学会优博获奖论文精华版

# 面向地理社会媒体的挖掘与应用

方 全 / 1. 中国科学院 2. 中国科学院自动化研究所

摘 要：本文针对目前海量的地理社会媒体数据与用户需求的信息之间存在“知识鸿沟”问题，面向地理社会媒体的挖掘与应用开展深入研究。特别地，把语义理解、知识挖掘与应用服务结合起来，贯穿到研究的问题中，具体做了三个方面的研究工作。① 地理位置计算。以地理位置为中心，旨在利用带地理标签的媒体数据，结合地理位置信息和多媒体语义，进行地理位置的建模和知识挖掘。分别提出了一种基于区域隐式支持向量机模型框架和系统性的可视化方法框架和概率图模型。② 用户理解。以用户为中心，利用用户产生的大量的媒体内容数据和丰富的用户网络行为数据进行用户分析建模。分别提出了一种关联性隐式支持向量机模型框架和基于超图学习的方法框架。③ 用户与地理位置结合的建模分析。从地理社会媒体数据挖掘得到的地理位置和用户的知识，需要通过有价值的应用送达给终端用户，满足用户的需求。针对用户在移动场景下的地理位置评分行为的特性，提出了一种场景感知回归混合模型对时空场景信息、用户兴趣、地理区域偏好、地理物品与内容进行统一建模并实现推荐。分别在真实数据集上测试，验证了所提方法框架的有效性，并进行了多种面向用户的应用，证明了地理社会媒体挖掘的实用性。

关键词：社会媒体；用户理解；知识挖掘；可视化；识别预测；探索发现；个性化服务

## 1. 引 言

信息技术及互联网的发展，尤其是移动互联网的兴盛，正在深刻地影响改变着人们的生活。社会媒体，一种新型的允许人们创造并分享媒体信息的工具和平台，在近年来得到了飞速的发展，吸引着全球数以亿计的用户参与其中。伴随社会媒体的兴盛，随着地理位置定位技术的发展，基于用户地理位置的服务（Location Based Service, LBS）成为主流应用。用户通过移动设备的 GPS、WiFi、通信基站等方式获取地理位置信息使用各种各样的服务。社会媒体和地理位置的结合形成了地理社会媒体（Georeferenced Social Media）。地理社会媒体涵盖各种形式的带地理位置特征的社会媒体网站和服务。地理社会媒体使得用户可以随时随地的获取和分享信息，产生了海量的带地理位置信息的社会媒体内容数据，并且此类数据的规模呈爆炸性增长。地理社会媒体具有多模态、数据异质、大规模、空时性等特点，根据这些特点，如何对其涵盖的数据进行有效地挖掘和利用，从数据中提取知识进行服务，成为未来互联网应用和发展的关键。地理社会媒体包含了地理位置、用户、数据三个重要的元素，三个元素相互联系作用。一方面，用户产生大量的带地理位置标签的媒体数据，用户在这一数据产生过程中充当传感器感知该地理位置区域，因此通过汇聚挖掘这些带地理位置标签的媒体数据可以理解相应的地理位置区域；另一方面，用户产生大量的在线交互活动行为数据，与媒体内容数据结合研究可以分析理解用户。地理社会媒体研究目的是从用户感知的地理位置数据和社交内容挖掘知识用于理解用户和地理位置，从而进行有价值的服务。

图 1 给出了面向地理社会媒体的挖掘与应用的研究框架，包括数据感知获取、地理位置与用户知识挖掘、面向用户的基于地理位置的应用服务提供三个环节。数据感知获取指的是获取关于物理世界的动态韵律和人们在线上线下行为活动状况，以各种形式的数据格式记录存储到网络空间。数据感知主要依靠大量传感器和设备（手机、定向传感器、车辆、人等）不断地自动运行感知物理世界来完成，人作为传感器智能灵活地感知线下世界而产生多媒体内容数据是一种重要数据感知方式。此外，用户在网络空间产生丰富的网络行为活动，比如发表评论、收藏和交友互动等。这些丰富的线下物理感知数据与用户在线行为数据组成了地理社会媒体数据。

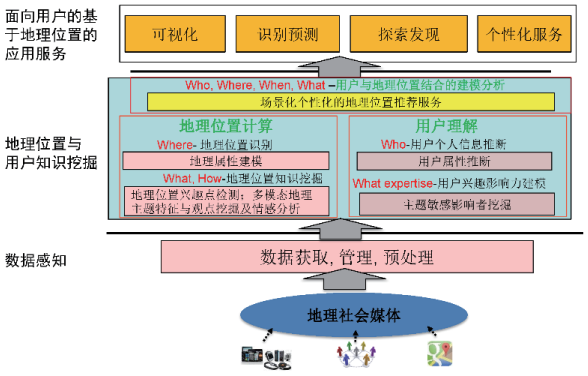



图 1 面向地理社会媒体的挖掘与应用的研究框架

在获取地理社会媒体数据后，对地理位置和用户进行建模分析、挖掘知识，以及理解地理位置和用户。这一阶段是面向地理社会媒体研究的核心环节。包括三个方面：① 以地理位置为对象，对带地理位置标签的数据进行挖掘分析，把多媒体语义理解与地理位置信息结合，提取地理位置知识，研究问题包括地理位置的建模识别、主题挖掘、观点挖掘及情感分析。地理位置的建模识别包括位置兴趣点（POI）的挖掘以及利用多媒体数据进行地理位置推断和定位，据此提出一种基于区域隐式支持向量机模型框架来挖掘一个地理区域中，具有代表性和判别性的地理属性帮助进行地理位置识别。此外，对挖掘的地理属性做了语义解释，可用于地理位置探索服务。地理位置主题挖掘指从带地理标签的数据中挖掘出地理位置上的兴趣主题（如旅游景点、购物、吃饭等）。由此提出一种自增量学习的地理兴趣点主题发现算法，其利用视觉文本信息能够自动地发现地理兴趣点的多个潜在主题，并进行可视化。地理位置观点挖掘及情感分析指的是从网络媒体数据中，挖掘地理位置实体的多模态主题特征及相对应的观点及情感极性，据此提出一种概率图模型来利用实体的多源媒体数据自动挖掘出实体的多模态主题特征与对应的观点，这样有效地丰富了地理实体的知识图谱维度。此外，对挖掘的观点做了情感分析，并联合主题特征观点进行了实体关联可视化。由挖掘的多模态主题特征与观点，设计实现了检索任务的应用。② 以用户为对象，挖掘用户产生的媒体内容数据和网络行为进行用户理解，具体解决的问题包括用户的属性推断和用户影响力分析。我们系统性地研究了性别、年龄、情感状况、职业、兴趣、情绪倾向六种用户属性，提出了一种关联性隐式支持向量机模型框架，并利用用户产生的媒体内容特征和属性关系进行用户属性推断。所提模型框架可以用于用户画像和基于属性的用户检索。同时提出一种基于超图学习的方法框架，并利用社会媒体网络中用户产生的媒体内容与链接关系挖掘主题敏感影响者。挖掘的主题敏感影响者可用于好友推荐和媒体信息推送服务。③ 用户与地理位置结合的建模分析，从地理社会媒体数据挖掘得到地理位置和用户的知识，需要把这种知识通过有价值的服务送达给终端用户，满足用户的需求。在这一方向，我们研究场景化个性化的地理位置探索推荐系统，对用户、场景地理信息、地理位置物品进行统一的建模，挖掘知识，为用户提供推送服务并帮助其进行地理位置探索。地理位置计算和用户建模挖掘得到的知识，可以很好嵌入到推荐服务系统中。针对用户在移动场景下的地理位置评分行为的特性，提出一种场景感知回归混合模型，对时空场景信息、用户兴趣、地理区域偏好、地理物品与内容进行统一建模并实现推荐。联合提出的模型，提出了一种用户意图和地理位置感知的概率矩阵分解模型，可以融入地理位置与用户的知识，有效地缓解了数据稀疏性问题从而提高了推荐性能，促使用户能够进行当地和新地点的探索发现。

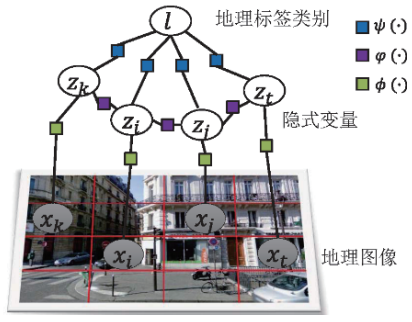
在获取关于地理位置和用户的知识基础上，设计了提供面向用户的应用服务，包括知识可视化展示、识别预测、媒体内容信息探索发现和面向用户的个性化服务等。这一阶段是面向地理社会媒体研究的重要环节。

## 2. 研究内容

### 2.1 地理位置计算

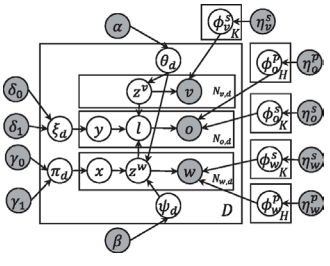
2.1.1 基于地理位置属性的地理位置识别

大规模带地理位置信息的媒体数据的出现，使得从媒体内容信息估计其地理位置属性成为可能。研究者们做了大量的工作来研究利用海量的带地理位置标签的图像数据进行地理位置识别。现有的基于媒体内容识别地理位置方法分成数据驱动的方法 [1] 和基于模型的方法 [2] 两类。数据驱动的方法采用相似性匹配简单有效，但是扩展性有限；基于模型的方法泛化识别，但缺乏解释性。因此有必要结合数据驱动和基于模型的方法来进行地理位置识别。通常来说，一个地理区域的图像包含一些独特的模式，其能够有效地帮助地理位置识别。如一些城市中典型独特的建筑元素（如屋檐、窗户等）可以帮助判断图片的地理位置属性。我们将挖掘城市区域的这些独特地理模式称作为地理属性，其需要满足两方面性质：① 判别性，它们能帮助区分出这个地理位置区域；② 代表性，它们在这个地理位置区域频繁存在并且是语义性可解释的。为挖掘一个地理区域的地理属性，对该地理区域的图像进行判别性分析，并提出一个基于图像区域隐式支持向量机模型，如图 2(a) 所示。对地理图像进行建模，每块图像区域被赋予一个隐变量来指示该区域是否对帮助识别该地理区域有帮助，即是否有地理判别性。RLSVM 通过在一个地理区域图像上建模学习，能得到每个图像区域是否地理判别性；然后选择有地理判别性的区域并聚类生成地理模式簇类作为地理属性；最后利用 Flickr 上用户标注的文本标签并提出区域标签相关性学习方法，对所挖掘的地理属性进行标注，使得挖掘的地理属性具有语义可解释性。在收集来自 GoogleStreetView 和 Flickr 上的城市地理图像数据库上，验证了所挖掘的地理属性能够有效地提升地理位置识别的效果。




(a) 基于区域的隐式支持向量机模型

(b) 多主题新加坡可视化



(c) mmAOM 的概率图模型

图 2 地理位置计算

2.1.2 用户感知的多主题地理位置可视化

通常用户到一个新的城市或者旅游景点，用户关心两个问题，一是有什么游玩兴趣点（Point Of Interest，POI）；二是各个兴趣景点的特色内容是什么？在本工作中，将挖掘用户产生的地理社会媒体数据来层次化多主题可视化城市，相比以前的工作 [3]，更进一步地理性和语义性来组织管理地理媒体数据。主题在这里指的是一个地理区域的特定有意思的内容或者代表性的模式。自然而然，一个地理区域具有多个可视化的主题。另一个重要的概念是地理兴趣点（Point of Interest, POI），其指的是热门拍照的地点并用质心来表示。我们提出三层次的方法框架。输入是一个可视化目标地理区域的图像数据，其附带有拍摄地点、文本元数据和账号用户信息。所提出的方法框架包括地理兴趣点主题发现、地理区域主题聚合和可视化三部分。由于只有部分图像具有地理坐标标签信息，首先利用文本标签和图像内容信息来推断缺失 GPS 信息的图像的所属地理兴趣点。地理兴趣点主题发现是核心部分，因此提出一个自增量式学习的算法来挖掘地理兴趣点多个主题。对于地理区域主题聚合和可视化，采用相似性主题的聚类方法实现。给出挖掘的地理兴趣点和地理主题，可以很容易地可视化一个地理兴趣点和进行地理区域。在关于新加坡的

Flickr 图像上进行了实验，实现了多主题新加坡可视化，如图 2(b) 所示。主客观评价验证了所提出可视化方法的有效性，证明了多主题可视化的潜在应用价值。

2.1.3 地理实体的多模态主题特征观点挖掘与情感分析

随着互联网的发展普及和社会媒体服务的兴盛，人们在网络上可以便捷地获取和分享丰富的社会多媒体信息。其结果是，社会媒体平台上聚集了海量的人们对物理实体的评论和情感信息。从大规模的用户生成内容中挖掘实体的主题观点和分析情感是知识挖掘中的重要任务。已有的主题特征观点挖掘的工作主要集中在文本内容处理上 [4]。在多媒体上来挖掘主题特征观点，目前还鲜有研究工作。实际上，一个地理实体的很多主题特征方面都是多模态表达的。比如，对于北京，观察到的地标和雾霾不仅能用文本来表达，还能很具体地用视觉图像来描述。我们称这样的主题特征具有视觉表达性。这种主题特征含有清晰和具体的视觉对应形态。同时，实体的一部分主题特征没有清晰和具体的视觉对应，例如经济、工业等。这样的主题特征不具有视觉表达性，其用文本描述而很难用视觉具体内容来表达。通过对实体的多模态主题的视觉表达性进行建模，并挖掘相应的主题特征及观点情感，能够更好地理解目标实体。本文研究从丰富的地理社会媒体数据中，挖掘一个地理实体的多模态主题特征及对应的观点情感。如图 2(c) 所示，我们形式化地理实体的多模态主题特征和观点挖掘为：输入是一个实体的相关多媒体文档，包括 Flickr 图像、Tripadvisor 评论和新闻文档。换言之，输入文档可以是一张图像、一篇新闻文档或一条评论。文档由视觉和文本特征词以及观点组构成。我们提出一个生成式概率图模型——多模态主题观点挖掘模型（multimodal Aspect-Opinion Model, mmAOM，如图 2(c)）来推断输出。mmAOM 对主题特征和观点词在文档的生成过程进行建模而学习文本和视觉模态之间的关联关系，来区分有视觉表达性的主题特征和非视觉表达性的主题特征，以及主题特征和观点之间的依赖关系来辨别主题特征及相应的观点。模型输出包括学习到的多模态主题特征、文档的主题分布、主题特征对应的观点。由派生的地理实体的多模态主题特征和对应的观点，设计了实体关联可视化和多模态主题特征检索的应用。实体关联可视化是要简洁地在图谱上可视化出实体关联的重要主题特征和对应的用户观点情感。多模态主题特征检索利用主题与观点之间的关联关系进行跨模态观点检索的任务。我们在真实的实体对象数据集中进行实验评价 mmAOM。除了在地理实体对象（北京、伦敦、巴黎、纽约）上实验，也在其他实体做了实验评测，包括人物（纳尔逊曼德拉、史蒂夫乔布斯）和品牌（阿迪达斯、耐克）。实验的结果证明了提出的 mmAOM 模型在挖掘实体多模态主题特征和观点的有效性，以及在可视化和检索方面的实用性。

## 2.2 用户理解

2.2.1 关联性用户属性推断

用户画像是个性化推荐的基础。在大多数的社交网络中，并不能得到准确和完整的用户属性。现有的工作通过利用用户产生的网络数据来推断用户属性 [5]，表明了用户网络行为活动对于预测用户属性的有效性。但是这些研究工作是独立研究用户属性的，并且只利用了文本内容。本文利用用户在社会媒体网络中产生的丰富在线多媒体内容信息和用户属性之间的关系，研究关联性用户属性推断。特别地研究性别、年龄、情感状况、职业、兴趣、情绪倾向六种类型的用户属性，每一种用户属性有多个值。为了能有效利用用户产生的内容信息和属性之间的关系，提出一个关联性隐式支持向量机（Relational Latent SVM, Relational LSVM）的模型框架进行用户属性推断。特别地，以 Google+ 为测试平台进行研究和实验。在 Google+，用户允许建立他们的个人档案在 About 版块和在 Posts 页面发布个人活动信息。我们把关联性用户属性推断形式化为输入是用户的社交网络内容数据，包括从 About 获取的档案数据和在 Posts 下载的发布活动信息。关联性隐式支持向量机从这些数据中监督性学习来推断输出，包括预测的用户属性和推断的用户属性关系。我们在来自 Google+ 的真实数据上评估提出的属性推断模型。实验结果证实了关联性隐式支持向量机推断用户属性的有效性和用户属性关系在用户相关应用中的用途。

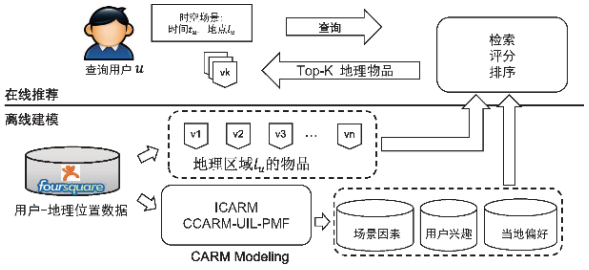
2.2.2 主题敏感影响者的挖掘

社会媒体网络的出现和快速流行为用户提供了

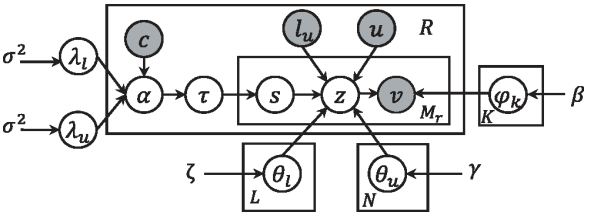
一个创造和分享兴趣内容的交互分享平台。最近，社会影响力分析已吸引了研究者充分的兴趣。相当量的工作已进行来验证影响力的存在[6]，或者在同质网络中的影响力建模[7]。但是，鲜有工作研究包含多模态兴趣内容的社交网络中主题敏感影响力量化的问题。本文探究在基于兴趣的社会媒体网络中的主题敏感影响者挖掘（Topic-Sensitive Influencer Mining, TSIM）的问题。TSIM 旨在挖掘网络中主题特定的有影响力的顶点。我们以最流行的图片分享网站之一的 Flickr 为测试平台，进行研究和实验；使用视觉文本的内容关系构建同质超边，用于主题学习和使用社交链接关系构建异质超边用于网络中影响力排序。所提出的解决 TSIM 的方法框架，主要包括超图学习、兴趣主题学习和主题敏感影响力排序三种学习阶段。首先，一个统一超图构建来对 Flickr 种的用户、图像和多种类型关系进行建模。在图像之间的视觉 - 文本信息用于构建同质超边。用户和图像之间的社交链接关系用于产生异质超边。其次，由于图像的稀疏社交链接的影响，有信息性的标签图像选取并在超图异质超边上通过超图正则化主题模型学习主题空间。我们通过协同表示相似性传播来获得所有的图像和用户的主题分布。最后，一种基于相似性传播的超图排序算法，在超图超边上运行获得用户和图像的特定主题的社会影响力得分。在从 Flickr 收集的真实数据上的实验验证了所提方法挖掘主题敏感影响者的有效性。

### 2.3 场景化个性化的地理位置推荐系统

基于地理位置的社会媒体网络服务的出现，例如 Foursquare、Facebook Places 和大众点评，为人们提供了一个产生和分享在物理位置进行评价的活动的便捷平台。全面地理解这种基于地理位置的用户评分行为对于进行很多应用十分重要，例如个性化推荐、地理位置探索和服务营销。文献 [8] 已经做了很多努力进行从用户评分历史数据中挖掘知识帮助用户找到有兴趣的地理物品。但是，利用用户的地理位置行为历史数据推断地理物品的评分进行推荐仍是一个具有挑战性的问题，包括数据稀疏性、用户内在兴趣与地理特色影响、时空性影响。本文通过探究用户地理位置评分行为研究场景化个性化的地理位置推荐问题。具体而言，提出了一个场景化个性化的地理位置推荐系统（context-aware personalized location recommendation system, CAPLRS），其利用用户与地理物品之间的关联关系、地理物品的地理位置与内容信息和时空场景信息来缓解数据稀疏性问题，并做出准确的地理位置推荐。如图 3(a) 所示，CAPLRS 由离线建模与在线推荐两部分组成。离线部分的核心模块是一个场景感知回归混合模型（context-aware regression mixture model, CARM, 如图 3(b) 所示），其设计用于对用户地理位置评分行为进行建模用于推断用户对地理物品的评分。CARM 通过同时考虑用户内在兴趣、地理区域偏好和时空场景影响，在一个统一的模型框架内，对用户在地理物品上的决策行为过程进行建模。CARM 能够自动地从用户的地理位置评分历史数据中，学习得到潜在主题、用户兴趣、地理区域偏好和场景影响因素。给定一个查询用户以及其对应的查询场景信息，即地理位置区域和时间节点，在线推荐部分为在地理区域的每一个地理物品计算一个排序分数。CARM 通过自动合并 CARM，离线学习得到的场景影响因素、用户的兴趣和地理区域的偏好进行推荐。我们在两个真实来源于 Dianping 和 Foursquare 的数据集进行了充分的实验评估提出的推荐系统性能。实验结果显示，所提的推荐系统 CAPLRS 在推荐效果和效率上的优越性。此外，实证分析结果也展示了 CAPLRS 有清晰的意义解释性，这对于增强人们对推荐系统的信任十分重要。




(a) 推荐框架图




(b) CARM 模型

图 3 场景化个性化的地理位置推荐系统

## 3. 结束语

本文在面向地理社会媒体的挖掘与应用的研究中，把语义理解、知识挖掘、应用服务结合起来贯穿到研究问题中，包括地理位置计算、用户理解，以及用户与地理位置结合的建模分析，做了一些有益的尝试和探索，取得了初步的研究成果。随着移动互联网的渗透式发展和数据的不断爆炸性增长，面向地理社会媒体的挖掘与应用研究面临更多机遇和挑战，有大量的亟待深入研究和解决的问题。例如，地理社会媒体计算是一个交叉性的课题，涉及到众多领域知识和技术手段，在数据感知获取、数据存储管理、数据挖掘分析算法、数据挖掘知识应用等方面有大量的问题需待继续研究；建立统一完整的地理社会媒体知识库，并探索智能系统应用服务； 探索研究面向地理社会媒体计算的理论框架， 帮助设计算法系统来处理大规模地理社会媒体数据。

参考文献


[1] J. Hays and A. A. Efros, “IM2GPS: estimating geographic information from a single image,” in CVPR 2008, 24-26.

[2] D. J. Crandall, L. Backstrom, D. P. Huttenlocher, and J. M. Kleinberg,“Mapping the world’s photos,” in Proceedings of the 18th InternationalConference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009, pp. 761-770, 2009.

[3] Y. Zheng, Z. Zha, and T. Chua, “Research and applications on georeferenced multimedia: a survey,” Multimedia Tools Appl., vol. 51, no. 1, pp.77-98, 2011.

[4] B. Liu and L. Zhang, “A survey of opinion mining and sentiment analysis,”in Mining Text Data, pp. 415-463, 2012.

[5] A. Mislove, B. Viswanath, P. K. Gummadi, and P. Druschel, “You are who you know: inferring user profiles in online social networks,” in WSDM, pp.251-260, 2010.

[6] A. Anagnostopoulos, R. Kumar, and M. Mahdian, “Influence and correlation in social networks,” in KDD, pp. 7-15, 2008.

[7] J. Tang, J. Sun, C.Wang, and Z. Yang, “Social influence analysis in large scale networks,” in KDD, pp. 807-816, 2009.

[8] J. Bao, Y. Zheng, D. Wilkie, and M. F. Mokbel, “A survey on recommendations in location-based social networks,” Geo Informatica, November 2014.



### 方 全

中国科学院博士毕业，师从许常胜教授；现为中国科学院自动化研究所助理研究员。曾获 2016 年度中国科学院优秀博士毕业论文。于多媒体与数据挖掘领域国际顶级会议和期刊上发表论文 10 多篇，包括 ACM MM、TMM、TIST、TOMM 等。获得国际顶级多媒体会议 ACM Multimedia 的最佳论文提名和国际重要多媒体会议 MMM 的最佳学生论文，以及微软亚洲研究院学者奖。主要研究方向为社会媒体数据挖掘与应用。