软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(4):1189−1211 [doi: 10.13328/j.cnki.jos.005616]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种空间上下文感知的提及目标推荐方法*

汤小月[1], 周 康[1], 王 凯[2]

[1](武汉轻工大学 数学与计算机学院,湖北 武汉 430023)

[2](武汉大学 计算机学院,湖北 武汉 430072)

通讯作者: 汤小月, E-mail: sharontang@whu.edu.cn

摘 要: 作为一种新兴的社交媒体用户交互服务,提及机制(mention mechanism)正在用户在线交互和网络信息传播方面扮演着重要角色.对用户提及行为的研究能够揭示用户的隐式偏好与其显式行为之间的联系,为信息传播监控、商业智能、个性化推荐等应用提供新的数据支撑.当前,对用户提及机制的探索多集中在其信息传播属性上,缺少从普通用户角度对其用户交互属性的学习.通过对普通用户提及行为的分析和建模构建一个推荐系统,为给定的社交媒体消息生成目标用户推荐.通过对大型真实社交媒体数据集的分析发现,用户的提及行为受其提及活动的语义和空间上下文因素的联合影响.据此,提出一个联合概率生成模型 JUMBM(joint user mention behavior model),模拟用户空间关联提及活动的生成过程.通过对用户语义和空间上下文感知的提及行为进行统一建模,JUMBM 能够同时发掘用户的移动模式、地理区域依赖的语义兴趣及其对应目标用户的地理聚集模式.此外,提出一种混合剪枝算法,加快推荐系统对在线 top-k 查询的响应速度.在大型真实数据集上的实验结果表明,所提方法在推荐有效性和推荐效率方面均优于对比方法.

关键词: 用户提及行为建模;目标用户推荐;空间上下文感知;综合概率模型;社交网络分析

中图法分类号: TP311

中文引用格式: 汤小月,周康,王凯.一种空间上下文感知的提及目标推荐方法.软件学报,2020,31(4):1189−1211. http://www.jos.org.cn/1000-9825/5616.htm

英文引用格式: Tang XY, Zhou K, Wang K. Spatial context-aware mention target recommendation method. Ruan Jian Xue Bao/Journal of Software, 2020,31(4):1189−1211 (in Chinese). http://www.jos.org.cn/1000-9825/5616.htm

Spatial Context-aware Mention Target Recommendation Method

TANG Xiao-Yue[1], ZHOU Kang[1], WANG Kai[2]

[1](School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan 430023, China)

[2](School of Computer Science, Wuhan University, Wuhan 430072, China)

Abstract: As a newly emerging social media user interactive service, mention mechanism is playing an important role in both information sharing and online social interacting. Researches on mention mechanism can provide us valuable resources to reveal the correlation between users' latent preferences and their explicit interacting behaviors and can be constructed as the data foundation for many applications such as information dissemination monitoring, business intelligence, and personalized recommendation. However, most of the previous works focused on the information diffusion aspect, lacking the in-depth study on its interaction attribute from the common users' perspective. This study aims to construct a recommendation system to automatically recommend target users for given social media posts based on the analysis and modeling of common users' mention behaviors. This study first analyzes two large-scale real-world datasets to explore the mention mechanism from the aspect of users' interactions and finds that, users' mention behaviors are impacted by

* 基金项目: 国家自然科学基金(61502362, 61401319, 61179032); 湖北省自然科学基金(2015CFA061, 2019CFB250)

Foundation item: National Natural Science Foundation of China (61502362, 61401319, 61179032); Natural Science Foundation of Hubei Province of China (2015CFA061, 2019CFB250)

收稿时间: 2018-02-11; 修改时间: 2018-04-29; 采用时间: 2018-06-10; jos 在线出版时间: 2019-11-06

CNKI 网络优先出版: 2019-11-06 11:48:52, http://kns.cnki.net/kcms/detail/11.2560.TP.20191106.1148.001.html

both the semantic and the spatial context of their mention activities. Secondly, based on a unified definition of the joint semantic and spatial context-aware mention behavior, a joint latent probabilistic generative model named JUMBM (joint user mention behavior model) is built to simulate the generating process of users' mention activities. Specially, JUMBM is able to simultaneously capture users' movement patterns, geographical area-dependent semantic interests, and the geographical clustering patterns of the targets users. Besides, a hybrid pruning algorithm is proposed to achieve a fast high-dimensional retrieval and facilitate the online top-*k* query answering. Extensive experiments on real-world datasets demonstrate the significant superiority of the proposed approach over the baseline methods to make more effective and efficient recommendations.

Key words: user mention behavior modeling; target user recommendation; spatial context-aware; joint probabilistic model; social networks analysis

随着社交网络服务的兴起,越来越多的人选择使用在线社交媒体系统分享信息.凭借着在内容生成方式、用户参与的广泛性与即时性、信息扩散模式与速度等方面的优势,社交媒体用户量在近些年呈现出爆发式的增长,每天都有海量的消息被用户发布.除了发布消息,用户也会采用不同的交互行为与其他用户进行在线交互.以 Twitter(https://twitter.com/)为例,用户可以"回复"和"转发"消息;为消息添加"标签";以"@用户名"的方式在消息中"提及"其他用户;或者"关注""订阅"感兴趣的人等等.这其中,得益于在缓解信息过载方面的优势[1,2],社交媒体中的"提及"机制(mention mechanism)正在成为用户在线交互和网络信息传播的主要功能载体.对这一独特的用户在线交互行为的分析和研究能够揭示用户的隐式偏好与其显式交互行为之间的联系,为信息传播监控[3,4]、商业智能[5−8]、个性化推荐[9]等应用提供新的数据支撑.

当前,社交媒体系统的活跃用户规模日渐庞大.2016 年末的统计显示(http://about.twitter.com/company/),Twitter 的月活跃用户量已超过 3 亿且平均每位用户拥有 208 位直接社交朋友.因此,当用户(即当前用户)需要在消息中提及其他用户(即目标用户)时,庞大的候选目标群会使得当前用户很难在短时间内找到其需要的目标用户.此时,为当前用户提供一个短小的、包含其最有可能提及的目标用户组成的列表,可以有效地帮助用户识别目标、节省搜索时间.目前,主流的社交媒体服务,如 Twitter、新浪微博(https://weibo.com/)和 Facebook(https://facebook.com/)等,会在用户输入提及符号"@"时,根据用户的部分输入内容和提及历史为其生成一个候选目标用户列表.然而,由于没有准确挖掘用户的提及行为偏好,其建议内容往往不符合用户的实际期望.例如,图 1 展示了新浪微博的目标用户建议的一个例子.可以看到,尽管输入内容中已明确指向了目标用户"2018 年俄罗斯世界杯",但其用户名仍未出现在微博给出的 top-10 推荐列表中.

Fig.1 An example of target user suggestion given by Sina Weibo

图 1 一个新浪微博目标用户建议的例子

当前,对用户提及行为数据的分析和使用吸引了来自人工智能及数据挖掘等不同领域研究者的广泛关注[2−4,8].然而,大多数当前研究仅聚焦于提及机制的信息传播属性,如寻找能够将消息更快散播的目标用户.事实上,社交媒体中的提及功能既是一个网络信息传播渠道,也是一个普通用户交互工具.直观上看,相对于少数的媒体工作者和广告商,提及功能的用户交互属性对占大多数的普通用户来说更为重要.从普通用户角度来看,

提及功能更广义的作用应是通知目标用户“查看”而不是“传播”某条消息[9].此外,已有研究表明,社交媒体中信息传播更多地依赖于用户的转发而非提及行为,寻找合适的转发者可以更加有效地提高消息在社交媒体中的扩散速度和深度[10−13].因此在本文中,我们通过对影响普通用户提及行为的因素进行分析以挖掘隐藏的用户行为偏好,并基于学习到的知识模型构建一个推荐系统,帮助当前用户搜索和识别目标用户.

尽管文本内容信息对于目标用户推荐的重要性已得到许多研究的证明[3,4,8,9],但尚未有工作调查其他影响用户目标选择的上下文因素.近年来,得益于可定位移动设备的普及,地理维度的信息正在社交媒体中迅速扩散,这为我们更好地理解用户的在线行为与其物理活动之间的关系提供了宝贵的资源.换句话说,用户物理维度的信息可以帮助我们更准确地捕捉用户的在线行为偏好[5,7].因此在本文中,我们探索用户提及活动的空间上下文信息对其目标用户选择的影响.通过对两个大型真实社交媒体数据集的分析(第 2.1 节)我们发现,被同一用户提及的目标用户呈现出地理聚集的趋势.这揭示了当前用户和目标用户之间的空间关联,激发了对空间上下文感知的用户提及行为建模和目标用户推荐的研究需求.

在本文中,我们从普通用户的角度研究社交媒体中的目标用户推荐问题.即,当用户需要在一条消息中提及其他用户时,本文研究如何找到最有可能被其提及的目标用户并生成推荐.具体而言,本文通过分析用户空间上下文感知的在线提及行为来研究该问题.我们提出一个联合概率生成模型 JUMBM(joint user mention behavior model),通过综合考虑语义和空间因素来模拟用户提及活动的生成(如图 3 所示).由于用户的提及行为受其语义和空间上下文因素的联合影响,JUMBM 引入了两个关键的潜在主题变量——语义主题和地理区域,分别负责生成用户活动的语义属性(如文本词语)和地理属性(如地理坐标).此外,不同于当前研究假设用户语义兴趣是固定不变的,JUMBM 利用目标用户的地理聚集区域来发掘当前用户空间依赖的语义兴趣.通过这种方式,JUMBM 能够统一地进行语义和空间感知的用户提及行为建模,从而发掘用户的移动模式、地理区域依赖的语义兴趣及其目标用户的地理聚集模式.此外,为了应对推荐中遇到的“维数灾难”问题并加快系统对在线查询的响应速度,本文提出一种混合剪枝算法对项目和属性空间进行综合剪枝,实现了高维大候选空间内的快速精确项目检索.

本文的主要贡献如下.

(1) 从普通用户角度对用户的提及行为进行了学习,并提出了一个概率生成模型来发掘语义和空间上下文因素对用户提及行为的联合影响.据我们所知,本文是第一个从普通用户角度进行空间感知的用户提及行为建模的工作.

(2) 设计了一种高效的混合剪枝算法,通过对项目和属性空间进行同时剪枝,实现了对在线查询的快速响应.

(3) 在两个大型真实社交媒体数据集上构建了一系列的实验.实验结果证明了在解决提及目标推荐问题时考虑用户空间上下文因素的必要性.同时,本文提出的方法在推荐有效性和效率方面均优于其他目标用户推荐方法.

本文第 1 节对相关工作的研究现状进行总结.第 2 节首先对本文使用的符号和研究问题做出形式化定义,然后对模型背后的研究动机进行描述,最后详细阐述 JUMBM 的模型结构和模型推理过程.第 3 节介绍如何基于学习到的知识进行高效率的目标用户推荐.本文在真实数据集上进行实验,在第 4 节中对提出的方法进行评估.最后,在第 5 节中对本文工作做出总结.

1 相关工作

本文的研究内容主要涉及到 3 个主题:(1) 社交网络用户交互行为分析;(2) 空间上下文感知的主题模型;(3) 社交媒体目标用户推荐.本文将分别总结相关工作的研究现状,并对其与本文方法的区别进行阐述.

1.1 社交网络用户交互行为分析

对用户在线交互模式的分析能够揭示用户的隐藏意图和真实行为之间的相关性.这种高层次的信息可以用于诸如内容传播分析、信息监控以及个性化推荐等等一系列应用.近年来,越来越多的研究者开始关注社交网络用户的在线交互行为.例如,Xu 等人的研究[14]关注于用户的在线消息“发布(posting)”行为.他们发现用户的

“发布”行为主要受到网络突发消息、个人社交邻居的历史记录和用户本身兴趣的影响.Qiu 等人[15]调查了用户与其生成文本间的交互行为,并提出一个概率模型发掘用户的主题兴趣和交互模式.Yin 等人[16]则关注社交媒体用户对项目的评分行为,其研究发现,用户的评分行为受到个人兴趣和公众集体倾向的共同影响.由于社交媒体上的用户“转发”行为往往指示了网络信息的流向,近年来对用户“转发”行为的研究也在不断涌现[10,17−24].如 Chen 等人[10]通过分析 Twitter 用户的转发模式构建了一个个性化的 Tweet 推荐系统.Bi 等人的研究[17]则聚焦于社交媒体中信息转发网络,他们发现,考虑用户间的转发关系能够使得基于文本的模型更准确地推理用户个人兴趣.除此之外,当前研究也关注于其他类型的用户交互行为,如“标签”[21−23]和“提及”[2,4,8,24].与这些研究相比,本文基于不同的研究目标,构建社交媒体用户的提及行为模型,通过挖掘用户的行为偏好为目标用户推荐提供知识模型.

1.2 空间上下文感知的主题模型

在过去的几十年里,主题模型(topic model)在文本挖掘、信息检索及自然语言处理领域得到了广泛的应用.主题模型不但可以发掘文本集合中隐藏的语义结构,还能有效地分析多种类型的离散数据.当前,主题模型已发展出了很多变体,如概率潜在语义分析(PLSA)、潜在狄利克雷分配(LDA)、层次狄利克雷过程(HDP)等等.当前,一些研究开始通过分析用户、位置和主题之间的关系来挖掘社交媒体中的地理知识[5,12,25−37].例如,Kurashima 等人[31]将主题模型和 Markov 模型进行了结合,根据用户上传照片的位置标签来推理一个摄影师访问该位置的概率.Hu 等人[5]则综合考虑了社交媒体消息的空间和文本内容,提出 ST 模型对用户移动模式、用户兴趣和地理位置的语义功能间的联系进行推理.Yin 等人[25]提出了一个位置-内容感知的概率主题模型,在空间项目推荐过程中对本地偏好和项目语义模式进行了量化.由 Wang 等人[29]提出的 Geo-SAGE 模型根据地理项目的内容和共现模式,将用户在特定区域内的个人兴趣和公众在区域的本地偏好进行了综合,提高了地理项目推荐精确度.Fang 等人[30]在其提出的时空上下文感知的推荐框架 STCAPLRS 中,对用户个人兴趣、本地偏好及时空上下文等因素进行了集成,为特定地理区域内的用户提供位置推荐.与上述工作相比,本文的主要工作是通过综合考虑用户提及交互行为活动的语义和上下文信息构建一个概率主题模型,对用户的提及行为偏好进行推理.

1.3 社交媒体目标用户推荐

当前,目标用户推荐正在成为社交媒体用户个性化推荐领域的一个研究热点.举例来说,Wang 等人[4]将这个任务视为一个学习-排序问题,通过抽取用户兴趣、内容依赖的用户关系以及用户影响力等特征来训练一个学习排序方法.他们的主要研究目的是通过寻找合适的目标用户来加快一条 Tweet 在 Twitter 中的传播.Zhou 等人[2]则从信息过载的角度研究了目标用户的排序问题.通过探索文本内容特征和用户个人吸引力、社区权威和对话内容等因素来构建一个排序模型来寻找合适的目标用户以减缓社交媒体信息过载问题.由 Tang 等人[8]开发的 CAR 推理框架在提取内容、社交、位置和时间这 4 种类型特征的基础上,采用支持向量机(SVM)模型来实现目标推荐.该工作关注的问题是如何在用户发表促销相关的消息时为其推荐合适的目标受众,从而为社交媒体中的市场工作者寻找具有高回应率的用户群.与之不同,本文从普通用户的在线交互角度,研究如何找到当前用户最有可能提及的目标用户.此外,CAR 使用了学习-排序方法来形式化并解决问题.而本文则基于概率生成方法来挖掘用户提及偏好并为目标推荐提供知识模型.最近,Gong 等人[9]提出了一个同时考虑了当前内容及目标用户的内容历史的主题翻译模型 A-UUTTM 以对用户行为偏好进行推理.该方法能够有效地识别目标用户名,是目前用于目标用户推荐效果最好的方法.但是,A-UUTTM 仅考虑了文本内容驱动的用户语义模式,忽略了空间因素对用户提及行为的影响.与之相比,本文提出的 JUMBM 模型则对用户地理区域依赖的语义兴趣、移动模式及目标用户的地理聚集模式进行了统一的建模和推理.第 4.4.1 节的实验结果则表明了在解决提及目标推荐问题时考虑用户空间上下文因素的必要性和 JUMBM 模型在挖掘用户提及偏好方面的有效性.

本文提出的方法与已有工作的区别总结如下.首先,本文旨在从普通用户的角度探索目标用户推荐问题,即寻找最有可能被当前用户通知“查看”当前消息的目标用户.也就是说,本文关注的社交媒体提及机制不再局限于其对某条消息传播速度的影响,而是其广义上的用户交互属性.其次,本文不仅基于消息内容,同时结合用户

提及活动的空间上下文信息对用户的提及行为进行空间感知的联合建模.实验结果(第 5.3.1 节)表明,增加对用户提及活动空间上下文信息的考虑能够显著提升推荐系统的性能.第三,本文尝试提供一个可部署的快速目标用户推荐系统,能够快速响应在线查询.为解决该推荐效率问题,本文设计了一种混合剪枝算法对在线检索空间进行剪枝.实验结果(第 5.3.3 节)表明了该剪枝算法的有效性.

2 空间上下文感知用户提及行为建模

本节首先对本文使用的符号和研究问题给出形式化定义,然后对本文对用户提及行为建模过程背后的研究动机进行描述,最后详细介绍本文提出的 JUMBM 模型结构、生成过程和参数推理过程.

2.1 问题定义

表 1 列举了用于 JUMBM 模型输入数据的符号.本文使用的数据结构定义如下.

定义 1(当前用户和目标用户). 如果用户 u 在至少一条消息中提及了用户 m,则 u 被称为当前用户,m 则称为目标用户.显然,本文中的“当前用户”和“目标用户”是两个相对的概念.一个用户既可以是一个提及实例的当前用户,也可以是另一个提及实例的目标用户.本文中,目标用户拥有两个属性:标识符 m 及其居住位置 l_m.

注意,社交媒体用户可以选择是否在其“个人资料”中公开其居住位置.对于那些提供了准确居住地点描述(例如 GPS 坐标或详细的位置描述)的用户,该位置将直接用作他们的居住位置.而对于那些没有明确给定居住地点或者对其居住位置描述过于粗糙的用户,我们采用一种当前广泛使用的 ground-truth 用户位置获取方法,为每个目标用户分配一个位置点(GPS 坐标)作为其居住位置(参见第 4.1 节).

定义 2(位置标记的用户提及活动). 本文使用一个五元组(u,W_d,l_d,M_d)表示一条位置标记的用户提及活动 d 即,当前用户 u 在发表于地点 l_d 的由词语 W_d 组成的消息中提及了目标用户 M_d.注意,由于用户可以在一条消息中采取多次提及行为,此处的 M_d 表示一个目标用户集合而非单个的目标用户.此外,我们使用 L_{M_d} 表示 M_d 中用户的居住位置集合.

定义 3(用户活动文档). 对于当前用户 u,其活动文档 D_u 是 u 的位置标记提及活动的集合.数据集 D 则由所有用户的提及活动文档组成,即 $D=\{D_u\}_{u=1}^{U}$. L_D 表示所有提及活动文档的发生位置集合,即 $L_D=\{l_d\}_{d=1}^{D}$. L_M 则表示所有目标用户的居住位置集合,即 $L_M=\{L_{M_d}\}_{d=1}^{D}$.

本文的研究问题可定义如下.

问题(目标用户推荐). 给定一个用户活动文档集 D 和一条由用户 u_q 在位置 l_q 处发表的由词语 W_d 组成的查询 q,即 $q=\{u_q,l_q,W_q\}$,本文为 u_q 推荐其最有可能提及的 top-k 个目标用户组成的列表,见表 1.

Table 1 Notations of input data of our model

表 1 模型输入数据符号

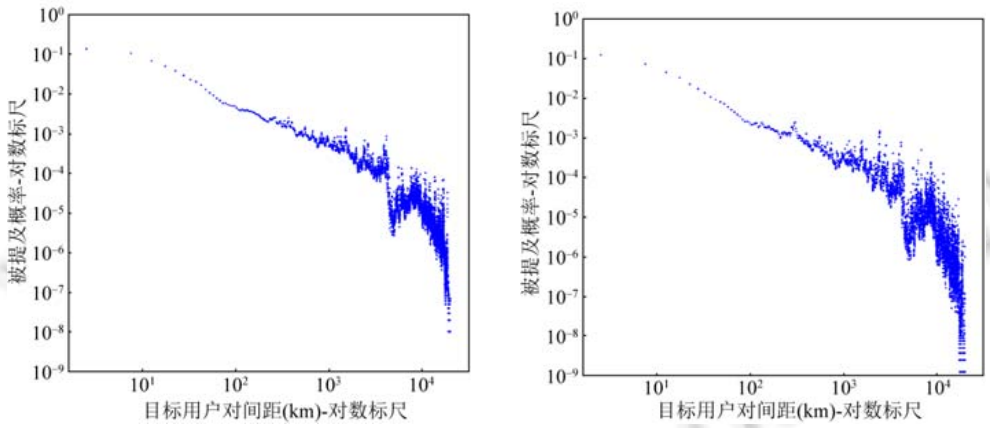
符号	含义
u,m,z,w,g,d	当前用户、目标用户、主题、词语、目标用户聚集区域、一条用户提及活动文档
l_m	目标用户 m 的居住位置
l_d	用户提及活动 d 的发生位置
T,G,W	主题集合、区域集合、唯一词语集合
K,R,N,N_M,V	主题、区域、当前用户、目标用户、唯一词语数量

2.2 JUMBM建模动机

(1) 目标用户地理分布.为了学习空间上下文信息对用户提及行为的影响,我们对采集自新浪微博的由 14 万位用户发布的 59 万条消息,以及采集自 Twitter 的由 13 万位用户发布的超过 55 万条消息组成的大型数据集进行了统计分析.数据集中的每一条消息都附加了一个位置标签且含有至少一个提及实例.同时,每一位目标用户均被分配了一个地理坐标(经度/维度)作为其居住位置.统计发现,被同一当前用户提及的目标用户呈现出地理聚集的趋势.具体来说,在 Ye 等人的启发下[38],我们首先计算了所有被同一用户提及的目标用户两两之间的

地理间距,然后计算出每个目标用户对在特定地理间距上被提及的概率并绘制概率-间距直方图,如图 2 所示.我们发现,目标用户对被提及概率呈现出幂律分布的趋势,且大部分目标用户对之间的地理间距都较短.该现象可以用以下理论倾向进行解释.首先,计算社会科学方向的研究[39]已经证实,在线社交关系往往形成于地理间距较短的节点间.McGee 等人的调查[40,41]也表明,用户与其大多数社交媒体朋友间的物理距离都较短.这些研究显示,用户主要与相近物理距离内的其他用户线上交互,且物理间距较短的用户之间的在线交互更加地频繁.这一结论与已有工作的研究结果是一致的[42,43].其次,地理范围聚焦的在线社区的演化加速了物理临近用户的在线聚集[39,44,45].同属一个地理范围聚焦的在线社区中的用户针对同一地理范围内有着相似的兴趣.因此,用户有可能对新的地理聚焦的社区产生兴趣,即使该社区关注的地理范围与用户所在位置距离较远.总的来说,这一地理现象揭示了当前用户和目标用户间的空间关联,激发了对空间上下文感知的社交媒体用户提及行为建模的研究需求.

(2) 用户位置感知的语义兴趣.前期工作已证实,文本内容表征的用户的语义兴趣是其选择提及目标的一个主要依据[2,8,9].因此,用户的历史消息文本对目标用户推荐来说至关重要.当前,对用户交互行为的研究多假设用户的语义兴趣是固定不变的[2,8,10,11].然而,关于社交媒体用户移动模式的研究发现,用户在不同的位置有不同的语义偏好[46−49].也就是说,用户所在的地理区域不同,其发布消息的语义兴趣也不同.举例来说,某用户在其居住城市发布的历史消息显示其语义兴趣偏向于“生活”或者“健身”,而当其到达一个新的城市时,其语义兴趣可能更偏向于“旅游”或者“美食”.自然地,用户语义兴趣的转变也会引起其在线交互行为模式的变化.因此,相对于基于历史内容的用户语义兴趣推理,发掘当前用户在其目标用户地理聚集区域内的语义兴趣,对目标用户推荐来说更加重要.



(a) Twitter (b) 新浪微博

Fig.2 The log-log geographical probability distribution of target user pairs

图 2 目标用户对的地理概率分布的对数直方图

2.3 模型结构

为了研究内容和空间上下文信息对用户提及行为的联合影响,本文提出了一个联合概率模型 JUMBM,通过综合考虑语义和空间因素模拟用户提及活动的生成过程.图 3 给出了 JUMBM 的图模型表示.其中,模型的观察变量,如词语 w 使用阴影圆圈表示,隐藏变量,如主题 z 则使用无阴影圆圈表示.此外,N、V、K、R 分别表示当前用户、词语、主题和区域的数量,$|D_u|$、$|M_d|$、$|W_d|$则分别表示 u 的提及活动文档数量、d 包含的目标用户数量以及 d 包含的词语数量.表 2 介绍了模型使用的参数符号及其含义.

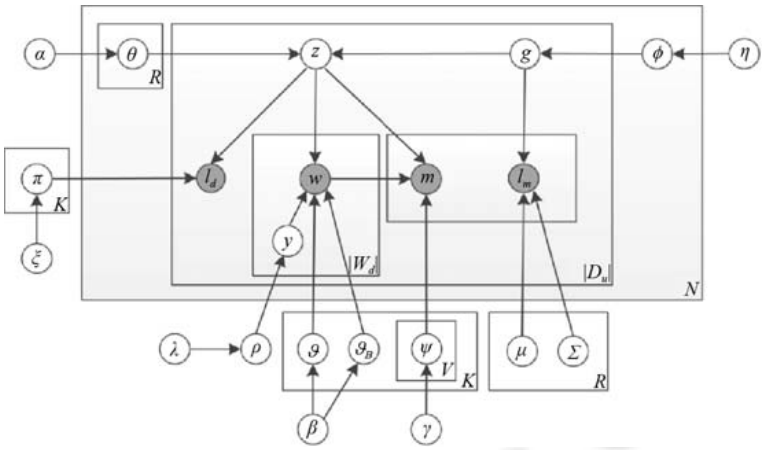


Fig.3 The graphic representation of JUMBM

图 3 JUMBM 的图模型表示

Table 2 Parameter notations in our model

表 2 模型参数符号

符号	含义
ϕ_u	目标用户聚集区域上的多项式分布,代表着 u 的目标用户的聚集区域的地理分布
$\theta_{u,g}$	对于给定的主题 z 和目标用户聚集区域 g,主题上的多项式分布
ϑ_z , ϑ_b	针对主题 z 和备用主题 b 的词语上的多项式分布
$\psi_{z,w}$	对于给定的主题 z 和词语 w,目标用户上的多项式分布
π_z	对于给定的主题 z,当前提及活动的地理位置上的多项式分布
μ_g,$\sum_g$	区域 g 的均值向量和协方差矩阵
y	开关变量,确定从哪个分布生成词语,y=0 或 1
α,γ,η,ξ	多项式分布 $\theta_{u,g}$、$\psi_{z,w}$、ϕ_u 、π_z 的 Dirichlet 先验
β	多项式分布 ϑ_z 和 ϑ_b 的 Dirichlet 先验
ρ	用于生成 y 的二项分布
λ	用于生成ρ的 Beta 先验,定义为λ={λ_1,λ_2}

由于用户的提及行为受到语义和空间上下文因素的联合影响,JUMBMs 引入了两个潜在变量——语义主题和地理区域,分别负责生成用户提及活动的语义属性(如文本词语)和地理属性(如地理坐标).基于这两个潜在变量,JUMBM 以一种统一的方式对用户地理区域依赖的语义兴趣、移动模式以及目标用户的地理聚集模式进行建模.接下来,我们将分别从这 3 个方面对 JUMBM 的模型结构进行阐述.

首先,基于研究动机 1,“目标用户地理聚集”现象表明被同一当前用户提及的目标用户呈现出地理聚集的趋势.该现象揭示了在线交互的用户之间的空间相关性.因此,在用户提及行为模型中融合对目标用户的地理聚集区域的建模是十分必要的.在本文中,我们首先将所有目标用户的居住位置划分进 R 个地理区域.此外,JUMBM 基于连续的地理位置点将一个区域 g 建模为一个地理高斯分布,从而避免了采用多项式分布引起的位置点分布过于稀疏的现象.也就是说,JUMBM 将目标用户 m 的居住地点 l_m 形式化为 $l_m \sim \mathcal{N}(\mu_g,\Sigma_g)$,其中,$\mu_g$ 和 Σ_g 分别表示区域 g 的均值向量和协方差矩阵.即:

$$P\left(l_m \mid \mu_g,\Sigma_g\right)=\frac{1}{2\pi\det\left(\Sigma_g\right)}\exp\left(-\frac{1}{2}\left(l_m-\mu_g\right)^T\sum\nolimits_g^{-1}\left(l_m-\mu_g\right)\right) \tag{1}$$

此外,我们使用一个地理区域上的多项式分布表示 ϕ_u 对 u 所对应的目标用户地理聚集区域分布.

其次,研究表明,用户的语义兴趣对其目标用户选择来说至关重要[2,8,10,11].受当前关于用户兴趣建模研究的启发[18,50,51],本文采用潜在的主题变量来表征用户的语义兴趣.传统的主题模型为文档中的每一个词语分配一个主题,不能有效处理长度较短且高噪声的社交媒体文本数据[9,13,21].因此对于一条提及活动 d,JUMBM 假设 W_d 中的所有词语具有相同的主题 z.此外,根据研究动机 2,JUMBM 基于潜在的主题和区域来揭示用户空间依赖的

语义兴趣.即,JUMBM 根据 D_u 的内容信息和 u 对应的目标用户的地理分布来推理 u 在特定区域 g 内对一组主题的兴趣分布 $\theta_{u,g}$.通过这种方式,我们将主题建模和地理聚集过程进行了统一.此外,在 JUMBM 建模过程中,如何提高主题的质量对于准确推理用户的语义兴趣来说至关重要.与传统的基于 LDA 模型的方法相比,主题翻译模型[9,21,52]结合了主题模型和翻译模型的优点,能够在从短文本和非标准文本数据中发掘隐藏的语义模式的同时,有效地应对社交媒体提及消息中存在的词典缺口问题.我们在建模过程中遵循基本的文本翻译建模机制,即将目标用户的生成过程看作是从内容到目标用户名的翻译过程.因此,JUMBM 中的主题 z 将同时负责生成词语 W_d 和其关联的目标用户 M_d.换句话说,JUMBM 中的每个主题 z 不仅与一个词语分布 ϑ_z 相关联,还与一个目标用户分布 $\psi_{z,w}$ 相关联.此外,JUMBM 引入一个联合潜在因素(主题-区域),按照主题翻译过程生成目标用户名.因此,JUMBM 将每一个主题-词语(z,w)与一个目标用户上的分布 $\psi_{z,w}$ 相关联.

第三,在以新浪微博、Twitter 和 Facebook 为代表的社交媒体服务中,用户可以选择在发表的消息中添加位置或者兴趣点(point of interest,简称 POI)标签来实现签到(check-in)功能.关于社交媒体用户空间移动模式的研究[13,26]表明,用户的签到活动呈现出较强的语义规则性.而在本文研究的问题中,用户位置标记的提及行为可以视为其签到行为在用户在线交互方面的扩展.也就是说,用户位置标记的提及活动中隐藏着与其签到活动类似的用户移动模式.在本文中,我们基于潜在主题来挖掘该用户移动模式.具体来说,主题 z 将负责用户提及活动的位置标签的生成,且每一个主题都与一个标签位置(即当前用户的当前位置)上的分布 π_z 相关联.基于该设计,我们对当前用户语义模式和移动模式的建模过程进行了结合,也使得主题和位置挖掘过程在一个统一的框架下相互影响和提高.

2.4 生成过程

JUMBM 假设用户 u 发出的提及活动 d 的生成过程如下.首先,u 根据其目标用户的地理分布 ϕ_u 选择一个区域 g.然后,u 根据其主题偏好 $\theta_{u,g}$ 选择一个区域 g 上的主题 z.给定主题后,组成文本内容的一系列词语 W_d 由主题-词语分布 ϑ_z 或备用词语分布 ϑ_b 生成.在主题和词语生成结束后,根据主题 z、词语 W_d 及概率分布 $P(.|z,W,\psi_{z,w})$ 选择目标用户 M_d.同时,每个目标用户 $m(m\in M_d)$的居住位置 $l_m(l_m \in L_{M_d})$则根据区域 g 上的地理分布 $\mathcal{N}(\mu_g,\Sigma_g)$ 生成.最后,根据标签位置上的多项式分布 $\pi_{z,a}$ 生成 d 的位置 l_d.模型的生成过程如算法 1 所示.

算法 1. JUMBM 生成过程.

1. 提取ρ~$Beta(\lambda_1,\lambda)$, $\vartheta_b \sim Drichlet(\beta)$;
2. **for each** $u\in$**do**
3. 抽样u对区域的分布 $\phi_u \sim Dirchlet(.|\eta)$;
4. **for each** $g\in G$ **do**
5. 抽样u在区域g上对主题的分布 $\theta_{u,g} \sim Dirchlet(.|\alpha)$;
6. **end for**
7. **end for**
8. **for each** $z\in T$ **do**
9. 抽样一个词语上的分布 $\vartheta_z \sim Dirichelt(.|\beta)$;
10. **for each** $w\in W$ **do**
11. 抽样一个主题和词语上的分布 $\psi_{z,w} \sim Dirichlet(.|\gamma)$;
12. **end for**
13. **end for**
14. **for each** $D_u\in D$ **do**
15. **for each** $d\in D_u$ **do**
16. 抽样一个目标用户聚集区域 $g \sim Multi(.|\phi_u)$;

17. 抽样一个主题 $z_d \sim Multi(.|\theta_{u,g})$;
18. **for each** $w \in W_d$ **do**
19. 抽样一个开关变量 $y_w \sim Bernoulli(\rho)$;
20. **if** y_w=0 **then**
21. 抽样一个词语 $w \sim Multi(.|\vartheta_z)$;
22. **else**
23. 抽样一个词语 $w \sim Multi(.|\vartheta_b)$;
24. **end if**
25. **end for**
26. **for each** $m \in M_d$ **do**
27. 抽样一个目标用户 $m \sim P(.|z_d, W_d, \psi_{z,w})$;
28. 抽样m的居住位置 $l_m \sim \mathcal{N}(\mu_g, \Sigma_g)$;
29. **end for**
30. 抽样u的当前位置 $l_d \sim Multi(.|\pi_z)$;
31. **end for**
32. **end for**

2.5 模型推理

通常来说,对潜变量概率模型进行模型参数推理需要计算使得观察变量的边缘概率最大的参数集,而该边缘分布的计算是与模型中的隐藏变量相关的.但是,由于隐藏变量间的耦合,精确地计算该边缘分布通常是不可实现的.因此,本文采用基于收缩吉布斯抽样[53]的近似学习方法来最大化公式(2)中的联合概率分布.本文假设JUMBM中除β外的所有先验变量均服从对称Dirichlet分布,而β则服从Beta分布.也就是说,模型中的先验变量均为多项式分布或Bernoulli分布的共轭先验.通过这种方式,我们可以在对参数θ、ϕ、π、ψ、μ、Σ、ϑ和ϑ_b进行推理的同时挖掘它们之间的不确定性关联.受前期研究的启发[12,53],我们将模型中的超参数设置成固定值以节省计算成本,即α=50/K、η=50/R、β=γ=ξ=0.01、λ_1=λ_2=0.5.根据算法1中的描述,JUMBM的观察和隐藏变量的联合概率分布可分解为

$$\begin{aligned}&\boldsymbol{P}\left(\boldsymbol{z},\boldsymbol{g},\boldsymbol{W_d},\boldsymbol{M_d},\boldsymbol{y},\boldsymbol{l_d},\boldsymbol{L_{M_d}}\mid\alpha,\lambda,\beta,\gamma,\eta,\xi,\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\\&=\boldsymbol{P}(\boldsymbol{g}\mid\eta)\boldsymbol{P}(\boldsymbol{z}\mid\boldsymbol{g},\alpha)\boldsymbol{P}(\boldsymbol{y}\mid\lambda)\boldsymbol{P}(\boldsymbol{W_d}\mid\boldsymbol{z},\boldsymbol{y},\beta)\boldsymbol{P}(\boldsymbol{M_d}\mid\boldsymbol{z},\boldsymbol{W},\boldsymbol{y},\gamma)\boldsymbol{P}(\boldsymbol{L_{M_d}}\mid\boldsymbol{g},\boldsymbol{\mu},\boldsymbol{\Sigma})\boldsymbol{P}(\boldsymbol{l_d}\mid\boldsymbol{z},\xi)\end{aligned}\tag{2}$$

使用吉布斯抽样方法进行参数推理需要抽样每一条提及活动中潜在变量z、g和y的后验概率分布.因此,我们首先根据条件概率 $P(g_{(u,d)}=x\mid\boldsymbol{g}_{\neg(u,d)},\boldsymbol{z},\boldsymbol{W_d},\boldsymbol{M_d},\boldsymbol{y},\boldsymbol{l_d},\boldsymbol{L_{M_d}},\boldsymbol{u}\cdot)$ 对区域g进行抽样.其中,(u,d)表示用户u的提及活动d.受篇幅所限,下文省略了具体的推理过程.根据贝叶斯规则,潜在变量g的抽样概率可以计算为

$$\begin{aligned}&P\left(g_{(u,d)}=x\mid\boldsymbol{g}_{\neg(u,d)},\boldsymbol{z},\boldsymbol{W_d},\boldsymbol{M_d},\boldsymbol{y},\boldsymbol{l_d},\boldsymbol{L_{M_d}},\boldsymbol{u}\cdot\right)\\&\propto\frac{n^{u,x}_{\neg(u,d)}+\eta}{\sum_{g'\in G}\left(n^{u,g'}_{\neg(u,d)}+\eta\right)}\frac{n^{u,x,z}_{\neg(u,d)}+\alpha}{\sum_{z'\in T}\left(n^{u,x,z'}_{\neg(u,d)}+\alpha\right)}\prod_{m\in M_d}P\left(l_m\mid\mu_g,\Sigma_g\right)\end{aligned}\tag{3}$$

其中,$n^{u,x}$是用户u生成的目标用户聚集区域为x的数目,$n^{u,x,z}$表示给定区域x之后主题z是根据用户u在区域x上的多项式分布生成的次数,$\neg(u,d)$表示所有的计数均未计入当前实例.此外,μ_g表示区域g的均值向量,Σ_g表示区域g的协方差矩阵,可根据如下公式抽样:

$$\mu_g=E(g)=\frac{1}{|C_g|}\sum_{m\in C_g}l_m\tag{4}$$

$$\Sigma_g=D(g)=\frac{1}{|C_g|-1}\sum_{m\in C_g}\left(l_m-\mu_g\right)\left(l_m-\mu_g\right)^T\tag{5}$$

其中,C_g表示被分配到区域g的目标用户集合.

在抽样区域g之后,我们可根据如下公式为同一条提及活动抽样出区域g依赖的主题分配:

$$P\left(z_{(u,d)}=k\mid \boldsymbol{z}_{\neg(u,d)},\boldsymbol{g},\boldsymbol{W_d},\boldsymbol{M_d},\boldsymbol{y},\boldsymbol{l_d},\boldsymbol{L_{M_d}},\boldsymbol{u}\cdot\right)$$
$$\propto \frac{n^{u,g,k}_{\neg(u,d)}+\alpha}{\sum_{z'\in T}\left(n^{u,g,z'}_{\neg(u,d)}+\alpha\right)}\frac{n^{k,l_d}_{\neg(u,d)}+\xi}{\sum_{l'\in L_D}\left(n^{k,l'}_{\neg(u,d)}+\xi\right)}\prod_{w\in W_d}\frac{n^{k,w}_{\neg(u,d)}+\beta}{\sum_{w'\in W}\left(n^{k,w'}_{\neg(u,d)}+\beta\right)} \tag{6}$$
$$\times\prod_{m\in M_d}\sum_{w\in W_d}\frac{n^{m,k,w}_{\neg(u,d)}+\gamma}{\sum_{m'\in M}\left(n^{m',k,w}_{\neg(u,d)}+\gamma\right)}$$

其中,g为已抽样的区域,$n^{u,g,k}$则表示用户u的提及活动中区域为g且主题为k的数目,n^{k,l_d}为位置l_d在给定主题k后根据位置多项式分布抽样生成的次数,$n^{k,w}$是词语w由主题为k的主题-词语分布抽样生成的次数,$\neg(u,d)$表示所有的计数均未计入当前实例.

此外,潜在开关变量y可根据如下后验概率抽样:

$$P\left(y_{(u,d,w)}=0\mid \boldsymbol{y}_{\neg(u,d,w)},\boldsymbol{z},\boldsymbol{g},\boldsymbol{W_d},\boldsymbol{M_d},\boldsymbol{l_d},\boldsymbol{L_{M_d}},\boldsymbol{u}\cdot\right)$$
$$\propto \frac{n^{y=0}_{\neg(u,d,w)}+\lambda}{\sum_{y'\in[0,1]}\left(n^{y=y'}_{\neg(u,d,w)}+\lambda\right)}\frac{n^{z,w,y=0}_{\neg(u,d,w)}+\beta}{\sum_{w'\in W}\left(n^{z,w',y=0}_{\neg(u,d,w)}+\beta\right)} \tag{7}$$

$$P\left(y_{(u,d,w)}=1\mid \boldsymbol{y}_{\neg(u,d,w)},\boldsymbol{z},\boldsymbol{g},\boldsymbol{W_d},\boldsymbol{M_d},\boldsymbol{l_d},\boldsymbol{L_{M_d}},\boldsymbol{u}\cdot\right)$$
$$\propto \frac{n^{y=1}_{\neg(u,d,w)}+\lambda}{\sum_{y'\in[0,1]}\left(n^{y=y'}_{\neg(u,d,w)}+\lambda\right)}\frac{n^{b,w,y=1}_{\neg(u,d,w)}+\beta}{\sum_{w'\in W}\left(n^{b,w',y=1}_{\neg(u,d,w)}+\beta\right)} \tag{8}$$

其中,$n^{y=0}$表示词语由主题-词语分布生成的次数,$n^{y=1}$表示词语由备用词语分布生成的次数,$n^{z,w,y=0}$表示词语w被分配为主题词语的次数,$n^{b,w,y=1}$则表示w被分配为备用词语的次数,$n_{\neg(u,d,w)}$意味着所有计数均未计入文档(u,d)的当前词语w.

经过足够次数的迭代后,即可通过统计z和g分配给提及活动的次数得到的近似后验概率来估计模型的参数.同时,我们可根据收缩吉布斯抽样方法得到如下对 JUMBM 参数的估计:

$$\theta^{u,g,z}=\frac{n^{u,g,z}+\alpha}{\sum_{z'\in T}\left(n^{u,g,z'}+\alpha\right)} \tag{9}$$

$$\phi^{u,g}=\frac{n^{u,g}+\eta}{\sum_{g'\in G}\left(n^{u,g'}+\eta\right)} \tag{10}$$

$$\vartheta^{z,w}=\frac{n^{z,w}+\beta}{\sum_{w'\in W}\left(n^{z,w'}+\beta\right)} \tag{11}$$

$$\psi^{z,w,m}=\frac{n^{z,w,m}+\gamma}{\sum_{m'\in M}\left(n^{z,w,m'}+\gamma\right)} \tag{12}$$

$$\pi^{z,l_d}=\frac{n^{z,l_d}+\xi}{\sum_{l'\in L_D}\left(n^{z,l'}+\xi\right)} \tag{13}$$

3 目标用户推荐

在本节中,我们介绍如何将学习到的 JUMBM 模型(即参数集$\hat{\Phi}=\left\{\hat{\theta},\hat{\vartheta},\hat{\psi},\hat{\phi},\hat{\pi},\hat{\mu},\hat{\Sigma}\right\}$)应用到目标推荐任务中.根据 JUMBM 的建模过程,给定一个查询$q=(u_q,l_q,W_q)$,u_q提及目标用户m的概率$P\left(m\mid u_q,l_q,W_q,\hat{\Phi}\right)$可计算为

$$P\left(m\mid u_q,l_q,W_q,\hat{\Phi}\right)=\frac{P\left(m,l_q\mid u_q,W_q,\hat{\Phi}\right)}{\sum_{m'\in M}P\left(m',l_q\mid u_q,W_q,\hat{\Phi}\right)}\propto P\left(m,l_q\mid u_q,W_q,\hat{\Phi}\right) \tag{14}$$

其中,概率 $P\left(m,l_q \mid u_q,W_q,\hat{\Phi}\right)$ 可进一步分解为

$$P\left(m,l_q \mid u_q,W_q,\hat{\Phi}\right)=\sum_{g\in G}\left[P(g)P\left(l_m \mid g,\hat{\Phi}\right)P\left(m,l_q \mid g,u_q,W_q,\hat{\Phi}\right)\right] \tag{15}$$

其中,区域 g 的先验概率 $P(g)$可计算为

$$P(g)=\sum_{u\in U}P(g \mid u)P(u)=\sum_{u\in U}P(u)\hat{\phi}_{u,g} \tag{16}$$

$$P(u)=\frac{|D_u|+\varepsilon}{\sum_{u'\in U}\left(|D_{u'}|+\varepsilon\right)} \tag{17}$$

其中,$P(u)$为 u 的先验概率;$|D_u|$表示 u 的提及活动数量.同时,我们在计算 $P(u)$时使用一个 Dirichlet 先验参数ε作为平滑参数,从而避免了过拟合现象的产生[12,13].接下来,概率 $P\left(m,l_q \mid g,u_q,W_q,\hat{\Phi}\right)$ 可进一步分解为

$$P\left(m,l_q \mid g,u_q,W_q,\hat{\Phi}\right)=\sum_{z\in T}\left[P\left(z \mid g,u_q,\hat{\Phi}\right)P\left(l_q \mid z,\hat{\Phi}\right)\prod_{w\in W_q}P\left(w \mid W_q\right)P\left(m \mid W_q,z,\hat{\Phi}\right)\right] \tag{18}$$

其中,$P(w|W_q)$表示词语 w 在 W_q 中的权重,在本文中我们使用词语的反文档频率(inverse documentfrequency,简称 IDF).此外,由于 $\psi_{m,z,w}$ (即 $P\left(m \mid W_q,z,\hat{\Phi}\right)$)矩阵的规模非常大(潜在的大小为 $K\cdot V\cdot N_M$),为了减轻数据稀疏性带来的参数估计困难,受 Gong 等人的启发[9,21],我们使用与主题无关的词对齐概率 $p(m|w)$对 $\psi_{m,z,w}$ 进行平滑操作,即:

$$\psi^{*}_{m,z,w}=\sigma\psi_{m,z,w}+\left(1-\sigma\right)p\left(m|w\right) \tag{19}$$

其中,σ是一个平滑参数且$\sigma\in[0,1]$;$p(m|w)$表示词 w 和目标用户 m 间的主题无关的词对齐概率,在本文中我们使用经典的 IBM 翻译模型 1[32]来计算.

根据公式(14)~公式(19),u_q 提及用户 m 的概率可通过公式(18)计算.通过这种方式,即可将计算得到的 top-k 概率值最大的目标用户作为推荐结果.

$$\begin{aligned}&P\left(m \mid u_q,l_q,W_q,z,\hat{\Phi}\right)\\&\propto\sum_{g\in G}\left[\sum_{z\in T}\left[P(g)P\left(l_m \mid g,\hat{\Phi}\right)P\left(z \mid g,u_q,\hat{\Phi}\right)P\left(l_q \mid z,\hat{\Phi}\right)\prod_{w\in W_q}P\left(w \mid W_q\right)P\left(m \mid W_q,z,\hat{\Phi}\right)\right]\right]\\&=\sum_{g\in G}\sum_{z\in T}\left[P(g)P\left(l_m \mid \hat{\mu}_g,\hat{\Sigma}_g\right)\hat{\theta}_{z,g,u_q}\hat{\pi}_{l_q,z}\prod_{w\in W_q}P\left(w \mid W_q\right)\hat{\psi}^{*}_{m,z,w}\right]\end{aligned} \tag{20}$$

3.1 高效top-k推荐

为了加快系统对在线查询的响应速度,基于公式(18),我们将目标用户 m 针对查询 q 的得分 $S(q,m)$的计算过程分成在线评分 $O(q,t)$计算和线下评分 $F(t,m)$计算两个部分:

$$S(q,m)\ =\sum_{t\in(z,a)}O(q,t)F(t,m)=\|\vec{q}\|\|\vec{m}\|\cos\left(\Delta_{\vec{q},\vec{m}}\right)\propto\cos\left(\Delta_{\vec{q},\vec{m}}\right) \tag{21}$$

$$O(q,t)=\hat{\theta}_{z,g,u_q}\hat{\pi}_{l_q,z}\prod_{w\in W_q}P\left(w \mid W_q\right)\hat{\psi}^{*}_{m,z,w} \tag{22}$$

$$F(t,m)=P(g)P\left(l_m \mid \hat{\mu}_g,\hat{\Sigma}_g\right) \tag{23}$$

其中,t 表示 $T\times G$ 集中的一个属性(即 $t=(z,g),z\in T$ 且 $g\in G$),$F(t,m)$表示线下计算的一个目标用户 m 关于属性 t 的分数,$O(q,t)$则表示在线计算的查询 q 在属性 t 上的权重.可以看到,除了 $F(t,m)$,$O(q,t)$中的主要部分,如 $\hat{\theta}_{z,g,u_q}$, $\hat{\pi}_{l_q,z}$ 和 $\hat{\psi}^{*}_{m,z,w}$ 也是通过线下计算的方式获得的.给定一个查询,$O(q,t)$的计算仅仅是对这些已得到的数值进行了线上整合.

一个直接的目标用户生成方法即通过扫描所有候选目标用户,基于公式(19)将 k 个得分最高的目标用户作为推荐结果.然而,该方法对于每一位候选用户均扫描一遍属性集,这会产生严重降低系统的在线推荐效率.在本文中,目标用户针对查询的得分是通过计算查询向量和目标用户向量的内积得到的(如公式(19)所示),且模型中的所有目标用户向量都具有相同的长度.因此,一个直观的针对该最大内积检索(maximum inner product search,简称 MIPS)问题的高效率解决方案,即首先通过一些数学转换方法[35]将该 MIPS 问题转化为一个 k 近邻(k-nearest neighbors,简称 KNN)检索问题,然后使用一些高效率近似 KNN 检索算法[6,7,34]来

寻找推荐结果.然而,JUMBM 模型潜变量分布的高维度问题会严重降低大多数低维检索算法的效率(即“维数灾难”问题).尽管一些高维检索框架基于不同的空间-文本索引在一定程度上减轻了高维度带来的检索困难,但其仅能应用于基于关键字的文本检索任务,不能很好地处理本文中的高语义相关度的用户活动检索问题[6,20].此外,这些算法都是针对近似而非精确检索任务设计的,直接应用在目标用户推荐中会严重降低检索准确率.近来,Yin 等人设计了基于聚类的分支定界算法 CBB[13]和属性剪枝算法 AP[12]以降低高维数据检索中的时间消耗,可以在仅扫描一部分项目(在本文的推荐任务中,项目即候选目标用户)或者属性的情况下找到使得最终得分最大的 top-k 个项目.具体来说,AP 算法通过筛选具有较低查询偏好和项目相关性的属性来约束属性检索空间,CBB 算法则通过选择在每个属性具有较高的上限得分项目组来对项目空间进行剪枝.但是,即使 AP 算法只扫描每个项目的一部分属性,其仍然需要遍历所有项目;CBB 算法则必须扫描每个选定项目的所有属性来计算完整的得分.换句话说,AP 和 CBB 算法仍需扫描大部分的项目或者属性空间来确定最终的推荐得分.

受 CBB 算法[13]和 AP 算法[12]的启发,我们设计了一种混合剪枝算法 HPA(hybrid pruning algorithm),以实现对项目和属性检索空间的联合剪枝.HPA 算法基于两个基本的观察事实:(1) 由于向量 $\vec{q}$ 和向量 $\vec{m}$ 的模是两个常量,其内积大小仅与两个向量的方向有关,如公式(19)所示;(2) 只有当查询 q 在属性 t 上的权重较高,且 m 关于属性 t 的分数也较高时,其对应的 $O(q,t)F(t,m)$值才会对最终得分有足够的影响.因此在本文中,我们首先使用球面 K-means 算法[36]根据目标用户向量的方向将目标用户聚类进 B 组中.然后,对每个桶 $b\in B$ 计算一个上限向量 $\vec{a}$, 使其在每一个属性上的 $F(t,m)$值最大(即 $F(t,a)=\max_{m\in b}F(t,m)$).显然,由于 $O(q,t)$的值非负,$S(q,a)$即为桶 b 中所有目标用户的上限得分.同时,对于每一个属性 t_i($i\in[1,K\cdot R]$)计算一个目标用户排序列表 L_{t_i},使其包含 k 个 $F(t_i,m)$值最高的目标用户.而对每个目标用户 m,我们根据 $F(t_i,m)$的值对其对应的属性进行降序排序.注意,所有上述计算和排序工作都是以线下计算的方式进行的.

给定一个在线查询 q,HPA 的在线处理流程如算法 2 所示.具体来说,HPA 首先根据向量 $\vec{q}$ 和每个桶的上限向量 $\vec{a}$ 的内积 $S(q,a)$的值对所有的桶进行排序(第 2 行).然后,基于查询和目标用户在属性上的权重找到 k 个候选目标用户(第 4 行~第 15 行).在此过程中,我们首先选择 top-N 个具有最小 N 值且覆盖了绝大部分的查询偏好的属性,即选择 $\sum_{i=1}^{n}O(q,t_i)>0.9\sum_{j=1}^{T\times G}O(q,t_j)$(第 4 行).对于每一个 top-$N$ 属性 t,我们从 L_t 中选择得分最高的目标用户作为候选用户(第 5 行~第 14 行).此外,HPA 采用二叉最小堆来实现列表 L,使得 L 的根项 m'(即堆顶)在 L 中具有最小的最终得分.此后,HPA 顺序地扫描所有的桶.对于一个未扫描的桶 b,若 $S(q,m')$不小于 b 的上限得分 $S(q,a)$,则 HPA 不再扫描后续桶,算法提前终止(第 18 行~第 20 行).否则,HPA 顺序扫描 b 中每一个目标用户来计算得分(第 21 行~第 37 行).在此过程中,对于每一个需要计算得分的目标用户 m,我们基于 Zhao 等人提出的区域剪枝技术[37]来检查能否避免遍历 m 对应的所有属性.具体来说,假设 HPA 已遍历了属性$\{t_1,t_2,...,t_i\}$,则可知这些属性对应的部分得分 S_p 的值为 $\sum_{j=1}^{i-1}O(q,t_j)F(t_j,m)$. 而对于第 i 项属性 t_i,其对目标用户 m 的上限得分为 $S_p+\sum_{j=i}^{T\times G}O(q,t_j)F(t_j,m)$. 由于 HPA 按照 $F(t,m)$值降序遍历所有的属性,则 $\{t_{i+1},...,t_{K\times R}\}$ 中的任一属性 t 对应的 $F(t,m)$值都比 $F(t_i,m)$值要小.也就是说,属性 $\{t_{i+1},...,t_{K\times R}\}$ 对应的部分得分值应小于等于 $\sum_{j=i}^{T\times G}O(q,t_j)F(t_j,m)$. 因此,$m$ 对应的所有属性的得分值 $S(q,m)$的上限即为 $S_p+\sum_{j=i}^{T\times G}O(q,t_j)F(t_j,m)$. 当堆顶 m'对应的得分值小于该上限得分时,HPA 将不再需要扫描任何 $\{t_{i+1},...,t_{K\times R}\}$ 中的属性(第 26 行~第 29 行).否则,HPA 将扫描 m'对应的所有属性来计算 m'的得分(第 31 行~第 36 行).

算法 2. 混合剪枝算法 HPA.

输入:

M:分成 B 个桶的目标用户集合

q:查询,$q=(u_q,l_q,W_q)$

L_t:每一个属性对应的有序目标用户列表

输出:

L:具有最高得分的 top-k 目标用户列表

1. 根据$S(q,a_b)$的大小对所有的桶进行排序;
2. 根据$O(q,t)$的大小对所有的属性进行排序;
3. 找到top-N满足 $N \leftarrow \min\left(\left\{n \mid \sum_{i=1}^{n} O(q,t_i) > 0.9\sum_{j=1}^{T\times G} O(q,t_j), T\times G\right\}\right)$ 的属性;
4. **for each** $t \in N$ **do**
5. **for each** $m \in L_t$且 $m \notin L$ **do**
6. **if** $L.size() < k$ **then**
7. $L.insert(\langle m, S(q,m)\rangle)$;
8. **else**
9. $m' \leftarrow L.top()$;
10. **if** $S(q,m) > S(q,m')$ **then**
11. $L.deleteTop()$;
12. $L.insert(\langle m, S(q,m)\rangle)$;
13. **end if**
14. **end if**
15. **end for**
16. **for each** $b \in B$ **do**
17. $m' \leftarrow L.top()$;
18. **if** $S(q,m') \geqslant S(q,a_b)$ **then**
19. break;
20. **else**
21. **for each** $m \in b$且 $m \notin L$ **do**
22. $S_p \leftarrow 0$; $O_p \leftarrow 0$, $Skip \leftarrow$ false;
23. **while** 对于m存在未扫描的属性 t **then**
24. $S_p \leftarrow S_p + O(q,t)F(t,m)$;
25. $O_p \leftarrow O_p + O(q,t)$;
26. **if** $S_p + \left(\sum_{i=1}^{T\times G} O(q,t_i) - OP\right)F(t,m) \leqslant S(q,m')$ **then**
27. $Skip \leftarrow$ true;
28. break;
29. **end if**
30. **end while**
31. **if** $Skip$ = false **then**
32. **if** $S(q,') > S(q,m')$ **then**
33. $L.deleteTop()$;
34. $L.insert(\langle m, S(q,m)\rangle)$;
35. **end if**
36. **end if**
37. **end for**
38. **end if**
39. **end for**

40. 逆排序L;
41. **return** L;

4 实验与分析

为了定量地分析 JUMBM 的性能,我们在两个真实数据集上构建了实验.本节将详细介绍实验构造和过程,并对实验结果进行展示和分析.

4.1 数据采集与实验设定

本文的实验构建在采集自新浪微博和 Twitter 的数据集上.数据集的采集策略如下.

微博数据集.受 Gong 和 Wang 等人[9,49]的启发,我们采用雪球抽样方法,从 5 个约有 1 000 个粉丝的初始用户开始,按照用户间的关注关系(即 following-follow 关系)抽取中国微博用户的个人信息和历史消息(受新浪微博 API 的限制,我们仅能采集每个用户的前 1 000 位关注用户).该采集策略能够保证得到微博完整数据图谱的一个子图,覆盖包括普通用户、明星用户、媒体工作者、微博广告商在内的全部微博用户类型,从而最大程度地降低数据采集偏置性.为了保证得到足够的位置数据,我们选择至少发布过 3 条带有位置标签的消息的用户.从 2014 年 1 月~2014 年 7 月,我们共抽取了 1 701 286 条用户个人信息和 26 994 838 条带有位置标签的消息.为了构建用户间的提及关系网络,我们抽取包含一个以上“@”实例的消息.然后,使用一种简单、有效的用户定位方法为所有可定位的目标用户分配一个居住位置(经度/维度).最终,我们得到由 147 621 名用户发布的 594 187 条带有位置标签和提及实例的消息,以及 251 054 名已知居住位置的目标用户构成的数据集.

Twitter 数据集.我们使用同样的抽取策略获得 Twitter 上美国用户的个人信息和历史消息来构建 Twitter 数据集.从 2016 年 10 月 15 日~2017 年 3 月 15 日,我们得到了一个由 1 620 081 名用户发表的 35 219 791 条带有位置标签的消息组成的原始数据集.在进行用户提及关系网络构建及可定位目标用户抽取处理之后,最终的 Twitter 数据集包含 553 145 条带有位置标签和提及实例的消息,以及由 807 330 条提及关系关联着的 133 625 位当前用户和 187 843 位居住位置已知的目标用户.表 3 列取了对本文数据集的统计信息.

Table 3 Statistics of Weibo and Twitter datasets

表 3 微博及 Twitter 数据集信息统计

统计	数据集	
	微博	Twitter
消息数	594 187	553 145
所有用户数	309 416	261 949
当前用户数	147 621	133 625
目标用户数	251 054	187 843
提及实例数	1 300 259	807 330
平均每位当前用户发出的提及实例数	8.81	6.04
平均每条消息中包含的提及实例数	2.19	1.46

需要注意的是,大多数微博和 Twitter 用户并未在其个人信息中公开居住位置,或者仅对其进行了粗粒度的描述(如居住城市).因此,我们采用一个被广泛使用的[40,42,43]ground-truth 用户位置推理方法为每一位目标用户分配一个居住位置(经度/维度).具体来说,对于用户 u,我们首先计算出 L_{D_u} 中所有位置确定的 l_1-多元几何中心点[36]作为 u 的初始位置.注意,我们选择使用几何中心点而非区域的中位点的原因是其对位置异常值是鲁棒的,尤其是当用户有在远离其通常活动的区域的活动历史时.然后,我们筛选出历史移动轨迹不正常的用户.具体来说,我们仅保留至少发布过 3 条且发布自 15km 地理半径范围内的消息的用户.同时,我们根据消息标记的时间戳和位置筛选出移动时速超过 1 000km/h 的用户.此外,对于指定了详细家庭位置(例如,GPS 坐标或详细位置描述)的用户,我们通过检索地理数据库获得其对应的位置点.本文中,我们使用 GeoName 数据库(http://www.geonames.org/)和高德地理编码 API(http://lbs.amap.com/api/javascript-api/guide/map-data/geocoding)将用户对位置的描述转化为经纬度坐标.

此外,对于一条用户提及活动 $d=(u,W_d,l_d,M_d)$,有$|M_d|\geqslant 1$.因此,我们将 d 中每一条(u,W_d,l_d,m),$m\in M_d$视为一条训练/测试用例.也就是说,本文数据集中的训练/测试用例的数量是由数据集包含的提及实例的数量而不是消息的数量决定的.受 Gong 等人[9]研究的启发,我们采用基于时间的分割方法将数据集分割为训练集和测试集.即,首先根据发表时间的先后将所有提及活动进行排序,然后按照 80/20 的比例对数据集进行划分.

4.2 评估指标

对于测试集中的一条测试用例(u,W_d,l_d,m),L_k 包含 k 个具有最高得分的目标用户.对于一个 ground-truth 目标用户 m^*,若有 $m^*\in L_k$,则计数为一个命中(hit),否则计数为一个未命中(miss).然后,我们采用准确率 Accuracy@k 来评估推荐的质量,即:

$$Accurcy@k=\frac{Hits@k}{N_{testcase}} \tag{24}$$

其中,$Hits@k$ 表示对于给定的 k 值,系统在测试集上的命中次数;$N_{testcase}$ 则表示所有测试用例的数目.除此之外,我们采用评估推荐系统性能常用的精确率(precision)、召回率(recall)和 $F1$ 值($F1$-meature)指标对目标推荐结果进行评估.同时,采用平均逆序排名(mean reciprocal rank,简称 MRR)来评估结果列表的排序效果.

4.3 对比方法

本文从推荐有效性和推荐效率两个方面对所有方法进行性能评估.

4.3.1 推荐有效性

在推荐有效性方面,我们采用以下方法作为对比方法.

(1) HIS.HIS 是一种基于用户提及历史的基准目标用户推荐方法[8,9].给定一条查询,HIS 方法推荐训练集中被当前用户提及次数最多的目标用户.

(2) CAR.CAR[8]是一个内容感知的目标用户推荐框架.通过抽取内容、社交、位置和时间信息相关的特征,CAR 使用学习排序(learning-to-rank)方法为一条促销类的消息推荐有高回应率的目标用户.在本文中,我们使用与原文[8]相同的特征集来实现 CAR 框架.因此,我们额外抽取了每一位目标用户在数据集限定的日期内发表的全部历史消息以及用户间的交互历史(用户间的提及与转发历史)来丰富我们的原始数据集.

(3) A-UUTTM.A-UUTTM 是由 Gong 等人提出的基于翻译模型的潜变量主题模型[9].该模型同时考虑了当前消息的内容和目标用户的内容历史,是目前用于目标用户推荐效果最好的方法.由于目标用户的内容历史对 A-UUTTM 来说至关重要,我们抽取了原始数据集中所有目标用户在被“@”时间之前发表的最新的 4 条消息.同时,我们使用与文献[9]相同的参数设置来实现 A-UUTTM.

此外,本文设计了 3 种 baseline 方法以验证在建模过程中分别考虑用户语义模式、目标用户的地理聚集区域以及当前用户的移动模式情况下的系统性能.

(1) JUMBM-NT.JUMBM-NT 是本文提出的 JUMBM 模型的一个变种,其仅考虑空间因素而忽略了用户的语义模式对其提及行为的影响,即在建模过程中忽略了用户提及活动中的文本词语 W.在 JUMBM-NT 模型中,给定一个主题,当前用户根据一个主题-目标用户分布来选择目标用户.

(2) JUMBM-NA.与 JUMBM 相比,JUMBM-NA 模型去除了对目标用户地理聚集区域的建模过程.在 JUMBM-NA 模型中,对于一条用户提及活动 d,主题 z 负责生成词语 W_d、目标用户 $m\in M_d$ 和位置 l_d,且每一个目标用户 $m\in M_d$ 都依据基本的主题-翻译过程生成.

(3) JUMBM-NP.JUMBM-NP 是 JUMBM 模型的另一个变种.JUMBM-NP 不考虑用户的移动模式对用户提及行为的影响,即对于一条用户提及活动中的位置 d,JUMBM-NP 忽略了对用户当前位置 l_d 的建模过程.

4.32 推荐效率

为了评估本文提出的 HAP 算法的效率,我们将 HAP 与以下 4 种算法进行比较.

(1) Brute Force Algorithm(BF).对于给定的查询,BF 算法通过扫描每个目标用户的所有属性来计算目标用户针对查询的得分,然后统计得分最高的 k 个目标用户作为推荐结果.BF 也是第 4.3.1 节中提到的对比方法

HIS、CAR 和 A-UUTTM 使用的推荐算法.

(2) Threshold Algorithm(TA).TA 算法[26]是对传统的基于阈值的方法的一个扩展,是当前效果较好的低维度检索算法.

(3) Clustering-based Branch and Bound Algorithm(CBB).CBB 算法[13]是一种项目搜索空间剪枝算法,能够在不扫描所有项目的情况下找到正确的 top-k 结果.

(4) Attribute Pruning Algorithm(AP).AP 算法[12]是一种用于属性空间剪枝的分支界定算法.对于一个项目,AP 算法能够在仅扫描部分属性的情况下确定该项目是否为正确的 top-k 结果.

4.4 结果与分析

4.4.1 推荐有效性

本节中,我们从推荐有效性方面对 JUMBM 进行评估.经过多次前期实验,将参数 K 和 R 的值设置为 K=70, R=90,以展现 JUMBM 的最优性能(参见第 4.4.3 节).此外,参照前期研究[9,21],我们在实验中将平滑参数σ设置为σ=0.8.表 4 和表 5 分别列取了在微博和 Twitter 数据集上,各方法在精确率、召回率、$F1$ 值、MRR、Accuracy@3 和 Accuracy@5 等指标下的性能表现.图 4 显示了不同推荐目标用户数量,即不同 k 值下的各种方法的实验精确率、召回率和 $F1$ 值.注意,图 4 中仅列出了 $k\in[1,7]$范围的结果值,这是因为,当 k>7 时各指标下的实验结果值已不再显著变化.

微博数据集上的推荐.在微博数据集上,本文提出的方法在所有评价指标下均取得了最优的实验结果.如表 4 所示,JUMBM 的推荐精确率为 0.647,召回率为 0.634,$F1$ 值为 0.641.与当前用于目标用户推荐效果最好的 A-UUTTM 方法相比,本文提出的基于 JUMBM 的推荐方法分别在精确率、召回率、$F1$ 值上取得了 11.2%、13.6%和 12.5%的相对提高.此外,Accuracy@3 和 Accuracy@5 的结果值显示,66.1%的 ground-truth 目标用户能够在 top-3 结果中找到,68.9%的 ground-truth 目标用户出现在了 top-5 结果中.同时,JUMBM 在 MRR 值方面也取得了最优结果,这表明,JUMBM 不仅能够生成准确的推荐,也能生成良好的推荐结果排序.图 4 所示的实验结果显示,随着被推荐目标用户的增多,JUMBM 方法的性能有所下降,但仍然在所有指标上均优于对比方法.当为一条消息推荐最佳目标用户时,JUMBM 可以获得最佳的 $F1$ 值.如果想获得最高的召回率,则可使用 JUMBM 推荐更多的目标用户.

Table 4 Experimental results on the Weibo dataset in different metrics

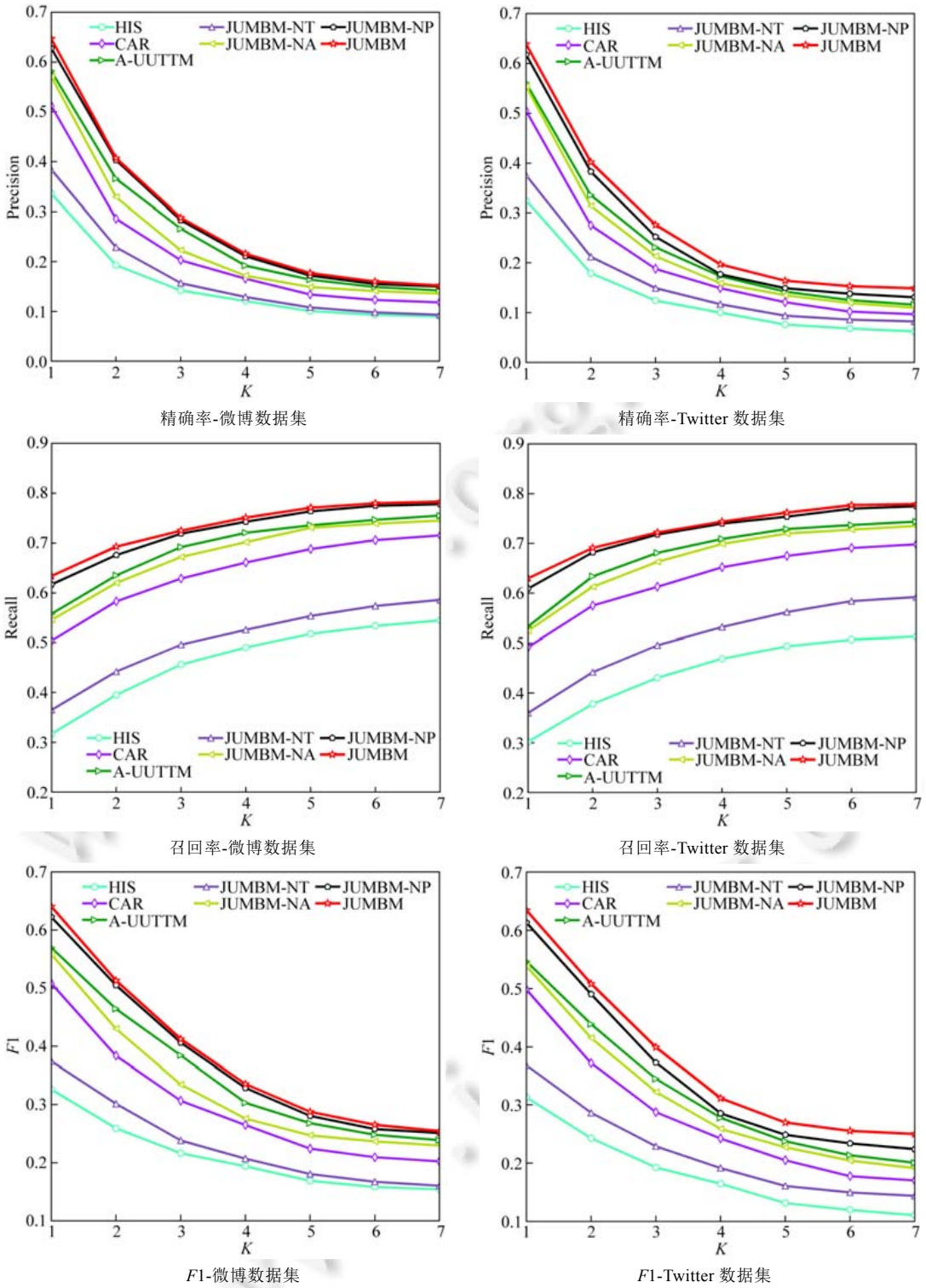
表 4 微博数据集上不同指标下的实验结果

方法	结果					
	精确率	召回率	$F1$	MRR	Accuracy@3	Accuracy@5
HIS	0.336	0.317	0.326	0.588	0.384	0.425
CAR	0.512	0.504	0.508	0.572	0.572	0.602
A-UUTTM	0.582	0.558	0.570	0.603	0.617	0.649
JUMBM-NT	0.384	0.365	0.374	0.398	0.393	0.424
JUMBM-NA	0.571	0.546	0.558	0.577	0.606	0.632
JUMBM-NP	0.628	0.617	0.622	0.656	0.641	0.665
JUMBM	**0.647**	**0.634**	**0.641**	**0.679**	**0.663**	**0.689**

Table 5 Experimental results on the Twitter dataset in different metrics

表 5 Twitter 数据集上不同指标下的实验结果

方法	结果					
	精确率	召回率	$F1$	MRR	Accuracy@3	Accuracy@5
HIS	0.325	0.302	0.313	0.575	0.353	0.392
CAR	0.505	0.492	0.498	0.552	0.544	0.586
A-UUTTM	0.561	0.533	0.547	0.581	0.604	0.625
JUMBM-NT	0.376	0.360	0.368	0.402	0.385	0.417
JUMBM-NA	0.554	0.524	0.539	0.569	0.575	0.612
JUMBM-NP	0.617	0.609	0.613	0.626	0.633	0.658
JUMBM	**0.640**	**0.631**	**0.635**	**0.665**	**0.652**	**0.675**



精确率-微博数据集　　精确率-Twitter 数据集

召回率-微博数据集　　召回率-Twitter 数据集

*F*1-微博数据集　　*F*1-Twitter 数据集

Fig.4　Results in precision, recall and *F*1 score with different values of top-*k* on the Weibo and Twitter datasets

图 4　微博和 Twitter 数据集上不同 top-*k* 值对应的结果精确率、召回率和 *F*1 值

从实验结果中可以观察到:

(1) HIS 方法在所有指标上均落后于其他方法.这表明,仅基于历史的目标用户建议方法可能会导致不准确的推荐结果.

(2) 与 CAR、A-UUTTM 和 JUMBM 相比,JUMBM-NT 的性能表现最差.这说明,使用用户提及活动的语义模式来挖掘用户提及行为偏好的必要性.与其他方法相比,JUMBM-NT 仅基于用户的提及活动的空间上下文信息来捕捉其行为偏好,忽略了文本内容信息的影响.

(3) JUMBM 和 A-UUTTM 的表现均优于 CAR,证明了一个严格设计的概率生成模型相对于一般的基于特征的学习-排序方法在表征用户提及行为方面的优势.

(4) 尽管 A-UUTTM 同时考虑了当前用户和目标用户的内容信息,但 JUMBM 仍然在所有指标上均取得了更好的结果,表明了引入用户提及活动空间上下文因素对其提及行为建模的必要性.与 A-UUTTM 相比,JUMBM 考虑了目标用户的地理分布和当前用户的移动模式,能够更好地表征和发掘用户的行为偏好,CAR 和 A-UUTTM 均忽略了这两个地理因素对用户提及行为的影响.

(5) 在我们的实验中,综合考虑了内容和空间因素的方法,其性能表现均优于仅考虑了单个因素的推荐方法.例如,当 k=5 时,JUMBM 的推荐准确率为 0.689.这一结果与仅考虑空间上下文因素的 JUMBM-NT 相比提高了 62.5%,与仅基于内容信息的 A-UUTTM 相比提高了 6.2%.显示出在建模过程中综合考虑内容和空间因素对准确挖掘用户提及行为偏好的必要性.

Twitter 数据集上的推荐.从表 5 和图 4 可以看出,Twitter 数据集上的实验与微博数据集上的实验有相同的结果趋势,即 JUMBM 在所有指标上一惯地优于对比方法.但是,在 Twitter 数据集上,各方法的性能表现均稍低于其在微博数据集上的性能表现.这可能是因为 Twitter 数据集中的用户平均提及活动数量较少,导致模型对用户提及行为偏好的推理不够准确.

4.4.2 各因素对推荐性能的影响

通过比较 JUMBM 及其变种模型的实验结果,我们可以进一步评估当前用户的语义模式(C)、目标用户的地理分布(G)以及当前用户的移动模式(S)等因素对 JUMBM 推荐性能的影响.表 6 和表 7 分别展示了 JUMBM 及其变种模型在微博和 Twitter 数据集上的实验 top-k 推荐准确率.其中,方法的推荐准确率越高,则各变种模型的缺失因素越不重要.注意,表 6 和表 7 仅展示了 k 小于 10 时各方法的性能表现.

Table 6 Top-k recommendation accuracy of JUMBM and its variant models on the Weibo dataset

表 6 微博数据集上 JUMBM 及其变种模型的 top-k 推荐准确率

方法	Accuracy@k			
	k=3	k=5	k=7	k=9
JUMBM-NT	0.393	0.424	0.441	0.458
JUMBM-NA	0.606	0.632	0.646	0.658
JUMBM-NP	0.641	0.665	0.686	0.697
JUMBM	**0.663**	**0.689**	**0.708**	**0.722**

Table 7 Top-k recommendation accuracy of JUMBM and its variant models on the Twitter dataset

表 7 Twitter 数据集上 JUMBM 和其变种模型的 top-k 推荐准确率

方法	Accuracy@k			
	k=3	k=5	k=7	k=9
JUMBM-NT	0.385	0.417	0.436	0.452
JUMBM-NA	0.575	0.612	0.627	0.641
JUMBM-NP	0.633	0.658	0.679	0.692
JUMBM	**0.652**	**0.675**	**0.696**	**0.713**

因为对于本文研究的推荐任务来说,更大的 k 值是没有意义的.从结果中可以看到,每个因素对推荐准确率的影响是不同的.总的来看,各因素对目标用户推荐性能影响程度的顺序为 $C \succ G \succ S$.具体而言,首先,语义模式在目标用户推荐性能方面起着主导作用.例如,仅考虑了两个空间因素(G 和 S)的 JUMBM-NT,比基于同样的空间因素但同时考虑内容信息的 JUMBM 的推荐准确率低 50%以上.这一结果与 Chen、Tang 和 Zhou 等人的研

究[2,8,10]是一致的.其次,考虑目标用户的地理分布对提高目标用户推荐准确率来说十分必要.例如,与 JUMBM-NA 相比,JUMBM 对目标用户的地理聚集区域进行了建模,这使其在微博数据集上的实验推荐准确率提高了约 9%(k=9 时).第三,考虑了当前用户移动模式的方法的推荐性能表现略优于未考虑这一因素的方法.如表 6 所示,对用户移动模式建模的 JUMBM 方法相比于未考虑用户移动模式的 JUMBM-NP 方法的推荐精确率提高了约 3.5%(k=9 时).

4.4.3 参数敏感性分析

我们构建了一系列的实验来研究模型参数的改变对推荐结果的影响.在 JUMBM 中,最重要的两个参数即主题数量K和区域数量R.本节我们仅展示使用不同参数的JUMBM在微博数据集上的实验结果,因为在Twitter数据集上的实验结果与之是相似的.

具体来说,我们展示了在主题数量 $K\in[50,90]$和区域数量 $R\in[70,110]$时 JUMBM 在 Accuracy@5 指标下的性能变化,见表 8.对于模型中的超参数,受前期研究的启发[12,53],我们将其设置成固定值以节省计算成本,即 $\alpha=50/K$、$\eta=50/R$、$\beta=\gamma=\xi=0.01$、$\lambda_1=\lambda_2=0.5$.表 8 展示了不同 K 和 R 值下的 Accuracy@5 结果.可以看到,在 K 和 R 值较小时,JUMBM 的推荐精确率随着 K 和 R 值的增加而迅速增大,但当其超过一定的阈值,即 K=70 且 R=90 时,JUMBM 的推荐精确率不再显著变化.这是因为,在 JUMBM 中,主题和区域的数量代表了模型的复杂度.当 K 和 R 值较小时,模型对数据的描述能力有限.相对地,当 K 和 R 增大到一定程度时,模型已复杂到足够处理当前数据.此后,随着 K 和 R 的进一步增大,数据稀疏性问题会变得越来越严重,继而导致模型过拟合并使得模型学习到的参数不再可靠.因此,在构造第 4.4.1 节中的实验时,我们采用设置 K=70、R=90 作为推荐有效性和推荐效率间的一个折衷.

Table 8 Accuracy@5 result with various K and R

表 8 不同 K 和 R 值下的 Accuracy@5 结果

		K				
		K=50	K=60	K=70	K=80	K=90
R	R=70	0.643	0.662	0.669	0.670	0.670
	R=80	0.657	0.674	0.680	0.681	0.683
	R=90	0.665	0.677	**0.689**	0.689	0.690
	R=100	0.666	0.677	0.689	0.690	0.691
	R=110	0.666	0.678	0.689	0.690	0.692

4.4.4 推荐效率

通过与第 4.3 节中提出的 4 种算法进行比较,我们在微博数据集上对本文提出的 HPA 算法的在线推荐效率进行评估.我们使用 Java 语言(JDK 1.7)实现所有在线检索算法,并将所有程序运行在一台配置了 Intel Xeon E5 处理器(2.4 GHz)、128 G 内存、操作系统为 Windows Server 2008 R2 的 PC 上.

我们在 6 300 维(即 K=70,R=90)的检索情况下,将 k 分别设置为 1、3、5、7 和 9 来展示各算法的性能.同时,我们使用数据集中的所有查询对算法进行测试.此外,经过若干次的预先实验,我们将 HPA 算法使用的桶的数量 B 设置为 430 以达到算法的最佳性能.图 5 给出了 BF、CBB、AP 和 HPA 算法的平均在线查询处理时间.注意,我们没有在图 5 中列出 TA 的处理时间.这是因为相比于其他算法,TA 的时间消耗过于严重.在我们的实验中,TA 需要 8 464.3ms 来产生 top-1 结果,同时需要近 11 000ms 来产生正确的 top-5 结果.TA 效率如此低下的原因是其作为一种单层次检索算法,必须频繁地更新阈值并维护排序列表的动态优先级队列,使其仅能处理低维数据检索问题.如图 5 所示,本文提出的 HPA 算法在推荐效率方面显著优于所有对比方法.例如,HPA 算法在 296.1ms 内从超过 25 万名的候选目标用户集合中找到了正确的 top-5 结果.这仅相当于 TA 算法时间消耗的 2.7%、BF 算法的 38.04%、CBB 算法的 54.8%以及现有最高效的 AP 算法的 59.03%.在我们的实验中,当 k=5 时,HAP 平均仅需扫描 643 个属性(约占全部属性的 10.2%).最终的统计结果显示,需要遍历所有属性来计算得分的候选目标用户数量为 11 046,仅占全部候选目标用户的 4.4%.尽管随着被推荐目标用户数量的增加,HPA 算法的推荐效率有所下降,但即使在 k=9 的情况下 HPA 算法仍能以相对其他算法更快的速度响应在线查询.可以看出,在数据维度很高的情况下,本文提出的 HPA 算法比遍历检索算法 BF、低维检索算法 TA、空间剪枝算法 CBB 和当前

效率最高的检索算法 AP 的计算效率都更高.

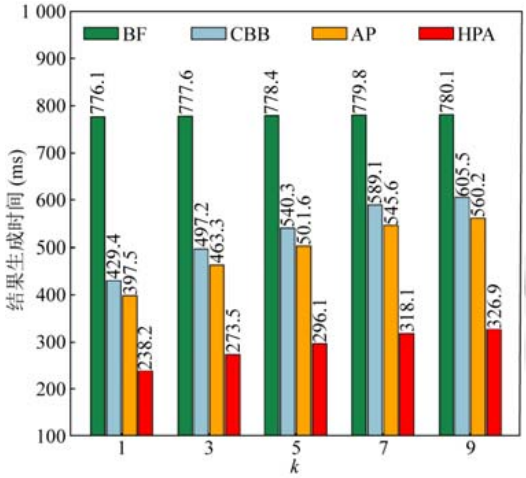


Fig.5 Recommendation efficiency of different methods on Weibo dataset

图 5 微博数据集上不同方法的推荐效率

5 总 结

当前,社交媒体中的提及机制正在用户在线交互和网络信息传播方面扮演着重要角色.本文从普通用户的在线交互角度着手,通过对用户提及行为的分析和建模构建一个推荐系统,为给定的消息自动生成候选目标用户,从而帮助用户识别目标、节省搜索时间.通过对大型真实社交媒体数据集的分析发现,用户的提及行为受其提及活动的语义和空间上下文因素的联合影响.针对此,本文提出一个联合隐式概率生成模型 JUMBM 来模拟用户提及活动的生成过程.JUMBM 通过一种统一的方式进行语义和空间上下文感知的用户提及行为联合建模,发掘用户的移动模式、地理区域依赖的语义兴趣及目标用户的地理聚集模式.同时,本文提出一个混合剪枝算法 HPA,用以应对推荐中遇到的"维数灾难"问题并加快推荐系统对在线 top-k 查询的响应速度.在大型真实数据集上的实验结果表明,本文提出的方法在推荐有效性和推荐效率方面均优于对比方法.

References:

[1] Bobadilla J, Ortega F, Hernando A, Gutiérrez A. Recommender systems survey. Knowledge-based Systems, 2013,46:109–32.

[2] Zhou G, Yu L, Zhang CX, Liu C, Zhang ZK, Zhang J. A novel approach for generating personalized mention list on micro-blogging system. In: Proc. of the 2015 IEEE Int'l Conf. Data Mining Workshop (ICDMW). IEEE, 2015. 1368–1374.

[3] Li Q, Song D, Liao L, Liu L. Personalized mention probabilistic ranking–recommendation on mention behavior of heterogeneous social network. In: Proc. of the Int'l Conf. on Web-age Information Management. Cham: Springer-Verlag, 2015. 41–52.

[4] Wang B, Wang C, Bu J, Chen C, Zhang WV, Cai D, He X. Whom to mention: Expand the diffusion of tweets by@ recommendation on micro-blogging systems. In: Proc. of the 22nd Int'l Conf. on World Wide Web. ACM, 2013. 1331–1340.

[5] Hu B, Ester M. Spatial topic modeling in online social media for location recommendation. In: Proc. of the 7th ACM Conf. on Recommender Systems. ACM, 2013. 25–32.

[6] Zhang D, Chan CY, Tan KL. Processing spatial keyword query as a top-k aggregation query. In: Proc. of the 37th Int'l ACM SIGIR Conf. on Research & Development in Information Retrieval. ACM, 2014. 355–364.

[7] Zhou K, Zha H. Learning binary codes for collaborative filtering. In: Pro. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2012. 498–506.

[8] Tang L, Ni Z, Xiong H, Zhu H. Locating targets through mention in Twitter. World Wide Web, 2015,18(4):1019–49.

[9] Gong Y, Zhang Q, Sun X, Huang X. Who Will You@. In: Proc. of the 24th ACM Int'l on Conf. on Information and Knowledge Management. ACM, 2015. 533–542.

[10] Chen K, Chen T, Zheng G, Jin O, Yao E, Yu Y. Collaborative personalized Tweet recommendation. In: Proc. of the 35th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2012. 661–670.

[11] Weng J, Lim EP, Jiang J, He Q. Twitterrank: Finding topic-sensitive influential Twitterers. In: Proc. of the 3rd ACM Int'l Conf. on Web Search and Data Mining. ACM, 2010. 261–270.

[12] Yin H, Cui B. Spatio-temporal Recommendation in Social Media. Singapore: Springer-Verlag, 2016.

[13] Yin H, Cui B, Zhou X, Wang W, Huang Z, Sadiq S. Joint modeling of user check-in behaviors for real-time point-of-interest recommendation. ACM Trans. on Information Systems (TOIS), 2016,35(2):11.

[14] Xu Z, Zhang Y, Wu Y, Yang Q. Modeling user posting behavior on social media. In: Proc.. of the 35th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2012. 545–554.

[15] Qiu M, Zhu F, Jiang J. It is not just what we say, but how we say them: LDA-based behavior-topic model. In: Proc. of the 2013 SIAM Int'l Conf. on Data Mining. Society for Industrial and Applied Mathematics, 2013. 794–802.

[16] Yin H, Cui B, Chen L, Hu Z, Zhou X. Dynamic user modeling in social media systems. ACM Trans. on Information Systems (TOIS), 2015,33(3):10.

[17] Bi B, Cho J. Modeling a retweet network via an adaptive bayesian approach. In: Proc. of the 25th Int'l Conf. on World Wide Web. Int'l World Wide Web Conferences Steering Committee, 2016, 459–469.

[18] Michelson M, Macskassy SA. Discovering users' topics of interest on twitter: A first look. In: Proc. of the 4th Workshop on Analytics for Noisy Unstructured Text Data. ACM, 2010. 73–80.

[19] Cheng J, Adamic L, Dow PA, Kleinberg JM, Leskovec J. Can cascades be predicted. In: Proc. of the 23rd Int'l Conf. on World Wide Web. ACM, 2014. 925–936.

[20] Wu D, Cong G, Jensen CS. A framework for efficient spatial Web object retrieval. The VLDB Journal—The Int'l Journal on Very Large Data Bases, 2012,21(6):797–822.

[21] Ding Z, Qiu X, Zhang Q, Huang X. Learning topical translation model for microblog hashtag suggestion. In: Proc. of the IJCAI. 2013. 2078–2084.

[22] Gong Y, Zhang Q, Huang X. Hashtag recommendation for multimodal microblog posts. Neurocomputing, 2018,272:170–7.

[23] Zhao F, Zhu Y, Jin H, Yang LT. A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. Future Generation Computer Systems, 2016,65:196–206.

[24] Zhang Y, Wang H, Yin G, Wang T, Yu Y. Social media in GitHub: The role of@-mention in assisting software development. Science China Information Sciences, 2017,60(3):032102.

[25] Koenigstein N, Ram P, Shavitt Y. Efficient retrieval of recommendations in a matrix factorization framework. In: Proc. of the 21st ACM Int'l Conf. on Information and Knowledge Management. ACM, 2012. 535–544.

[26] Yin H, Sun Y, Cui B, Hu Z, Chen L. Lcars: A location-content-aware recommender system. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2013. 221–229.

[27] Liu B, Fu Y, Yao Z, Xiong H. Learning geographical preferences for point-of-interest recommendation. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2013. 1043–1051.

[28] Cheng Z, Shen J. Just-for-me: An adaptive personalization system for location-aware social music recommendation. In: Proc. of the Int'l Conf. on Multimedia Retrieval. ACM, 2014.

[29] Wang W, Yin H, Chen L, Sun Y, Sadiq S, Zhou X. Geo-sage: A geographical sparse additive generative model for spatial item recommendation. In: Proc. of the 21th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2015. 1255–1264.

[30] Fang Q, Xu C, Hossain MS, Muhammad G. Stcaplrs: A spatial-temporal context-aware personalized location recommendation system. ACM Trans. on Intelligent Systems and Technology (TIST), 2016,7(4):59.

[31] Kurashima T, Iwata T, Irie G, Fujimura K. Travel route recommendation using geotags in photo sharing sites. In: Proc. of the 19th ACM Int'l Conf. on Information and Knowledge Management. ACM, 2010. 579–588.

[32] Brown PF, Pietra VJD, Pietra SAD, Mercer RL. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 1993,19(2):263−311.

[33] Dhillon IS, Modha DS. Concept decompositions for large sparse text data using clustering. Machine Learning, 2001,42(1):143−75.

[34] Koenigstein N, Ram P, Shavitt Y. Efficient retrieval of recommendations in a matrix factorization framework. In: Proc. of the 21st ACM Int'l Conf. on Information and Knowledge Management. ACM, 2012. 535−544.

[35] Shrivastava A, Li P. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In: Advances in Neural Information Processing Systems. 2014. :2321−2329.

[36] Vardi Y, Zhang CH. The multivariate $L1$-median and associated data depth. Proc. of the National Academy of Sciences, 2000,97(4): 1423−6.

[37] Zhao K, Cong G, Yuan Q, Zhu KQ. SAR: A sentiment-aspect-region model for user preference analysis in geo-tagged reviews. In: Proc. of the 31st IEEE Int'l Conf. on Data Engineering (ICDE 2015). IEEE, 2015. 675−686.

[38] Ye M, Yin P, Lee WC, Lee DL. Exploiting geographical influence for collaborative point-of-interest recommendation. In: Proc. of the 34th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2011. 325−334.

[39] Takhteyev Y, Gruzd A, Wellman B. Geography of Twitter networks. Social networks, 2012,34(1):73−81.

[40] McGee J, Caverlee J, Cheng Z. Location prediction in social media based on tie strength. In: Proc. of the 22nd ACM Int'l Conf. on Information & Knowledge Management. ACM, 2013. 459−468.

[41] McGee J, Caverlee JA, Cheng Z. A geographic study of tie strength in social media. In: Proc. of the 20th ACM Int'l Conf. on Information and Knowledge Management. ACM, 2011. 2333−2336.

[42] Compton R, Jurgens D, Allen D. Geotagging one hundred million twitter accounts with total variation minimization. In: Proc. of the Big Data (Big Data). IEEE, 2014. 393−401.

[43] Jurgens D. That's what friends are for: Inferring location in online social media platforms based on social relationships. In: Proc. of the ICWSM. 2013,13(13):273−82.

[44] Brown C, Nicosia V, Scellato S, Noulas A, Mascolo C. The importance of being placefriends: Discovering location-focused online communities. In: Proc. of the 2012 ACM Workshop on Online Social Networks. ACM, 2012. 31−36.

[45] Lim KH, Chan J, Leckie C, Karunasekera S. Detecting location-centric communities using social-spatial links with temporal constraints. In: Proc. of the European Conf. on Information Retrieval. Cham: Springer-Verlag, 2015. 489−494.

[46] Ye M, Shou D, Lee WC, Yin P, Janowicz K. On the semantic annotation of places in location-based social networks. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2011. 520−528.

[47] Soliman A, Yin J, Soltani K, Padmanabhan A, Wang S. Where Chicagoans Tweet the most: Semantic analysis of preferential return locations of Twitter users. In: Proc. of the 1st Int'l ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics. ACM, 2015. 55−58.

[48] Yin H, Zhou X, Cui B, Wang H, Zheng K, Nguyen QVH. Adapting to user interest drift for poi recommendation. IEEE Trans. on Knowledge and Data Engineering, 2016,28(10):2566−2581.

[49] Wang K, Yu W, Yang S, Wu M, Hu YH, Li SJ. Location inference method in online social media with big data. Ruan Jian Xue Bao/Journal of Software, 2015,26(11):2951−2963(in Chinese). http://www.jos.org.cn/1000-9825/4907.html [doi: 10.13328/j.cnki.jos.004907]

[50] Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X. Comparing Twitter and traditional media using topic models. In: Proc. of the European Conf. on Information Retrieval. Berlin, Heidelberg: Springer-Verlag, 2011. 338−349.

[51] Zhao Z, Cheng Z, Hong L, Chi EH. Improving user topic interest profiles by behavior factorization. In: Proc. of the 24th Int'l Conf. on World Wide Web. Int'l World Wide Web Conferences Steering Committee, 2015. 1406−1416.

[52] Huang W, Kataria S, Caragea C, Mitra P, Giles CL, Rokach L. Recommending citations: Translating papers into references. In: Proc. of the 21st ACM Int'l Conf. on Information and Knowledge Management. ACM, 2012. 1910−1914.

[53] Griffiths TL, Steyvers M. Finding scientific topics. Proc. of the National academy of Sciences, 2004,101(1):5228−35.

附中文参考文献:

[49] 王凯,余伟,杨莎,吴敏,胡亚慧,李石君.一种大数据环境下的在线社交媒体位置推断方法.软件学报,2015,26(11):2951−2963. http://www.jos.org.cn/1000-9825/4907.htm [doi: 10.13328/j.cnki.jos.004907]

汤小月(1983－),女,湖北武汉人,博士,讲师,CCF 专业会员,主要研究领域为数据挖掘,社交网络,智能优化算法.

周康(1965－),男,博士,教授,主要研究领域为生物计算,智能优化算法,数据库,数据挖掘.

王凯(1988－),男,博士生,主要研究领域为空间数据挖掘与推理,智慧城市.