

基于位置的社交网络中双重异质社区的聚类与关联方法

龚卫华¹⁾ 沈 松¹⁾ 裴小兵²⁾ 杨旭华¹⁾

¹⁾(浙江工业大学计算机科学与技术学院 杭州 310023)

²⁾(华中科技大学软件学院 武汉 430074)

摘 要 近年来,异质信息网络特别是基于位置的社交网络(Location-Based Social Networks, LBSN)中的社区发现已成为新兴的研究热点. 然而,目前大多数社区发现研究仅考虑基于同质结构的社交网络,显然都已无法有效融合 LBSN 这种异质网络所包含的多模实体及其多维关系. 为了应对该挑战性问题,本文提出了一种新的双重社区聚类与关联方法(Communities Clustering and Associating Method, CCAM),该方法先在 LBSN 的社交媒体层上,通过信息熵度量用户发布主题之间的相似性,进而再将相似用户兴趣聚类问题转换成求解基于模糊聚类的目标函数以获得重叠的兴趣主题簇结构. 然后在地理位置层中,将用户-位置签到关系网络形成的二分图转换为超图模型,并采用超边聚类方式得到用户关于地理位置的兴趣点特征簇. 最后,在兴趣主题簇与地理位置簇之间借助中间用户层的社交关系建立这两层异质簇间的关联性表示模型,并通过随机梯度下降法求解模型的局部最优解. 在两个真实数据集 Foursquare(NYC)和 Yelp 上的实验结果表明,本文提出的 CCAM 方法有效融合了用户-媒体发布关系、用户间社交关系、用户-位置签到关系等多维度关系,能准确获得 LBSN 中紧密关联的用户兴趣主题簇与地理位置簇,使得这双层社区结构不仅在外部分布特征与兴趣内聚性指标上都优于传统算法,并且还在兴趣主题推荐与位置兴趣点推荐方面的平均准确率提高至少 32%.

关键词 基于位置的社交网络;异质社区发现;多维关系;超图聚类

中图法分类号 TP393 **DOI号** 10.11897/SP.J.1016.2020.01909

Clustering and Associating Method of Dual Heterogeneous Communities in Location Based Social Networks

GONG Wei-Hua¹⁾ SHEN Song¹⁾ PEI Xiao-Bing²⁾ YANG Xu-Hua¹⁾

¹⁾(School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023)

²⁾(School of Software, Huazhong University of Science and Technology, Wuhan 430074)

Abstract In recent years, community discovery in heterogeneous information networks, especially in location-based social networks (LBSN), has become an emerging research hotspot attracting more and more attentions. However, since presently most of the traditional community discovery studies in social networks only focus on homogeneous network structures, there exists one obvious drawback in these studies that they cannot effectively integrate the multi-mode entities and their multi-dimensional heterogeneity relations included in LBSN. Therefore, in order to overcome this challenging problem, this paper proposes a novel dual heterogeneous communities clustering and associating method, which is called CCAM to fully fuse with multi-mode entities and their multi-relations. The main idea of CCAM is below: firstly, in the upper social media layer of LBSN,

收稿日期:2019-07-03;在线发布日期:2019-11-24. 本课题得到国家自然科学基金(61502420,61802346)、浙江省自然科学基金(LY13F020026,LQ18F020008)、中国博士后科学基金(2015M581957)资助. 龚卫华,博士,副教授,硕士生导师,中国计算机学会(CCF)会员,主要研究领域为社交网络、机器学习. E-mail: whgong@zjut.edu.cn. 沈 松,硕士研究生,主要研究方向为社交网络、推荐系统. 裴小兵,博士,教授,硕士生导师,中国计算机学会(CCF)会员,主要研究领域为机器学习、数据挖掘. 杨旭华,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为网络科学、机器学习.

this method measures the similarity of publishing document topics between users by means of information entropy, and further transforms the similar interests clustering problem into solving the objective function based on fuzzy clustering to identify the overlapping topic-based communities of users. Then, in the bottom geographical location layer, the bipartite graph between users and locations with check-ins relationships is converted to the hyper graph model, and the proposed hyper edge clustering approach based on graph partition is exploited to obtain the point of interest clusters about geographical features of users. Finally, the representation model for associating the upper topic-based communities and the bottom geographical location clusters is established via users' social relations in the middle user layer of LBSN, and the local optimal solution of association function for the dual heterogeneous clusters is gained by using stochastic gradient descent method. The experimental results in two real LBSN datasets such as Foursquare (NYC) and Yelp show that our proposed CCAM method can effectively fuse three types of relations in LBSN, for example, social relations of users, social media publishing relations, and users check-ins relations. Consequently, this method can accurately obtain the closely correlated dual heterogeneous clusters such as the user's interest clusters and geographical location clusters, which not only makes the external structural characteristic and the internal interest cohesive index better than some traditional community clustering algorithms, but also outperforms these algorithms at least 32% through measuring the mean average precision in the both field of online interest topics recommendation and offline point of interest recommendation.

Keywords location-based social networks; heterogeneous community discovery; multi-dimensional relations; hypergraph clustering

1 引 言

近年来,随着移动设备与 GPS 技术的快速发展,基于位置的社交网络 (Location-Based Social Networks, LBSN) 受到工业界和学术界的广泛关注,目前比较流行的 FacebookPlaces、Foursquare、Yelp 和 WeChat 等移动 APP 不仅具备传统社交网络的社交功能,还能提供许多与位置相关的服务如位置签到或热点推荐等. 作为一种新型的异质信息网络, LBSN 中不仅包含了大量的用户、媒体和位置等多模实体,还拥有丰富多样的异质关系,比如用户社交关系、用户与位置的签到关系以及用户发表的社交媒体信息等,这些数据混合在一起形成非常复杂的结构特征,尤其以用户为中心的各类社区结构已成为传统社交网络及其扩展的 LBSN 所普遍存在的一种共同特征,它有助于人们深刻认知线上与线下社会的群体结构与行为规律.

目前,在传统社交网络领域已有许多社区发现成果^[1], 总体上看,现阶段主流的社区发现方法大致有 4 种:基于团渗透的方法^[2]、基于标签传播的方法^[3-6]、

基于局部扩展与优化的方法^[7-9]、基于链接划分的方法^[10-12]等, 这些方法虽然能准确地发现以节点为中心或以边为中心的社区结构,但都存在的问题是仅适用于具有同类实体和单一关系的社交网络,而无法有效应对和发现包含多维关系的异质网络社区结构,特别是对于 LBSN 这种新型的异质网络既有用户、媒体和位置等多模实体,又有因这些实体间多维关系构成的异质社区,比如用户不仅在社交媒体空间活跃而形成兴趣主题社区,还在地理空间上由签到行为形成兴趣点簇,并且这两种结构之间还存在着一定的空间映射关系或关联性. 上述社区发现研究显然都不能有效融合多模实体及其多维异质关系,无法有效发现 LBSN 中所特有的社区结构. 另一方面,异质网络的社区发现研究也一直是个挑战性难题,当前仅有少数研究尝试提出了一些融合多维数据的方法,例如将多维网络集成为一个加权网络再扩展现有的社区发现方法揭示社区结构,或采用合并多个单维社区结构的方法来融合多种信息,但这些方法仍未能准确、真实地反映 LBSN 这种新型异质网络中的社区结构及关联性.

为此,本文针对 LBSN 中异质社区的发现难题,

提出一种融合多维关系的双重社区聚类与关联方法 CCAM (Communities Clustering and Associating Method), 该方法先分别在社交媒体层和地理位置层挖掘得到两种异质社区结构, 然后再建立起这两层异质簇间的关联性. 概括来看, 本文的主要贡献如下:

(1) 提出了一种双重异质社区聚类方法, 即先在社交媒体空间采用模糊聚类方式得到重叠的用户兴趣主题社区, 然后在地理位置空间上运用超图聚类方法获得兴趣点特征簇.

(2) 采用基于矩阵分解的表示模型建立了用户主题簇与地理位置簇之间的关联性, 由此展现用户兴趣主题在地理空间维度上的分布与关联规律.

(3) 本文充分融合了 LBSN 中包含的用户-媒体发布关系、用户间社交关系、用户-位置签到关系等 3 种维度的关系, 在两个真实 LBSN 数据集上的实验结果表明本文的 CCAM 方法所获得紧密关联的双重异质社区不仅在结构特征与兴趣内聚性指标上优于其他算法, 并且还在兴趣主题推荐与位置兴趣点推荐方面都具有最好的推荐效果.

本文第 2 节介绍社区发现的研究现状; 第 3 节描述 LBSN 的模型结构, 并提出双重异质社区的聚类与关联方法; 第 4 节进行仿真实验及结果分析; 最后, 第 5 节总结全文.

2 相关工作

现有信息网络中的社区结构研究工作主要有两大类: 面向同质网络的社区发现与面向异质网络的社区发现.

首先, 在同质网络中, 当前大多数的社区发现方法都聚焦于此, 该类研究归纳起来可分为以节点为中心的社区和以边为中心的社区. 前者比较典型的有基于团渗透的方法、基于标签传播的方法、基于局部扩展与优化的方法等, 其中, 基于团渗透的方法是将社区视为由一些团(全连通子图)构成的集合, 这些团之间通过共享节点而紧密连接, 代表性的算法如 CPM^[2]. 基于标签传播的方法是一类启发式算法, 主要过程是初始时给每个节点分配唯一的标签, 然后根据每个节点及其邻居节点的相似度传播标签, 最终使得具有相同标签的节点被划分到相同的社区中, 常见的算法有 LPA^[3]、SLPA^[4]、LPPB^[5] 和 LPANNI^[6] 等. 基于局部扩展与优化的方法是利用网络的局部特性不断挖掘网络中的社区结构, 例

如 EAGLE^[7]、GCE^[8]、RNM^[9] 等都是该方法的典型代表. 而对于以边为中心的社区, 目前相关研究还比较少, 主要有基于链接划分的方法, 它将链接而不是节点作为考虑对象, 通过设计适当的划分策略来获取链接社区结构. 典型的有 Ahn 等人^[10] 提出了一种基于边相似性划分的层次聚类算法来找出重叠的边社区, 还有一些研究如 Shi 等人^[11] 先采用遗传算法检测重叠的链接社区, 然后再将链接社区转化成节点社区. Zhou 等人^[12] 提出了一种基于链接密度聚类的重叠社区发现算法 DBLC.

其次, 在异质网络中, 社区发现的关键问题是如何有效融合多模实体及其多维关系, 从已有的研究成果来看, 主要分为两种: 一种研究是尝试用网络集成或社区聚合方法^[13] 来抽取社区结构, 网络集成是指将多维网络集成为一个加权网络, 再扩展现有的社区发现方法揭示社区结构, 但该方法容易丢失一些重要信息; 而社区聚合是指先在每层网络发现社区, 然后再合并这些社区, 最终得到多层网络中的社区结构. 例如 Tang 等人^[14] 先将多维网络分割成多个单维网络, 然后再集成各单维度上的聚类结果以获得最终的社区结构, 但该方法没有考虑不同网络层节点间的连接关系. Tagarelli 等人^[15] 提出一种集成聚类方法发现多层网络中的共识社区结构, 该方法先通过对节点的共关联矩阵直接聚类得到各层的社区结构, 然后采用爬山策略来优化基于模块度的层间社区间的连接性和层内社区间的连接性. 还有 Liu 等人^[16] 先采用传统模块度方法评价各子网络社区划分优劣, 再提出一种基于线性加权的复合模块度优化方法来集成每个单维子网络中的社区模块度. 最近 Pramanik 等人^[17] 也通过扩展模块度概念提出了一种新的多层模块度指标 Q_M 来评价多层网络中的社区发现质量, 本文实验中也采用该指标评价社区整体结构特性.

另一种研究是采用网络联合聚类方法直接融合多维数据来发现异质网络中的社区, 例如 Gao 等人^[18] 针对星型结构的高维异质数据提出了一种联合聚类方法 CBGC, 该方法将由 X 、 Y 和 Z 三种对象组成的三分图谱聚类转化为两对 X - Y 与 Y - Z 二分图联合划分问题, 并通过半定规划方法求得最优解. Mei 等人^[19] 在科研合作网中对多种关联的数据对象提出一种模糊聚类方法 FC-MR, 使得具有较大排序值的不同对象聚在同一簇中, 但这些多维数据对象实际上对应实体的多维属性. Hu 等人^[20] 提出了基于模糊聚类的重叠社区发现方法 FCAN, 该方

法仅考虑结合了网络中节点的属性特征与链接关系.最近,基于矩阵或张量分解^[21]的聚类方法成为处理多类型、多语义信息的重要手段,例如 Ding 等人^[22]最早提出了正交非负矩阵三分解的联合聚类方法,其表示形式如 $X \approx FSG^T$, 其中 F 表示行聚类的指示矩阵,而 G 表示列聚类的指示矩阵.另有 Pei 等人^[23]也进一步提出了含有图正则化项的非负矩阵三分解聚类框架 NMFT,该框架同时对社交网络中的用户与内容进行联合聚类,通过图正则化融合了 3 种类型的信息:用户相似性、消息相似性和用户交互关系.Comar 等人^[24]针对多重网络提出了一种基于非负矩阵分解的联合聚类方法,同时还考虑结合关于不同网络的社区间相关关系的先验信息来提高聚类质量,但该方法所采用的三因子矩阵分解方式仅适于处理两层异质网络结构,即包含两种实体与三种异质关系边.Ma 等人^[25]提出一种半监督的联合非负矩阵分解方法 S2-jNMF,以检测多层网络中的社区结构,但缺乏考虑异质关系的融合.

迄今为止,针对 LBSN 这种新型异质网络的社区发现研究比较少见,Wang 等人^[26]提出了以边为中心的联合聚类框架发现 LBSN 中重叠的用户社区结构,既考虑了用户之间的社交关系,又考虑了用户与地点的签到连接,同时还加入了这两种实体的属性信息,有效提高了社区发现质量.还有 Zhao 等人^[27]将 LBSN 这种多维异质网络建模成超图结构,超图中的顶点有用户、地点、评论或图片等多模实体,超边表示多方实体间形成的 4 种异质关系,然后提出基于稠密子图的软聚类算法来发现重叠社区,条件是要求稠密子图中超边权重能达到最大化.

综上所述,目前大量的社区发现成果都集中在传统社交网络领域,大多数方法仅能有效发现一些同质或单一的社区结构.虽然已有一些研究尝试采用集成或联合聚类方法来融合异质网络中的多维信息,但这些社区发现结果都有一定的局限性,有些方法考虑的数据维度偏少,容易丢失重要信息,而另一些方法则忽略了多维网络中不同层间的交互关系或关联性.针对异质网络特别是 LBSN 中的社区发现至今仍是一个挑战性难题.因此,本文将提出一种充分融合多维关系的双重异质社区聚类与关联方法,该方法先分别在社交媒体层和地理位置层挖掘得到用户兴趣主题簇与地理位置簇,再采用基于矩阵分解的表示模型建立这两层异质簇间的关联性.

3 LBSN 建模与双重异质社区的发现及关联方法

3.1 相关概念及形式化描述

LBSN 是一种由多模实体及其多维关系复合而成的异质网络,其模型结构如图 1 所示,主要有三层网络结构:社交媒体层(主题集)、用户层(用户集)和地理位置层(位置集),这三层都是以用户活动为纽带联系起来形成多模网络,社交媒体层是由用户发布或评论的媒体信息所形成的各种兴趣主题,用户层则是用户间交互形成的社交关系网络,地理位置层则展现了由用户与位置签到关系所构成的空间簇.

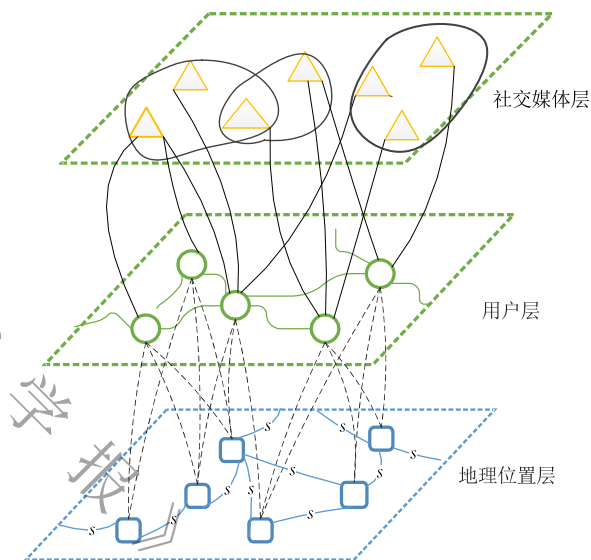


图 1 LBSN 的三层结构模型

为了更好地理解,下面给出 LBSN 的多模异质网络具体定义.

定义 1. 用户-媒体关系网络(G). 用户在社交媒体层上发表或评论的媒体信息可抽象成二部图结构,记为 $G=(U, T, W)$, 其中, U 表示用户集, T 表示媒体的主题集, W 表示用户对媒体的兴趣关联, 即 $W=\{(u_i, t_j) | u_i \in U, t_j \in T, U \cap T = \emptyset\}$.

定义 2. 用户社交网络(O). 在用户层中, 用户间社交关系形成的网络可表示成一个无向图结构, 记为 $O=(U, E)$, 其中, U 表示用户集, E 表示用户社交关系的边集, 即为 $E=\{(u_i, u_j) | u_i, u_j \in U\}$.

定义 3. 用户-位置签到网络(P). 用户在地理位置上的签到行为形成的关系网络可表示成二部图结构, 记为 $P=(U, L, C)$, 其中 U 表示用户集, L 表示地理位置集, C 表示用户与位置间的签到关系集, 即 $C=\{(u_i, l_j) | u_i \in U, l_j \in L, U \cap L = \emptyset\}$.

3.2 社交媒体层的兴趣主题簇聚类方法

在 LBSN 的社交媒体层中,为了获得用户的兴趣主题簇,首先需要计算用户各主题之间的相关性,本文采用文献[16]中的互信息方法来度量任意两个主题词间的关联性:

$$MI(t_i|t_j) = \log\left(\frac{p(t_i|t_j)}{p(t_i)}\right) \quad (1)$$

式(1)中, $p(t_i|t_j)$ 表示给定主题词 t_j 条件下包含主题词 t_i 的概率,而 $p(t_i)$ 表示主题词 t_i 在所有主题词中的占比.不难发现, $MI(t_i|t_j)$ 表明了主题词 t_i 对主题词 t_j 的依赖性.同理,也可通过计算 $MI(t_j|t_i)$ 得出主题词 t_j 对主题词 t_i 的依赖性.

进一步地,为了降低主题词间关联的偶然性,我们采用更精确的计算指标 $DMI(t_i|t_j)$ 如式(2)所示:

$$DMI(t_i|t_j) = MI(t_i|t_j) - MI(\bar{t}_i|t_j) \quad (2)$$

在此基础上,两个主题词之间的关联强度可计算为

$$strength(t_i, t_j) = \frac{DMI(t_i|t_j) + DMI(t_j|t_i)}{2} \quad (3)$$

由式(3)可知, $strength(t_i, t_j)$ 是通过计算两个主题词相互依赖性的平均值而得到主题词间的相关性指标.

然后,我们通过计算任意两个用户所发表主题词之间的平均关联度而得到他们的兴趣相似性,如下式:

$$r(u_i, u_j) = \frac{\sum_{p=1}^{|u_i|} \sum_{q=1}^{|u_j|} strength(u_{ip}, u_{jq})}{|u_i| |u_j|} \quad (4)$$

式(4)中, $|u_i|$ 与 $|u_j|$ 分别表示用户 i 与用户 j 的主题数,用户主题的关联强度 $strength(u_{ip}, u_{jq})$ 由式(3)得出.

下面,进一步构造用户相似性矩阵 $\mathbf{A}$,其中 $a_{ij} \in \mathbf{A} (1 \leq i, j \leq n, i \neq j)$ 且令 $a_{ij} = r(u_i, u_j)$, $a_{ii} = 0$. 另外,假设矩阵 $\mathbf{S}$ 表示用户在兴趣社区簇中的分布情况, $s_{if} \in \mathbf{S}$ 表示用户 i 对社区 f 的隶属度.我们将用户兴趣社区的聚类问题建模成如下目标函数:

$$\max J(\mathbf{S}) = Tr(\alpha \mathbf{S}^T \mathbf{A} \mathbf{S}) - \theta Tr(\mathbf{S}^T \mathbf{B}) - \frac{\varphi}{2} \|\mathbf{S}\|_F^2$$

$$\text{s. t. } \mathbf{S} \mathbf{1} = \mathbf{1}, \mathbf{S} \geq 0 \quad (5)$$

其中,参数 α, θ, φ 都在区间 $[0, 1]$ 中, $\mathbf{1}$ 表示的是元素都为 1 的向量. $\mathbf{B} = [b_{if}]$ 是一个 $n \times c$ 矩阵,其中

$$b_{if} = \sum_{j=1}^n \sum_{g=1, g \neq f}^c a_{ij} s_{jg}, c \text{ 表示簇参数. 不难看出,式(5)}$$

中第 1 项使得同一簇内的用户紧密相关,第 2 项计算用户及其邻居隶属分布的不一致性,第 3 项是正则化项.如果目标函数的值越大,表明用户兴趣簇的特征结构越显著.

为了求解上述目标函数,我们假设向量 $\boldsymbol{\lambda}^T = (\lambda_i) (1 \leq i \leq n)$ 和矩阵 $\boldsymbol{\Omega} = [\omega_{if}]_{n \times c}$ 都是拉格朗日算子,将式(5)改写成拉格朗日公式得到:

$$\max R(\mathbf{S}, \boldsymbol{\lambda}, \boldsymbol{\Omega}) = J(\mathbf{S}) + \boldsymbol{\lambda}(\mathbf{1} - \mathbf{S} \mathbf{1}) + Tr(\mathbf{S}^T \boldsymbol{\Omega}) \quad (6)$$

式(6)的 KKT 条件为

$$\partial \mathbf{R} / \partial \mathbf{S} = 2\alpha \mathbf{A} \mathbf{S} - \theta \mathbf{B} - \varphi \mathbf{S} - \boldsymbol{\lambda} + \boldsymbol{\Omega}^T = 0 \quad (7)$$

$$\partial \mathbf{R} / \partial \boldsymbol{\lambda} = \mathbf{1} - \mathbf{S} \mathbf{1} = 0 \quad (8)$$

$$\boldsymbol{\Omega} \circ \mathbf{S} = 0 \quad (9)$$

$$\boldsymbol{\Omega} \geq 0 \quad (10)$$

若存在一个解 $(\mathbf{S}^*, \boldsymbol{\lambda}^*, \boldsymbol{\Omega}^*)$ 满足以上所有 KKT 条件,则 $\mathbf{S}^*$ 被认为是式(5)的最优解,采用如下迭代更新规则得到:

$$\mathbf{S}_i^{(l+1)} = \frac{1}{\varphi} (2\alpha \mathbf{a}_i \mathbf{S}^{(l)} - \theta (\mathbf{b}_i^{(l+1)})^T - \boldsymbol{\lambda}_i^{(l+1)} \mathbf{1} + (\boldsymbol{\omega}_i^{(l+1)})^T) \quad (11)$$

$$\boldsymbol{\lambda}_i^{(l+1)} = \frac{1}{c} ((2\alpha \mathbf{a}_i \mathbf{S}^{(l)})^T \mathbf{1} - \theta \mathbf{b}_i^{(l+1)} \mathbf{1} - \varphi + \boldsymbol{\omega}_i^{(l+1)} \mathbf{1}) \quad (12)$$

$$b_{if}^{(l+1)} = \sum_{j=1}^n \sum_{g=1, g \neq f}^c a_{ij} s_{jg}^{(l)} \quad (13)$$

$$\omega_{if}^{(l+1)} = \begin{cases} 0, & S_{if}^{(l)} > 0 \\ 1, & S_{if}^{(l)} = 0 \end{cases} \quad (14)$$

其中, $S_{if}^{(l+1)}$ 的正则化公式如下:

$$S_{if}^{(l+1)} = \begin{cases} \frac{S_{if}^{(l+1)}}{\sum_{f=1}^c S_{if}^{(l+1)}}, & S_{if}^{(l+1)} > 0 \\ 0, & S_{if}^{(l+1)} = 0 \end{cases} \quad (15)$$

根据上述规则,迭代更新矩阵与各参数值,直到收敛至设定的阈值,由此便可得到最优的用户兴趣隶属矩阵 $\mathbf{S}^*$,具体而言,针对社交媒体层的用户兴趣主题聚类算法(SMC)过程如下.

算法 1. 社交媒体聚类(Social Media Clustering).

输入:用户-媒体关系网络 G , 簇数 c , 参数 $\alpha, \theta, \varphi, \epsilon$

输出:用户兴趣主题簇的归属矩阵 $\mathbf{S}$

1. 通过式(4)计算得到用户相似性矩阵 $\mathbf{A}(\mathbf{r})$;
2. 令 $i=1$, 用随机非负数初始化矩阵 $\mathbf{S}^{(i)}$ 并由式(15)标准化
3. REPEAT
4. 通过式(13)、(14)分别计算 $\mathbf{B}^{(i+1)}$ 和 $\boldsymbol{\Omega}^{(i+1)}$ 值.
5. 通过式(12)更新 $\boldsymbol{\lambda}^{(i+1)}$ 的值
6. 由式(11)更新 $\mathbf{S}^{(i+1)}$, 再由式(15)标准化
7. $\epsilon^{(i+1)} = |J(\mathbf{S}^{(i+1)}) - J(\mathbf{S}^{(i)})|$

8. $i=i+1$
9. UNTIL $\epsilon^{(i)} \leq \epsilon$ 或 $i > i_{\max}$
10. RETURN $S^{(i)}$

算法 1 的主要过程是先通过计算用户间的主题词关联性而得到用户的兴趣相似矩阵,再经过迭代求解目标函数以获得最优的用户隶属矩阵,即重叠的用户兴趣主题社区.该算法的主要计算代价包括用户之间的主题相似性计算和用户兴趣主题簇归属矩阵的迭代优化.首先,假设用户数为 n ,主题数为 m ,步骤 1 由式(4)计算用户相似矩阵时,需要先获得用户自身主题间的关联性,再计算用户对之间的相似度,所以该过程的整体时间复杂为 $O(n^2 \cdot m^2)$.然后,步骤 3~9 主要实现了矩阵及参数值的迭代更新,若迭代次数为 l ,则该算法总的时间复杂度为 $O(lcn + n^2 m^2)$,其中 c 是常数.

3.3 地理位置层的超图聚类方法

在地理位置层上,不同用户在地理空间上的签到行为形成了如定义 3 所示的用户-位置签到网络 P ,可表示成如图 2 所示的二部图.不难发现,二部图结构仅能描述用户与位置这两类节点间的二元对应关系,缺陷是失去了节点“同质性”,因而很难直观、完整地表达网络结构特征,例如用户的共同兴趣点及地理位置间的关联性无法直接体现.为了更好地解决该问题,本文将引入超图概念,并提出二部图转化为超图模型的方法,以便深入研究用户的位置兴趣点簇分布特征及关联性.

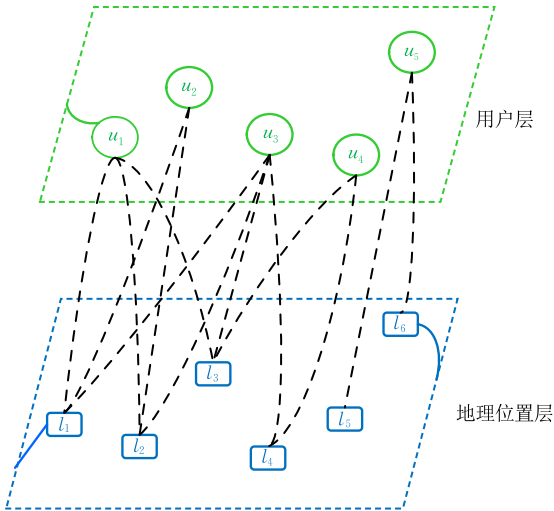


图 2 用户-位置签到网络的二部图结构

定义 4. 超图(H). 设 $U = \{u_1, u_2, \dots, u_n\}$ 是一个有限集,若 $D_i \neq \emptyset (i=1, 2, \dots, d)$ 且 $\bigcup_{i=1}^d D_i = U$, 则称二元关系 $H = (U, D)$ 为超图,其中 U 称为超图的顶点集合, $D = \{D_1, D_2, \dots, D_d\}$ 称为超图的边集合,

D 中的元素 $D_i = \{u_{i_1}, u_{i_2}, \dots, u_{i_j}\}, 1 \leq j \leq n$ 称为超边.

由定义 4 可知,超图中的用户顶点集仍与二部图中的用户集定义相同,而超边则是由闭合曲线包含多个顶点来表示具有共同的关系.于是,本文提出将二部图转化为超图的规则是:用户集 U 仍然保持不变,超边 D_{ij} 表示在二部图中兴趣点 l_i 和兴趣点 l_j 上同时签到的用户交集,其构成如下:

$$D_{ij} = \{u_k \mid (u_k, l_i) \in C \wedge (u_k, l_j) \in C, \exists u_k \in U, l_i, l_j \in L\}$$

$$\text{且 } D = \bigcup_{i \neq j} D_{ij} \quad (16)$$

根据上述规则,我们将用户-位置签到网络从图 2 所示结构转换为对应的超图模型如图 3 所示,该超图以用户为顶点,超边是由关联兴趣点表示的闭合圈,包含了在该超边位置上共同签到的用户.与图 2 对比发现,超图结构实质上是把二部图中的二元对应关系转化成超图里的包含关系,其优点是不仅能更简洁地表达两个甚至多个用户之间的共同签到关系,还能直观地体现超边所代表的兴趣点在地理空间上的分布与关联特征.

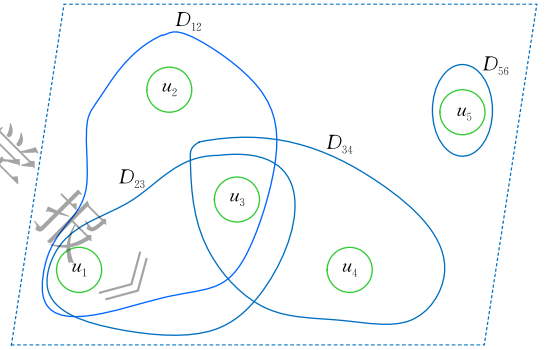


图 3 用户-位置签到网络的超图表示方法

在超图 H 中,任意两条超边 D_x 与 D_y 的相似性可采用 *Sorgenfrei* 系数来度量,如下式:

$$\text{sim}(D_x, D_y) = \frac{|D_x \cap D_y|^2}{|D_x| |D_y|} \quad (17)$$

由式(17)可知,两条超边的相似值越大表明他们包含的共同用户数越多,这样超边所代表的兴趣点就越聚集,从而就形成了关于用户的兴趣点聚簇结构.为了获得该结构特征,我们将相似超边的聚类问题建模成如下目标函数:

$$\text{Obj} = \arg \max_C \sum_{i=1}^g \sum_{D_{ij} \in D \wedge D_{ij} \notin C_i} \text{sim}'(D_{ij}, C_i) \quad (18)$$

其中, g 表示簇数, $C = \{C_1, C_2, \dots, C_g\}$ 表示划分的簇,而 $\text{sim}'(D_{ij}, C_i)$ 则表示超边 D_{ij} 与某簇 C_i 的相似性,如下式:

$$\text{sim}'(D_{ij}, C_i) = \frac{1}{|C_i|} \sum_{D_{xy} \in C_i} \text{sim}(D_{ij}, D_{xy}) \quad (19)$$

不难得知,上述目标函数表示的意义是将任意超边划分到与其最相似的超边簇中,使得同一社区簇内的超边具有最大的相似值.具体地,在地理位置层上的超边聚类算法过程如下.

算法 2. 超边聚类方法(Hyper-edge Clustering).

输入:地理位置层的超图 H ,超边聚簇数 g

输出:用户在位置簇上的归属矩阵 $\mathbf{P}$

1. 随机选取 g 个超边作为初始化簇 C'
2. 初始化 $\max sim_i = 0, obj_{old} = 0$
3. REPEAT
4. 新簇 $C \leftarrow \emptyset$
5. FOR each $d_i \in D$ //对于每条超边
6. FOR $k=1$ to g //对于每个簇
7. 由式(19)计算 $sim'(d_i, C_k)$
8. IF $sim'(d_i, C_k) > \max sim_i$ THEN
9. $\max sim_i \leftarrow sim'(d_i, C_k)$
10. END IF
11. END FOR
12. 将超边 d_i 加入到 $\max sim_i$ 值对应的新簇 C_k 中
13. END FOR
14. 新簇 C 的 $obj_{new} \leftarrow \sum \max sim_i$
15. $\Delta = |obj_{new} - obj_{old}|$
16. 更新簇 $C' \leftarrow C, obj_{old} \leftarrow obj_{new}$
17. UNTIL $\Delta \leq \epsilon$
18. FOR $k=1$ to g //对于每个超边簇
19. FOR each $d_i \in D$ //对于每条超边
20. FOR each $u_x \in d_i$ //对于超边中的每个用户
21. IF $d_i \in C_k$ THEN
22. $\mathbf{P}_{xk} = 1$
23. END FOR
24. END FOR
25. END FOR

算法 2 的时间复杂度主要分为超边聚类和计算用户对位置簇的归属矩阵这两部分,假设用户数为 n ,超边数为 e ,步骤 3~17 计算超边与各社区簇间的相似性,如果收敛迭代次数为 l ,那么其时间复杂度为 $O(le^2)$.步骤 18~25 是为超边中的用户分配所属位置簇,假设每超边的平均用户数为 n/e ,那么其时间复杂度为 $O(g \cdot e \cdot n/e)$,最终该算法总的时间复杂度约为 $O(le^2 + gn)$,其中 g 是常数.

3.4 社交媒体层与地理位置层的聚簇关联方法

通过对社交媒体层与地理位置层的聚类分析,我们既得到了用户兴趣主题簇,又获得了用户关于位置的感兴趣点特征簇,这两层结构分别考虑了用户-媒体发布关系与用户-位置签到关系.除此之外,在 LBSN 模型中处于用户层的各用户之间本身还有一维社交关系,本文将利用社交关系建立上述两层结

构之间关联性,如图 4 所示,社交媒体层上的用户兴趣簇属于重叠的社区结构,同样,在地理位置层上的位置簇表明用户在地理空间上的兴趣点分布特征,而用户社交关系则成为联系这两层异质簇的纽带,由此便能深入考察他们之间的关联性.

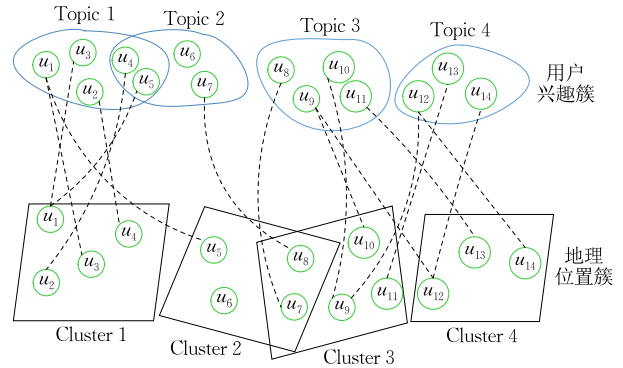


图 4 用户兴趣簇和位置簇间的社交关联

由图 4 可知,用户分别在社交媒体层和地理位置层上被划分成 c 个主题簇与 g 个位置簇.首先,假设矩阵 $\mathbf{S} = [\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_n]^T \in \mathbb{R}^{n \times c}$,其中 $\mathbf{S}_i$ 表示第 i 个用户在 c 个主题簇中的隶属特征向量.同样地,令矩阵 $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_n]^T \in \mathbb{R}^{n \times g}$,其中 $\mathbf{P}_j$ 表示第 j 个用户在 g 个位置簇中的隶属特征向量.另外,再假设矩阵 $\mathbf{R} \in \mathbb{R}^{c \times g}$ 表示这两种簇之间的关联度矩阵.

对于 LBSN 模型的用户层中社交关系网络 $O = (U, E)$,其邻接矩阵可表示成 $\mathbf{V} = (v_{ij})_{n \times n}$,元素 v_{ij} 的定义如下:

$$v_{ij} = \begin{cases} 1, & (u_i, u_j) \in E \\ 0, & \text{其他} \end{cases} \quad (20)$$

于是,我们进一步对用户社交关系构建基于矩阵分解的表示模型如下:

$$J = \frac{1}{2} \sum_{v_{ij} \neq 0} (v_{ij} - \mathbf{S}_i \mathbf{R} \mathbf{P}_j^T)^2 + \frac{\alpha}{2} \|\mathbf{R}\|_F^2 \quad (21)$$

为了最小化上述损失函数,一般采用随机梯度下降法(SGD)求解得到局部最优解,相应的矩阵变量 $\mathbf{R}$ 的迭代更新规则为

$$\mathbf{R}^{\tau+1} = \mathbf{R}^{\tau} + \gamma \left[\sum_{v_{ij} \neq 0} \mathbf{S}_i^T \cdot (v_{ij} - \mathbf{S}_i \mathbf{R}^{\tau} \mathbf{P}_j^T) \cdot \mathbf{P}_j - \alpha \mathbf{R}^{\tau} \right] \quad (22)$$

式(22)中,参数 γ 表示迭代更新步长的学习率, α 表示正则化因子, τ 表示梯度下降迭代次数.当两次迭代更新矩阵 $\mathbf{R}$ 的绝对值误差不大于 ϵ 时,则终止迭代.具体的计算过程如算法 3 所示.

算法 3. 两层异质簇间的关联方法(Dual-layer Heterogeneous Clusters Associating).

输入:用户层社交关系网络 O 的邻接矩阵 $\mathbf{V}$,社交媒体层聚簇数 c 和地理位置层聚簇数 g

输出: 兴趣簇矩阵 S 与地理位置簇矩阵 P 的关联矩阵 R

1. 通过算法 1 得到用户兴趣主题簇矩阵 S .
2. 通过算法 2 得到地理位置簇矩阵 P .
3. 初始化关联矩阵 R , 令矩阵的每个值都为 1.
4. REPEAT:
5. 由式(22)迭代计算得矩阵 R^{r+1}
6. UNTIL $|R^{r+1} - R^r| \leq \epsilon$ // 函数收敛至 ϵ ;

4 仿真实验与结果分析

4.1 实验数据集描述

本文选择两个典型的 LBSN 数据集 Foursquare (NYC) 和 Yelp 进行实验, 由于数据集本身存在一些噪声数据以及访问稀疏的位置点, 我们需要先对数据集进行过滤使得每个用户在 LBSN 模型的 3 个层次上都有较完整的关系或信息. 对于 Foursquare(NYC) 数据集, 筛选出了超过 2 人签到的地理位置, 评论数超过 5 条的用户及其所拥有的社交关系. 对于 Yelp 数据集, 筛选出了签到数超过 50 的地理位置数, 以及评论超过 5 条的用户及其所拥有的社交关系. 预处理完成后, 各数据集中的用户数、位置数、评论数以及社交关系和签到关系等信息统计结果如表 1 所示.

表 1 两种 LBSN 数据集的基本信息

数据集	用户数	签到数	社交链接数	位置数	评论数
Foursquare(NYC)	4811	59 027	44 397	27 303	113 254
Yelp	16 184	515 563	314 341	22 005	454 086

由表 1 可知, 数据集 Foursquare(NYC) 与 Yelp 上的用户签到密度都非常低, 分别为 $4.49E-4$ 和 $1.44E-3$, 这反映了 LBSN 中的兴趣点有较大的数据稀疏性, 而在用户社交关系上, 两个数据集的用户社交关系平均度约为 9.2 和 19.4, 这表明社交用户间的交互关系比较密切.

4.2 对比算法与评价指标

为了验证本文提出的双重异质社区聚类方法的有效性, 我们选取进行实验对比的算法描述如下:

(1) CNM: 是一种传统的基于模块度的社区发现算法^[28], 在 LBSN 网络中只考虑用户节点的社交关系.

(2) CBGC: 是由 Gao 等人^[18]提出的一种代表性的 K-分图谱聚类方法, 该方法将由 X、Y 和 Z 三种对象组成的三分图聚类问题转化为两个成对的 X-Y 与 Y-Z 联合聚类子问题, 对 LBSN 而言, CBGC 不仅考虑了媒体、用户与位置等多模对象, 还融合了

用户-媒体发表关系和用户-位置签到关系, 但其有一个缺点是忽略了用户间的社交关系.

(3) NMFTF: 是 Ding 等人^[22]提出的非负矩阵三因子分解的联合聚类方法, NMFTF 的表示形式如 $X \approx FSG^T$, 该方法应用于 LBSN 异质网络的社区发现时仅考虑结合了用户社交关系与用户签到关系这两种维度信息.

(4) CCAM: 是本文提出的双重社区聚类及关联方法, 充分考虑并融合了用户社交关系、用户-位置签到关系和用户-媒体发表关系等 3 种维度关系, 既能得到用户的兴趣主题簇, 又能获得用户关于地理位置的兴趣点簇.

为了评价两层异质社区簇之间的结构特征, 我们采用一种改进的模块度指标, 称为综合模块度 Q_M^C 来衡量异质簇的划分质量, 其方法如下:

$$Q_M^C = \sum_{i,j \in (M_1 \cup M_2)} \left[\frac{1}{3} \left\{ \frac{1}{2|E_1|} \sum_{i,j \in V_1} \frac{S_i \cdot S_j}{|S_i| |S_j|} \left(v_{ij} - \frac{(h_i \times h_j)}{2|E_1|} \right) + \frac{1}{|E_{12}|} \sum_{i \in M_1, j \in M_2} \left(v_{ij} - \frac{(c_i \times c_j)}{|E_{12}|} \right) + \frac{1}{2|E_2|} \sum_{i,j \in M_2} \frac{P_i \cdot P_j}{|P_i| |P_j|} \left(v_{ij} - \frac{(h_i \times h_j)}{2|E_2|} \right) \right\} \right] \quad (23)$$

式(23)中, M_1 表示在社交媒体层上的用户社区簇, M_2 表示在地理位置层的位置簇, v_{ij} 表示用户社交关系的邻接矩阵元素, E_1 表示 M_1 层中社区簇的边集, E_2 表示 M_2 层上位置簇的边集, 而 E_{12} 则表示这两层簇间的链接集合, h_i 表示社区簇内节点 i 的度, c_i 表示在 M_1 层上的社区簇 i 与 M_2 层中所有簇间的链接度, 同样地, c_j 表示在 M_2 层上的社区簇 j 与 M_1 层中所有簇间的链接度. S_i 表示在 M_1 层上节点 i 在各个簇之间的分布向量, P_i 表示在 M_2 层上节点 i 在各个簇之间的分布向量. 由式(23)可知, 综合模块度 Q_M^C 由三部分组成: 在社交媒体层上的模块度、地理位置层上的模块度, 以及这两层簇间的模块度.

除了外在结构上的度量, 兴趣内聚性是一种衡量社区划分内在特性的重要指标, 描述社区内所有关联用户的兴趣相似度定义如下:

$$IC = \frac{1}{2|E_C|} \sum_{(u_i, u_j) \in E_C} r(u_i, u_j) \quad (24)$$

式(24)中, $|E_C|$ 表示社区的边数, $r(u_i, u_j)$ 表示社区内用户对 u_i 与 u_j 间的兴趣相似性, 可由式(4)计算而得. 显然, 当 IC 值越高时, 表明社区的感兴趣内聚性越好, 社区内在结构也越紧密.

程度影响了社区结构的大小,具体在 Foursquare (NYC)数据集上,前 10 大社区的用户仅占总数的 32%.而在 Yelp 数据集中,占总数 63%的用户主要集中在前 5 大社区中,其余是大量稀疏的小社区.对于 CBGC 与 NMTF 算法而言,两者都考虑了 2 种类型的关系,获得了头部效应明显的社区,他们在 Foursquare(NYC)上分别有占总数 76%与 85%的用户集中在前 6 大社区中.在 Yelp 上的实验结果也类似,两种算法都有 80%以上的用户都分布在前 7 大社区中.相比之下,由于本文的 CCAM 算法全面地考虑了 3 种类型的关系,因而包含更完整、更真实的社区特征,我们在两种数据集上所获得的兴趣主题簇表现出更强的不平衡性,说明在一些热点主题上容易聚集大量用户,而在地理位置簇上的用户分布与 CBGC 算法的结果基本类似,但在地理位置上分布的用户群体相对均匀些.

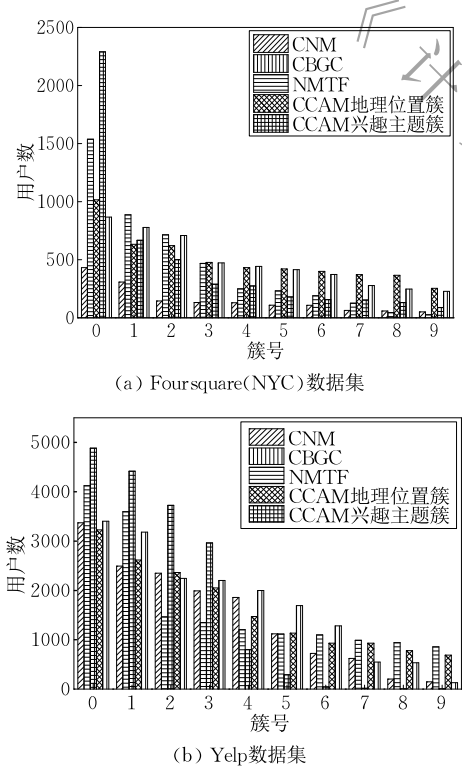


图 9 Top-10 社区的用户规模统计

图 10 显示了图 9 中各算法所对应的社区用户平均度分布情况.由图 10 可知,CNM 算法在两种数据集上所得的社区用户平均度都较低,绝大多数在 5 以内,表明社区内部的密度非常稀疏.较高的 CBGC 与 NMTF 算法在 Foursquare(NYC)数据集上的用户平均度大致相近,而在 Yelp 数据集上,后者表现出更高的社区密度.CCAM 算法在 Foursquare(NYC)上所获得的兴趣主题簇与地理位置簇中用户平均度

分布都比较均匀,整体上比 CBGC 与 NMTF 更高一些,这表明本文的 CCAM 算法获得的双重社区结构内部联系更紧密一些.而在 Yelp 上,CCAM 的地理位置簇大多都具有最高的社区密度特征,即使在用户数较少的簇上也拥有较高的平均度.对于 CCAM 的兴趣主题簇,只有前 4 簇拥有较高的平均度,有较明显的社区密度特征,这是因为大部分用户都集中在头部的兴趣主题簇中,而其余较小的兴趣主题簇上用户间联系比较稀疏.实验结果表明,融合了多维关系的 CCAM 算法所获得的双重社区簇总体上表现出最高的社区密度特征.

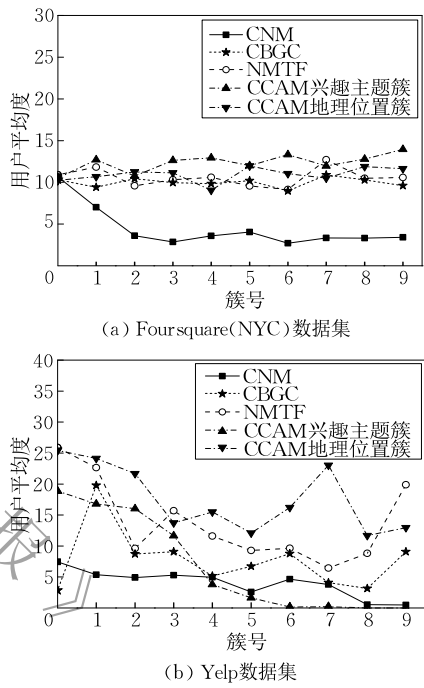
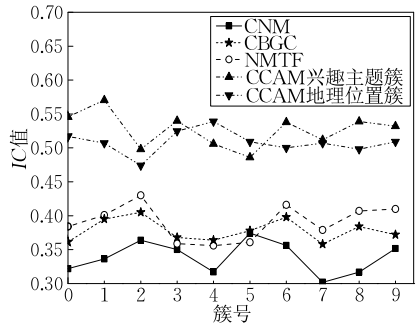
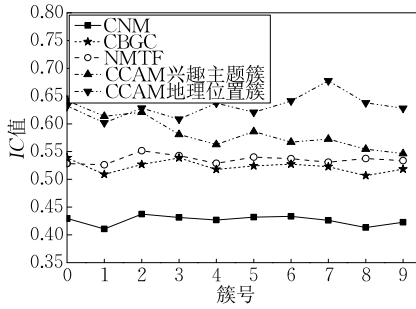


图 10 Top-10 社区内用户平均度

为了评价社区内在的兴趣内聚性,本文比较了 4 种算法在两个数据集上获得 Top-10 社区的 IC 指标,结果如图 11 所示.不难看出,在 Foursquare (NYC)中,CCAM 算法所得的兴趣主题簇上 IC 值比地理位置簇略高,这两种簇的兴趣相似度都普遍高于 0.5,都明显都优于其他 3 种算法,CNM 算法由于仅考虑单维的社交关系而导致社区兴趣内聚性最低,而 CBGC 与 NMTF 算法都考虑了不同的二维关系,两者整体的兴趣内聚性差距不大.在 Yelp 中,CNM 算法的各社区兴趣相似度虽平稳地维持在 0.4 左右,但低于 CBGC 与 NMTF 算法至少 10%以上,相比而言,CCAM 算法的双重社区簇一直都具有最高的兴趣 IC 值.通过上述实验,我们得到的结论是 CCAM 算法获得的双重社区簇不仅表

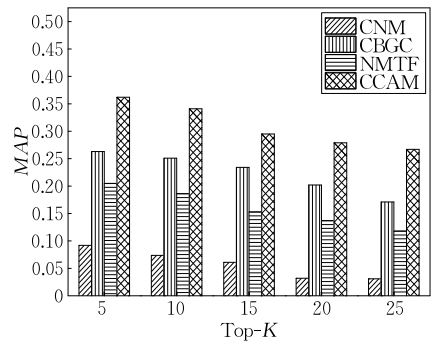


(a) Foursquare(NYC)数据集

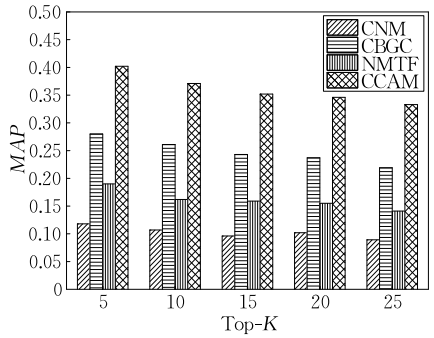


(b) Yelp数据集

图 11 Top-10 社区的兴内聚性



(a) Foursquare(NYC)数据集



(b) Yelp数据集

图 12 线上兴趣主题推荐性能

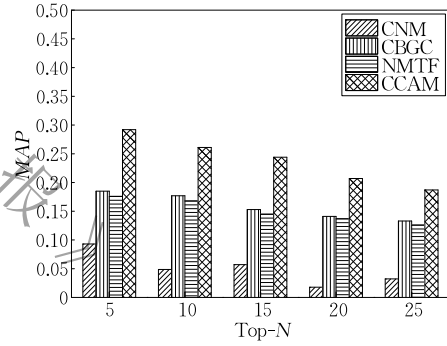
现出紧密的外在结构特征,还有最好的兴趣内聚性.

(3) 基于社区的推荐应用性能比较

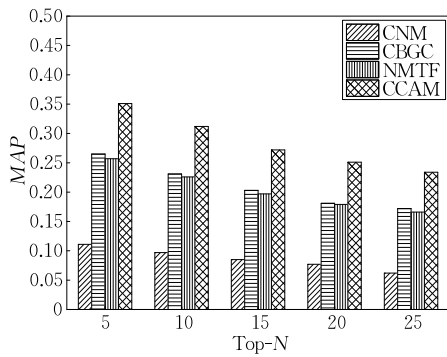
为了分析社区结构对推荐性能的影响,我们将 Foursquare(NYC)和 Yelp 数据集分别按照 4:1 比例分成训练集和测试集,然后比较 4 种算法所得的社区结构在线上兴趣主题推荐和线下位置兴趣点推荐应用中的平均准确率(MAP)指标,MAP 取值在 0 到 1 之间,当该值越大时,说明推荐效果越好.

图 12 给出了不同 Top-K 兴趣主题推荐下各算法所对应的 MAP 结果.从图 12 可以看出,CNM 算法仅靠单维社交关系形成的社区进行兴趣主题推荐时表现较差,其 MAP 值在两个数据集上都没有超过 0.12,而考虑两维关系的 CBGC 比 NMTF 算法表现更好,这是由于 CBGC 在社区聚类中正好考虑了社交媒体关系从而具有较好的推荐效果.相比之下,CCAM 算法的 MAP 值在两个数据集上都明显高出其他算法 40%左右,原因是该算法不仅融合了更多关系,并且在社交媒体层上获得了基于模糊聚类的重叠社区,这有助于提高推荐准确率.

图 13 显示了不同 Top-N 位置兴趣点推荐下各算法的 MAP 值.与图 12 相比,由于线下位置具有较大的稀疏性,图 13 中 4 种算法的兴趣点推荐性能都没有兴趣主题推荐的效果好.从图 13 中可以看出,在两个数据集上,CBGC 与 NMTF 算法的兴趣点推荐精度基本相似,两者都明显优于 CNM 算法,



(a) Foursquare(NYC)数据集



(b) Yelp数据集

图 13 线下位置兴趣点推荐性能

这是因为他们在社区聚类中都考虑了用户签到关系数据.综合来看,本文的 CCAM 算法比 CBGC 和 NMTF 算法的 MAP 值至少高 32%以上,其原因是该算法不仅考虑用户签到关系,还与社交媒体层的

兴趣主题簇具有紧密的关联性,因而能更准确地发现用户线下位置的兴趣点需求。总之,上述实验结果表明,融合多维关系的社区结构对改善推荐性能有较大的影响,本文提出的 CCAM 算法所获得的双重社区簇结构同时在兴趣主题推荐与兴趣点推荐方面表现出最好的推荐效果。

5 总 结

本文针对 LBSN 这种新型异质网络中的社区结构开展研究,提出一种充分融合多维关系的双重社区聚类与关联方法 CCAM。具体地,我们先将 LBSN 建模成由社交媒体、用户和地理位置等实体构成的三层异质网络模型,然后针对社交媒体层,通过信息熵度量用户发布主题之间的相似性,进而再将相似用户兴趣聚类问题转化求解基于模糊聚类的目标函数以获得重叠的兴趣主题簇结构。再次,在地理位置层中,将用户-位置签到关系网络转化为超图模型,并采用超边聚类方式得到用户关于地理位置的兴趣点特征簇。最后,在兴趣主题簇与地理位置簇之间借助中间用户层的社交关系建立这两层异质簇间的关联性表示模型,并通过随机梯度下降法求解模型的局部最优解。在两个 LBSN 数据集上的实验结果表明,本文提出的 CCAM 方法能有效地融合 3 种维度的异质关系,所获得的紧密关联的双重社区簇不仅在结构特征与兴趣内聚性指标上优于其他算法,并且还在线上兴趣主题推荐与线下位置兴趣点推荐方面都具有最高的推荐能力。

在未来工作中,我们将进一步优化异质社区的聚簇参数,使得多层社区簇能自动呈现最佳的关联性,以后还将考察 LBSN 中异质社区结构的动态演化规律。

参 考 文 献

- [1] Muhammad A J, Muhammad S Y, Siddique L, et al. Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, 2018, 108: 87-111
- [2] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structures of complex networks in nature and society. *Nature*, 2005, 435(7043): 814-818
- [3] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 2007, 76(3): 036106
- [4] Xie Jierui, Szymanski Boleslaw K, Liu Xiaoming. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process//*Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW'11)*. Vancouver, Canada, 2011: 344-349
- [5] Liu Shi-Chao, Zhu Fu-Xi, Gan Lin. A label-propagation-probability-based algorithm for overlapping community detection. *Chinese Journal of Computers*, 2016, 39(4): 717-729(in Chinese)
(刘世超, 朱福喜, 甘琳. 基于标签传播概率的重叠社区发现算法. *计算机学报*, 2016, 39(4): 717-729)
- [6] Lu Meilian, Zhang Zhenglin, Qu Zhihe, et al. LPANNI: Overlapping community detection using label propagation in large-scale complex networks. *IEEE Transactions on Knowledge and Data Engineering*, 2019, 31(9): 1736-1749
- [7] Shen Huawei, Cheng Xueqi, Kai Cai, et al. Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, 2009, 388(8): 1706-1712.
- [8] Lee C, Reid F, McDaid A, et al. Detecting highly overlapping community structure by greedy clique expansion//*Proceedings of the 4th SNA-KDD Workshop (SNA-KDD'10)*. Washington, USA, 2010: 33-42
- [9] Zhang Zehua, Miao Duoqian, Qian Jin. Detecting overlapping communities with heuristic expansion method based on rough neighborhood. *Chinese Journal of Computers*, 2013, 36(10): 2078-2086(in Chinese)
(张泽华, 苗夺谦, 钱进. 邻域粗糙化的启发式重叠社区扩张方法. *计算机学报*, 2013, 36(10): 2078-2086)
- [10] Ahn Y Y, Bagrow J, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*, 2010, 466(7307): 761-764
- [11] Shi Chuan, Cai Yanan, Fu Di, et al. A link clustering based overlapping community detection algorithm. *Data & Knowledge Engineering*, 2013, 87: 394-404
- [12] Zhou X, Liu Y H, Wang J, et al. A density based link clustering algorithm for overlapping community detection in networks. *Physica A: Statistical Mechanics and Its Applications*, 2017, 486: 65-78
- [13] Loe C W, Jensen H. Comparison of communities detection algorithms for multiplex. *Physica A: Statistical Mechanics and its Applications*, 2015, 431: 29-45
- [14] Tang L, Wang X, Liu H. Community detection via heterogeneous interaction analysis. *Data Mining and Knowledge Discovery*, 2012, 25(1): 1-33
- [15] Tagarelli A, Amelio A, Gullo F. Ensemble-based community detection in multilayer networks. *Data Mining and Knowledge Discovery*, 2017, 31(5): 1506-1543
- [16] Liu X, Liu W, Murata T, et al. A framework for community detection in heterogeneous multi-relational networks. *Advances in Complex Systems*, 2014, 17(6): 1-21

- [17] Pramanik S, Tackx R, Navelkar A, et al. Discovering community structure in multilayer networks//Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics. Tokyo, Japan, 2017: 611-620
- [18] Gao B, Liu T, Zheng X, et al. Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering//Proceedings of the 11th ACM SIGKDD (KDD'05). Chicago, USA, 2005: 41-50
- [19] Mei J, Chen L. A fuzzy approach for multi-type relational data clustering. IEEE Transactions on Fuzzy Systems, 2012, 20(2): 358-371
- [20] Hu L, Chan K C. Fuzzy clustering in a complex network based on content relevance and link structures. IEEE Transactions on Fuzzy Systems, 2016, 24(2): 456-470
- [21] Li Xutao, Ng Michael K, Ye Yunming. MultiComm: Finding community structure in multi-dimensional networks. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(4): 929-941
- [22] Ding C, Li T, Peng W, et al. Orthogonal nonnegative matrix tri-factorizations for clustering//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 126-135
- [23] Pei Y, Chakraborty N, Sycara K P. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks//Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15). Buenos Aires, Argentina, 2015: 2083-2089
- [24] Comar P, Tan P N, Jain A K. A framework for joint community detection across multiple related networks. Neurocomputing, 2012, 76(1): 93-104
- [25] Ma Xiaoke, Dong Di, Wang Quan. Community detection in multi-layer networks using joint nonnegative matrix factorization. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(2): 273-286
- [26] Wang Z, Zhang D, Zhou X, et al. Discovering and profiling overlapping communities in location-based social networks. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2014, 44(4): 499-509
- [27] Zhao Y, Chen Q, Yan S, et al. Detecting profilable and overlapping communities with user-generated multimedia contents in LBSNs. ACM Transactions on Multimedia Computing, Communications, and Applications, 2013, 10(1): 1-22
- [28] Clauset A, Newman M, Moore C. Finding community structure in very large networks. Physical Review E, 2004, 70(6): 066111



GONG Wei-Hua, Ph. D., associate professor, M. S. supervisor. His research interests include social network, machine learning.

SHEN Song, M. S. candidate. His research interests include social network, recommending system.

PEI Xiao-Bing, Ph. D., professor, M. S. supervisor. His research interests include machine learning, data mining.

YANG Xu-Hua, Ph. D., professor, Ph. D. supervisor. His research interests include network science, machine learning.

Background

Nowadays location based social networks (LBSN) becomes more and more popular and have attracted much attention from researchers. As one typical kind of heterogeneous information networks, LBSN not only has multi-mode entities such as users, locations and social media, but also contain many multi-dimensional relationships, e. g. user social relations, user-location check-in relations, and user-media tag relations. Examples of LBSN include Yelp, Facebook and so on. In such heterogeneous networks, how to detect complicated communities is an important and fundamental problem. However, presently most of the traditional community detection methods in social networks are almost based on

homogeneous network structures, only considering single dimensional relation or entity. Although some recent studies try to fuse multi-relations, they still cannot efficiently integrate the multi-mode entities and their heterogeneity relations contained in LBSN.

To address the limitations of existing methods, this paper provides multi-layer heterogeneous communities detection and association method with multi-dimensional relations fused in LBSN, which is called CCAM. This method can not only accurately embody the topics of user interest preferences in social media dimension, but also really reflect the user's distribution for points of interest clusters in geographic space

dimension. The main idea of CCAM is below: firstly, in the upper social media layer of LBSN, fuzzy clustering is proposed to obtain different overlapped user interest communities. Then, in the bottom geographical location layer, the users and locations and their relationship composed of bipartite graph are transformed into hyper graph structures, and the user's interest clusters about geographical location is obtained by using hyper edge clustering method. Finally, the upper user topic communities and the lower geographic location clusters are established the correlating representation model between the two layers by the social relations among

users in the middle user layer of LBSN. The extensive experimental results on two real datasets show that our proposed method in this paper outperforms the other state-of-art methods and can also effectively mine dual associated communities between online social space and geographical space.

This work is supported by the Natural Science Foundation of Zhejiang Province (Nos. LY13F020026, LQ18F020008), the National Natural Science Foundation of China (Nos. 61502420, 61802346), and the China Postdoctoral Science Foundation (No. 2015M581957).

《计算机学报》