod_{w\in d}P(w|\boldsymbol{\Phi}_{r,z}) \quad \text{(V)}\end{aligned} \tag{2}$$

其中,r 是所有文档的地域变量,z 是所有文档的话题变量,SN 表示来自社交网络的语料中用户、位置和词项的集合. 将模型联合概率分布的表达式拆分,利用多项分布和狄利克雷分布的共轭关系可以推导出多项分布 $\boldsymbol{\theta}$、$\boldsymbol{\Phi}$ 和 $\boldsymbol{\pi}$. 式(2)中的(I)和(III)不依赖于多项分布参数,(II)依赖于 $\boldsymbol{\pi}$,(IV)依赖于 $\boldsymbol{\theta}$,(V)依赖于 $\boldsymbol{\Phi}$. 因此可以进一步化简公式(II),(IV)和(V),推导规则见式(3)~(5):

$$\begin{aligned}\text{(II)}&=\prod_{u\in\boldsymbol{U}}P(\boldsymbol{\pi}_u|\eta)\times\prod_{u\in\boldsymbol{U}}\prod_{d\in\boldsymbol{D}_u}P(r|\boldsymbol{\pi}_u)\\ &=\prod_{u\in\boldsymbol{U}}P(\boldsymbol{\pi}_u|\eta)\times\prod_{d\in\boldsymbol{D}_u}P(r|\boldsymbol{\pi}_u)\\ &=\prod_{u\in\boldsymbol{U}}\frac{1}{Beta(\eta)}\prod_{r\in\boldsymbol{R}}\boldsymbol{\pi}_{u,r}^{\eta_r-1}\prod_{r\in\boldsymbol{R}}\boldsymbol{\pi}_{u,r}^{n_r^{(u)}}\\ &\propto\prod_{u\in\boldsymbol{U}}\prod_{r\in\boldsymbol{R}}\boldsymbol{\pi}_{u,r}^{n_r^{(u)}+\eta-1}\end{aligned} \tag{3}$$

$$\begin{aligned}\text{(IV)}&=\prod_{u\in\boldsymbol{U}}\prod_{r\in\boldsymbol{R}}P(\boldsymbol{\theta}_{u,r}|\alpha)\times\prod_{u\in\boldsymbol{U}}\prod_{d\in\boldsymbol{D}_u}\prod_{w\in d}P(z|\boldsymbol{\theta}_{u,r})\\ &=\prod_{u\in\boldsymbol{U}}\prod_{r\in\boldsymbol{R}}P(\boldsymbol{\theta}_{u,r}|\alpha)\prod_{z\in\boldsymbol{Z}}P(z|\boldsymbol{\theta}_{u,r})^{n_z^{(u,r)}}\\ &=\prod_{u\in\boldsymbol{U}}\prod_{r\in\boldsymbol{R}}\frac{1}{Beta(\alpha)}\prod_{z\in\boldsymbol{Z}}\boldsymbol{\theta}_{u,r,z}^{\alpha_z-1}\prod_{z\in\boldsymbol{Z}}\boldsymbol{\theta}_{u,r,z}^{n_z^{(u,r)}}\\ &\propto\prod_{u\in\boldsymbol{U}}\prod_{r\in\boldsymbol{R}}\prod_{z\in\boldsymbol{Z}}\boldsymbol{\theta}_{u,r,z}^{n_z^{(u,r)}+\alpha_z-1}\end{aligned} \tag{4}$$

$$\begin{aligned}\text{(V)}&=\prod_{r\in\boldsymbol{R}}\prod_{z\in\boldsymbol{Z}}P(\boldsymbol{\Phi}_{r,z}|\beta)\times\prod_{u\in\boldsymbol{U}}\prod_{d\in\boldsymbol{D}_u}\prod_{w\in d}P(w|\boldsymbol{\Phi}_{r,z})\\ &=\prod_{r\in\boldsymbol{R}}\prod_{z\in\boldsymbol{Z}}P(\boldsymbol{\Phi}_{r,z}|\beta)\times\prod_{r\in\boldsymbol{R}}\prod_{z\in\boldsymbol{Z}}\prod_{w\in\boldsymbol{V}}P(w|\boldsymbol{\Phi}_{r,z})^{n_w^{(r,z)}}\\ &=\prod_{r\in\boldsymbol{R}}\prod_{z\in\boldsymbol{Z}}\frac{1}{Beta(\beta)}\times\prod_{w\in\boldsymbol{V}}\boldsymbol{\Phi}_{r,z,w}^{\beta_w-1}\prod_{w\in\boldsymbol{V}}\boldsymbol{\Phi}_{r,z,w}^{n_w^{(r,z)}}\\ &\propto\prod_{r\in\boldsymbol{R}}\prod_{z\in\boldsymbol{Z}}\prod_{w\in\boldsymbol{V}}\boldsymbol{\Phi}_{r,z,w}^{n_w^{(r,z)}\beta_w-1}\end{aligned} \tag{5}$$

其中 $n_r^{(u)}$、$n_z^{(u,r)}$ 和 $n_w^{(r,z)}$ 的含义依次表示:地域 r 中出现用户 u 的次数、话题 z 在用户地域对 $\langle u,r\rangle$ 中出现的次数以及词项 w 在地域话题对 $\langle r,z\rangle$ 中出现

的次数.

公式推导的目的是为了得到当前词项的话题变量和地域变量.根据模型中马尔可夫链的性质和贝叶斯定理,用 z_w 和 r_w 表示当前词项 w 的话题变量和地域变量,可以将式(2)中隐变量 z 和 r 的联合概率写为式(6)的形式:

$$P(r_w,z_w|\boldsymbol{r}_{-w},\boldsymbol{z}_{-w},SN;\tau,\alpha,\beta,\eta,\sigma)=$$
$$P(u|\tau)\times P(l|\sigma,l_r)\cdot\times\frac{n_{r_w}^{(u_w)}+\eta_{r_w}}{\sum_{r\in\boldsymbol{R}}(n_r^{(u_w)}+\eta_r)}\times$$
$$\frac{n_{z_w}^{(u_w,r_w)}+\alpha_{z_w}}{\sum_{z\in\boldsymbol{Z}}(n_z^{(u_w,r_w)}+\alpha_z)}\cdot\times\frac{n_w^{(r_w,z_w)}+\beta_w}{\sum_{w'\in\boldsymbol{V}}(n_{w'}^{(r_w,z_w)}+\beta_{w'})}\tag{6}$$

根据式(6),可得到话题变量 z_w 的采样式(7)和地域变量 r_w 的采样式(8).

$$P(z_w|z_w,\boldsymbol{r}_{-w},\boldsymbol{z}_{-w},SN;\tau,\alpha,\beta,\eta,\sigma)\propto$$
$$\frac{n_{z_w}^{(u_w,r_w)}+\alpha_{z_w}}{\sum_{z\in\boldsymbol{Z}}(n_z^{(u_w,r_w)}+\alpha_{z_w})}\times\frac{n_w^{(r_w,z_w)}+\beta_w}{\sum_{w\in\boldsymbol{V}}(n_w^{(r_w,z_w)}+\beta_{w'})}\tag{7}$$
$$P(r_w|z_w,\boldsymbol{r}_{-w},\boldsymbol{z}_{-w},SN;\tau,\alpha,\beta,\eta,\sigma)\propto$$
$$P(l|\sigma,l_r)\times\frac{n_{z_w}^{(u_w)}+\eta_{r_w}}{\sum_{r\in\boldsymbol{R}}(n_r^{(u_w)}+\eta_r)}\times$$
$$\frac{n_{z_w}^{(u_w,r_w)}+\alpha_{z_w}}{\sum_{z\in\boldsymbol{Z}}(n_z^{(u_w,r_w)}+\alpha_z)}\times\frac{n_w^{(r_w,z_w)}+\beta_w}{\sum_{w\in\boldsymbol{V}}(n_w^{(r_w,z_w)}+\beta_{w'})}\tag{8}$$

得到地域和话题隐变量的转移概率推导公式后,可推出各多项分布参数的参数更新规则.由于已知所有分配给用户 u 的地域,记为 $\boldsymbol{r}_u$,用户的地域概率分布见式(9):

$$P(\boldsymbol{\pi}_u|\boldsymbol{r}_u;\eta)=Dir(n^{(u)}+\eta)\tag{9}$$

由于 $\boldsymbol{\pi}_u$ 的后验概率是 Dirichlet 分布,那么用户 u 出现在地域 r 的似然概率恰好为该 Dirichlet 分布的期望,如式(10)所示:

$$\boldsymbol{\pi}_{u,r}=P(r|u)=E[P(\boldsymbol{\pi}_u|r_u;\eta)]=\frac{n_r^{(u)}+\eta_r}{\sum_{r\in\boldsymbol{R}}(n_r^{(u)}+\eta_r)}\tag{10}$$

同理可得参数 $\boldsymbol{\theta}$ 以及参数 $\boldsymbol{\Phi}$ 的更新规则如式(11)、(12):

$$\boldsymbol{\theta}_{u,r,z}=\frac{n_z^{(u,r)}+\alpha_z}{\sum_{z\in\boldsymbol{Z}}(n_z^{(u,r)}+\alpha_z)}\tag{11}$$

$$\boldsymbol{\Phi}_{r,z,w}=\frac{n_w^{(r,z)}+\beta_w}{\sum_{w\in\boldsymbol{V}}(n_w^{(r,z)}+\beta_w)}\tag{12}$$

基于上述的采样公式及参数更新公式,GTTD 模型的吉布斯采样算法如算法 1 所示.算法主要分为三个阶段:初始化阶段、*Burn-in* 阶段和参数更新阶段.在初始化阶段,为每个单词随机采样一个地域编号和话题编号.在 *Burn-in* 阶段,根据采样式(7)和(8)建立话题变量和地域变量的马尔可夫链.在参数更新阶段,统计全局变量并更新多项分布参数.

算法 1.　GTTD 模型的吉布斯采样算法.

输入:$\boldsymbol{D}$,$|\boldsymbol{Z}|$,$mergeDist$,σ,α,β,η

输出:$\boldsymbol{\theta}$,$\boldsymbol{\Phi}$,$\boldsymbol{\pi}$

```
I := Iterations;
    computeRegion(mergeDist);
FOREACH u∈U DO
FOREACH d∈D_u DO
FOREACH w∈d DO
r~uniform();
z~uniform();
        assign w to r,z;
END
END
END
FOREACH i=1 to I DO
FOREACH u∈U DO
FOREACH d∈D_u DO
FOREACH w∈d DO
z~Eq(22);
r~Eq(23);
END
END
END
END
UPDATE θ,Φ,π;
RETURN θ,Φ,π;
```

5.4　模型复杂度分析

本节分析了 GTTD 模型和文献[13-14]提出的 Geofolk 和 LGTA 模型的算法时间复杂度.之所以选择上述两个模型分析对比,原因在于 GTTD 模型与上述模型的基本假设相同,且模型生成过程类似.不同点在于用户的位置偏好(地域分布)也被视为变量,与地域-话题多项分布和话题-词项多项分布类似,也需要通过 Dirichlet 先验分布产生.

GTTD 模型采样算法的主要部分是在 I 层循环中对每个用户的每篇文档的词项进行变量采样,因此算法的整体时间复杂度为 $O(I\times|\boldsymbol{U}|\times|\boldsymbol{D}|\times(|\boldsymbol{Z}|+|\boldsymbol{R}|))$.

Geofolk 模型采用的吉布斯采样算法与本文 GTTD 模型使用的参数估计方法一致,算法时间复杂度在同一量级.

LGTA 模型采用变分 EM 算法进行参数估计,

在算法的 E-step，地域变量的后验分布算法时间复杂度为 $O(I_1(KN|\boldsymbol{V}|+N|W|))$；在算法的 M-step，算法时间复杂度为 $O(N|\boldsymbol{D}|+I_2KN|\boldsymbol{V}|)$. 从而，算法总体的时间复杂度为 $O(I_1(KN|\boldsymbol{V}|+N|W|+N|\boldsymbol{D}|+I_2KN|\boldsymbol{V}|))$，其中，$K$ 表示待发现的话题数量，N 表示地域数量，W 表示语料中所有词项出现的总次数，$\boldsymbol{V}$ 表示词表，I_1 为 EM 算法的迭代次数，I_2 是求话题和词项后验分布达到收敛所需要的迭代次数. 可见，本模型算法的时间复杂度相比 LGTA 模型稍低.

6 实验设计与结果分析

为了说明 GTTD 模型能够有效地发现在线社交网络中的地域性话题，本文共设置两组话题发现实验以及一组对比实验. 针对话题发现实验，本文分两个方面说明模型的效果：一方面，通过词项的聚类结果表明 GTTD 模型能够发现 Tweet 文本的话题；另一方面，从空间角度说明 GTTD 模型能抽取地域与话题间的关联. 针对模型性能的对比实验，本文选取文献[13]和[14]中的 Geofolk 和 LGTA 模型作为比较对象，以困惑度作为评价指标，说明 GTTD 模型在地域性话题发现的性能上具有优势. 实验数据采用第 3 节抓取的 USA、NYC、SFO 和 Basketball 数据集，实验分别是：

(1) GTTD 模型在 USA 数据集上的实验；

(2) GTTD 模型在 Basketball 数据集上的实验；

(3) GTTD、Geofolk 和 LGTA 模型在 NYC、SFO 和 Basketball 数据集上的对比实验.

6.1 评价指标

困惑度(perplexity)是评价主题模型性能最常用的指标，因此本文选取困惑度作为评价指标. 具体到话题发现模型，常使用训练集构造模型，然后计算模型在测试数据集上的困惑度以评估模型的效果. 在 GTTD 模型中，数据集中的一篇文档 d 包括文本 w_d 和 l_d，困惑度的计算方法如式(13)：

$$perplexity_{text,location}(\boldsymbol{D}_{test})=\exp\left\{-\frac{\sum_{d\in\boldsymbol{D}_{test}}\log P(w_d,l_d)}{\sum_{d\in\boldsymbol{D}_{test}}N_d}\right\} \tag{13}$$

其中，

$$P(w_d,l_d)=\sum_{u\in\boldsymbol{U}}P(u|\tau)\times\sum_{r\in\boldsymbol{R}}P(r|\boldsymbol{\pi}_v)P(l|r,\sigma)\times\sum_{z\in\boldsymbol{Z}}P(z|\boldsymbol{\theta}_{u,r})P(w|\boldsymbol{\Phi}_{z,r}) \tag{14}$$

在实验中，吉布斯采样的迭代次数 *Burn-In Steps* 难以通过经验确定，因此在实验前先进行一次参数测准实验，计算吉布斯采样每一轮迭代时模型的困惑度，图 5 展示了模型运行在 NYC 数据集中困惑度随迭代次数的变化情况，由图可知，当迭代次数处于 870～900 之间时，困惑度逐步下降，之后则保持相对平稳，可以认为采样算法在此期间收敛. 通过该实验即可以确定吉布斯采样的迭代次数.

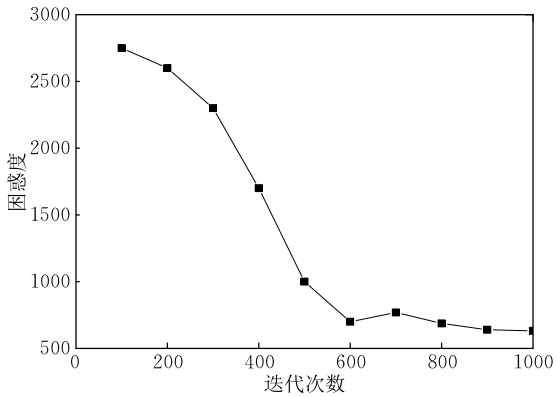

图 5 困惑度随迭代次数的变化趋势

表 4 展示了 GTTD 模型输入参数的设置情况. 文献[21]指出，Dirichlet 先验分布的超参数选取很大程度上决定了模型区分话题的性能. α、β 的值本质上决定了发现话题的粒度，较大的 β 值对应了相对较粗的话题粒度，即倾向于发现较少数目的话题，而 α 的值最好取 $50/|\boldsymbol{Z}|$. 此外，根据文献[22]的结论，话题个数应设置为 5 的倍数. 本文依据文献[22]的超参数设置方法，取 β 和 η 的值均为 0.1，α 设为 $50/|\boldsymbol{Z}|$，话题个数 $|\boldsymbol{Z}|$ 的设定往往依赖于经验值，需要根据实验目的调整. σ 为混合高斯分布的标准差，其与 *mergeDist* 的值决定了地域的大小，根据实验目的调整.

表 4 实验参数设置

数据集	σ	*mergeDist*	$\lvert\boldsymbol{Z}\rvert$	$\lvert\boldsymbol{R}\rvert$	*Burn-In*
USA	0.033	0.066	50	1600	890
Basketball[1]	0.033	0.066	30	50	840
NYC	0.0033	0.0165	40	195	880
SFO	0.0033	0.0165	40	125	850
Basketball[2]	0.0033	0.0165	10	50	850

6.2 GTTD 在 USA 数据集上的实验

USA 数据集是全美范围内的采样，地域范围广、内部话题分布混杂，在该数据集上进行实验能够检验模型的综合效果. 模型的输出参数 $\boldsymbol{\Phi}_{r,z}=P(w|r,z)$，$w\in\boldsymbol{V}$，$\boldsymbol{\Phi}_{r,z}$ 表示词项 w 在地域话题对 $\langle r,z\rangle$ 上的概率分布，是一个 $\boldsymbol{R}\times\boldsymbol{Z}$ 维的矩阵，根据其边缘概率分

布可求得特定话题的词项分布，词项的概率大小代表了该词项与该话题的相关程度，计算公式如式(15)：

$$P(w|z)=\sum_{r\in \boldsymbol{R}}P(w|r,z) \tag{15}$$

实验结果表明大部分话题具有较为鲜明的含义，容易进行汇总和分析. 在表 5 中展示了 6 个典型话题，每个话题给出分布概率最大的 10 个词语.

表 5　GTTD 模型发现到的典型话题

Topic-1		Topic-2		Topic-3		Topic-4		Topic-5		Topic-6	
词	概率	词	概率	词	概率	词	概率	词	概率	词	概率
hot	0.071360	temperature	0.054630	market	0.046746	cake	0.094101	suck	0.046327	play	0.070069
captain	0.042259	days	0.097239	trading	0.078350	sauce	0.069701	happy	0.075203	paul	0.069328
game	0.085984	hpa	0.009084	stock	0.062781	meat	0.080464	people	0.017562	injury	0.096244
goal	0.016088	humid	0.031899	delists	0.014519	foot	0.063608	hope	0.055807	spurs	0.055784
team	0.057532	time	0.053282	sell	0.064764	rare	0.093430	swear	0.077152	game	0.017881
full	0.060971	wind	0.034099	limit	0.027116	icing	0.093999	man	0.082501	chrispaul	0.087471
tackle	0.057763	sun	0.032456	today	0.028983	like	0.067687	feel	0.051094	clippers	0.053914
coach	0.022306	rain	0.009108	deal	0.085939	medium	0.068587	low	0.085173	nba	0.087883
man	0.078193	meter	0.092928	balance	0.036080	pork	0.084947	much	0.066130	playoff	0.068554
kick	0.092244	wave	0.003286	bid	0.036285	cooking	0.049129	waste	0.051538	jump	0.004097

从表 5 描述的典型话题词项分布情况(纵向排列为同一话题下的单词及其出现概率)分析，这些话题应该与运动、天气、食物、情绪等有关. 这表明，虽然 Twitter 平台的文本数据零碎且有较大噪音，但是基于主题概率思想的 GTTD 模型依然具有较好的话题发现能力.

模型的输出参数 $\boldsymbol{\theta}_{u,r}=P(z|u,r)$，$z\in \boldsymbol{Z}$，表示特定话题在用户地域对$\langle u,r\rangle$上的概率分布，是一个 $\boldsymbol{U}\times\boldsymbol{R}$ 维的矩阵，由其边缘概率分布可求得特定地域中每个话题出现的概率分布. 计算公式如式(16)：

$$P(z|r)=\sum_{u\in \boldsymbol{U}}P(z|u,r) \tag{16}$$

图 6 和图 7 给出了两个地域的话题分布情况，其中地域 A 中话题 10 和话题 14 的概率较高，而地域 B 中话题 7 和话题 19 的概率较高. 这表明，Twitter 平台上不同地域的话题分布存在差异性，体现在：给定一个地域，不同话题在该地域内出现的概率不同，表现为 Twitter 平台用户讨论该话题的频度；反过来，给定一个话题，其在不同地域的流行程度也不

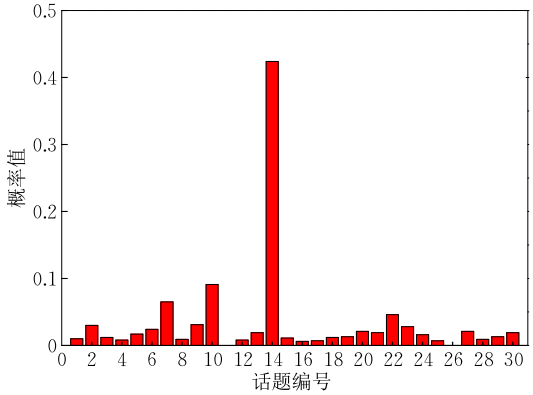

图 6　地域 A 内的话题分布情况

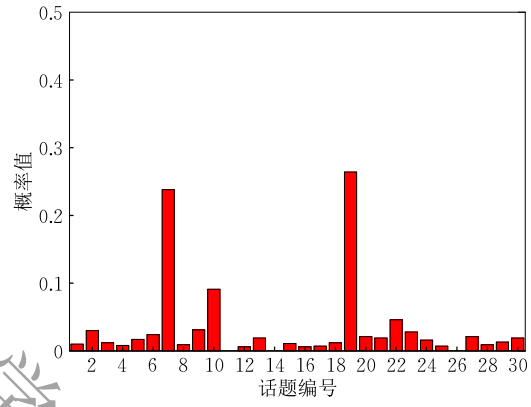

图 7　地域 B 内的话题分布情况

同. 这正符合本文地域性话题的特征，说明本文提出的 GTTD 模型能够将地域作为影响话题产生的因素，并发现地域性话题.

6.3　GTTD 在 Basketball 数据集上的实验

由模型在 USA 数据集上的实验结果可见，话题处在一个较高的层次上，这是由于社交网络中用户群体的数量很大，群体的共性较明显，个性化信息的提取相对困难. 然而，这种粗粒度的话题发现不能满足实际舆情感知应用的需求. 本文对特定领域的话题进行更细致的分析，既可以缩小文本语料的地域分布范围，又能选取特定内容的信息.

本实验在 Basketball 数据集上进行. 文献[21]指出，话题个数能够影响待发现话题的粒度，并且话题数目的确定有赖于实验语料和实验目的，即一定程度上取决于先验知识. 考虑到获取本实验 basketball 数据集的过滤条件，以及现实生活中美国职业篮球联赛(NBA)有 30 只球队的现状，出于验证数据集中话题存在地域性分布考虑，我们不妨设定本实验中的话题个数为 30. 实验结果中各话题

基本与 NBA 的球队或者球星相关，其中关于圣安东尼奥马刺队的话题较为明显，选取与“马刺队”话题相关概率最大的前 300 个文档，调用 Google MapsAPI① 以热力图的形式在地图上展示出来，如图 8 所示. 可见在德克萨斯州这个话题的热度最高，在美国其他人口较为密集的地方也有人关注. 具体到圣安东尼奥市，文档则分布在马刺队的主场 AT&T 中心体育场附近，如图 9 所示. 根据 NBA 的实际赛程，2015 年 3 月 5 日马刺队主场迎战萨克拉门拖国王队，并以 27 分的领先取得胜利，社交网络上关于马刺大胜的讨论较多. 模型输出的结果恰好印证了这种话题的地域性分布. 需要指出的是，虽然本实验中设定话题个数为 30，符合 NBA 球队的真实数目，但并不表示本方法能够达到显著区分 30 支 NBA 球队的效果.

图 8 全美范围内“马刺队”话题的相关文档的分布

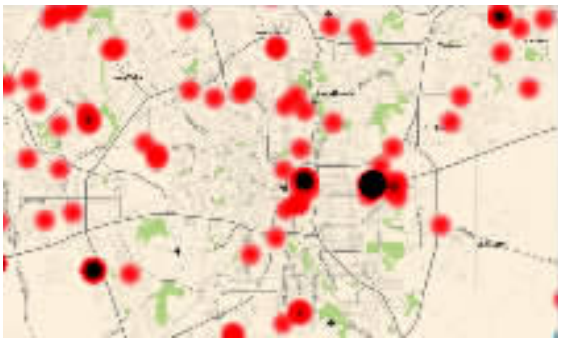

图 9 圣安东尼奥市“马刺队”话题的相关文档的分布

6.4 GTTD 与 Geofolk 和 LGTA 模型的对比实验

为了更加全面地验证地域性话题发现模型的性能，将本文提出的 GTTD 模型与文献[13]提出的 Geofolk 模型以及文献[14]中提出的 LGTA 模型就困惑度指标进行对比实验.

对比算法 1：Geofolk. 该算法同样运用文本和地理位置信息进行语义建模，相比于单纯基于文本分析的算法，Geofolk 算法在标签推荐、内容分类方面具有较好表现. 然而，Geofolk 算法认为每个话题仅有一个地理信息，这导致模型无法描述话题在多地域上的分布情况.

对比算法 2：LGTA. 该算法从地域聚类和话题建模两方面构建地域性话题发现模型，能够获得同一话题在不同地域的概率分布. 其认为所有地域具有共同的话题集合，而且未考虑用户层面的偏好信息.

3 种模型在 NYC、SFO 和 Basketball 数据集上运行，对于每个数据集随机抽取 80% 的数据对算法进行模型训练，使用剩余的数据进行困惑度计算. 为了避免数据集划分带来的误差，实验进行了十折交叉验证. 图 10 展示了实验结果，其中 Y 轴是交叉验证得到的平均困惑度.

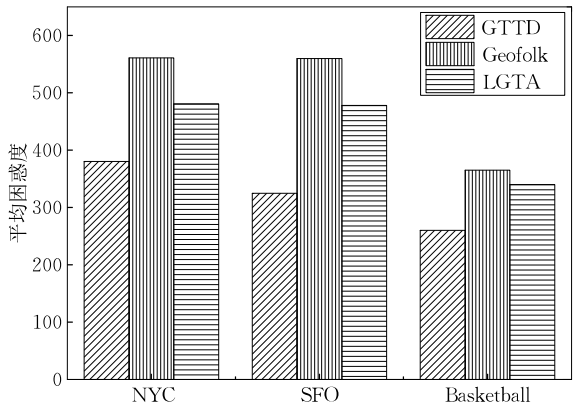

图 10 各数据集上 3 种模型的困惑度结果对比

根据实验结果可知，在 Basketball 数据集上，3 种算法的困惑度明显低于在 NYC 和 SFO 数据集上的运行结果. 这是由于 Basketball 数据集是在 USA 数据集上经过数据筛选得到的，数据内部的话题分类较为明晰. 可以注意到，GTTD 模型在 3 个数据集上的效果比 Geofolk、LGTA 都有较大提升. 与 Geofolk 相比，GTTD 模型将地域和话题的概念区分开来，能够发现分布在多个地域上的普遍性话题，另外 Geofolk 算法认为话题的地域分布服从高斯分布，也导致其话题发现能力受到局限. 与 LGTA 算法相比，GTTD 算法引入用户层面的个性化信息，其发现个性化话题的能力更强.

最后需要指出，本文提出的 GTTD 模型具有良好的扩展性. 虽然针对 Twitter 在线社交平台，但只要社交媒体数据符合本文第 5.1 节定义的文档形式 $d=\{u,w_d,l_d\}$，均可借助 GTTD 模型进行社交平台地域性话题发现工作. 符合要求的社交平台包括 Foursquare、Flickr 等.

① https://developers.google.com/maps/

7　总结与展望

针对传统舆情感知技术较少考虑到地域相关因素的不足，本文结合在线社交网络中丰富的位置信息提出一种基于主题模型的地域性话题发现模型.首先，本文基于用户、位置和话题间的复杂关联详尽分析了社交网络的地域性和话题性特征；其次，根据主题模型思想构建了 GTTD 模型，将各种因素引入话题发现，使模型同时具备语义和空间描述能力；最后，本文使用从 Twitter 平台上获取的真实数据对地域性话题发现模型 GTTD 进行分析与评价.实验结果表明，GTTD 模型能有效地发现所抓取 Twitter 数据中的地域性话题.此外，针对其它含有地域位置维度的在线社交网络，如 Foursquare 和 Flickr 等，GTTD 模型依然适用.本文开展的研究工作尚可结合社交网络中话题追踪任务深入进行，如引入时间因素对话题库动态维护等.此外，本文提出的模型还能结合用户位置预测等问题继续研究.

致　谢　审稿专家和编辑在论文投稿及审稿过程中提出了宝贵的意见和建议，在此表示感谢！

参　考　文　献

[1] 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

[2] Fang Y, Zhang H, Ye Y, et al. Detecting hot topics from Twitter: A multiview approach. Journal of Information Science, 2014, 40(5): 578-593

[3] Song Z, Martin E, et al. Discovering more meaningful regions: A regularized geographical topic model//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 231-240

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

[5] Hofmann T. Probabilistic latent semantic indexing//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, USA, 1999: 50-57

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

[7] Hong L, Davison B D. Empirical study of topic modeling in Twitter//Proceedings of the 1st Workshop on Social Media Analytics. Washington, USA, 2010: 80-88

[8] Huang B, Yang Y, Mahmood A, et al. Microblog topic detection based on LDA model and single-pass clustering//Yao J T, Yang Y, Slowinski R, et al, eds. Rough Sets and Current Trends in Computing. Berlin Heidelberg, Germany: Springer, 2012: 166-171

[9] AlSumait L, Barbará D, Domeniconi C. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking//Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy, 2008: 3-12

[10] Vosecky J, Jiang D, et al. Dynamic multi-faceted topic discovery in Twitter//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. San Francisco, USA, 2013: 879-884

[11] Guo X, Xiang Y, et al. LDA-based online topic detection using tensor factorization. Journal of Information Science, 2013, 39(4): 1-11

[12] Eisenstein J, et al. A latent variable model for geographic lexical variation//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts, USA, 2010: 1277-1287

[13] Sizov S. Geofolk: Latent spatial semantics in web 2.0 social media//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. Cambridge, UK, 2010: 281-290

[14] Yin Z, Cao L, Han J, et al. Geographical topic discovery and comparison//Proceedings of the 20th International Conference on World Wide Web. Lyon, France, 2011: 247-256

[15] Yuan Q, Cong G, et al. Who, where, when and what: Discover spatio-temporal topics for Twitter users//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 605-613

[16] Liu Y, Ester M, Hu B, et al. Spatio-temporal topic models for check-in data//Proceedings of the 15th IEEE International Conference on Data Mining. Atlantic City, USA, 2015: 889-894

[17] Zhang L, Sun X, Zhuge H. Location-driven geographical topic discovery//Proceedings of the 9th IEEE International Conference on Semantics, Knowledge and Grids. Beijing, China, 2013: 210-213

[18] Hu B, Jamali M, Ester M. Spatio-temporal topic modeling in mobile social media for location recommendation//Proceedings of the 13th IEEE International Conference on Data Mining. Dallas, USA, 2013: 1073-1078

[19] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford University, USA, 2009

[20] Böhm, Christian, et al. Querying objects modeled by arbitrary probability distributions//Proceedings of the 10th International Symposium on Spatial and Temporal Databases. Boston, USA, 2007: 294-311

[21] Griffiths T L, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences, 2004, 101(Supplement 1): 5228-5235

[22] Yang B, Manandhar S. Community discovery using social links and author-basedsentiment topics//Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Beijing, China, 2014: 580-587

CAO Jiu-Xin, born in 1967, Ph. D., professor, Ph. D. supervisor. His research interests include complex network, social network, mobile Internet and cloud resource management and security.

XU Shuai, born in 1991, Ph. D. candidate. His research interest is social computing.

CHEN Gao-Jun, born in 1988, M. S. His research interest is social computing.

ZHAO Li-Yang, born in 1990, M. S. candidate. His research interest is social computing.

ZHOU Tao, born in 1989, Ph. D. candidate. His research interest is social computing.

LIU Bo, born in 1975, Ph. D., associate professor. Her research interests include pervasive computing and social computing.

Background

As the most active communication platform, online social network has influenced various aspects of the society with its revolutionary power, and drawn great attention to the social media phenomenon. The highly active and self-organizing mode of information diffusion in online social networks has made traditional opinion monitoring techniques limited, thus analysis on personal information combining with new user attributes has become a new research direction. In this paper, in order to detect geographical topics in Twitter platform, we firstly analyzed the semantic and spatial properties of users in online social networks based on the relationship between user, region and topics is further explored; secondly, the concept of geographical topic is proposed and then a probabilistic model is constructed on the basis of topic modeling. Experiments on Twitter datasets indicate that the proposed model is capable of detecting geographical topics effectively meanwhile showing better performance in the criteria of perplexity than other models.

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61272531, 61202449, 61272054, 61370207, 61370208, 61300024, 61320106007 and 61472081), the National High Technology Research and Development Program (863 Program) of China (2013AA013503), the Jiangsu Technology Planning Program (SBY2014021039-10), the Jiangsu Provincial Key Laboratory of Network and Information Security under Grant No. BM2003201 and the Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grant No. 93k-9.