ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 1,Jan. 2016,pp. 18－27
DOI:10.16337/j.1004-9037.2016.01.002

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

# 基于社会媒体的旅游数据挖掘与分析

高新波　沈钧戈

(西安电子科技大学电子工程学院,西安,710071)

**摘　要**:信息技术和社会媒体的快速发展使得网上旅游数据日益庞大。随着游客对旅游需求呈现多样化和定制化趋势,旅游信息服务成为目前的研究热点。社会媒体中的社区和自媒体平台存在着海量的旅游信息资源,基于社会媒体的旅游数据挖掘可以充分利用这些资源,将其应用在智慧旅游推荐上,进而推动"互联网＋旅游信息化"的快速发展。本文全面分析和讨论了目前"互联网＋旅游信息化"的研究背景和发展历程,分析了目前社会媒体中的旅游数据的特点,进一步介绍了"互联网＋旅游信息化"背景下的热点研究应用,最后总结了基于社会媒体旅游数据的挖掘与应用的研究难点,并且对未来可能的研究方向进行了探讨。

**关键词**:社会媒体;旅游异质信息;旅游信息挖掘;旅游信息重排序;旅游信息推荐

**中图分类号**:TP391　　**文献标志码**:A

## Social Media Based Travel Data Mining and Analysis

Gao Xinbo, Shen Junge

(School of Electronic and Engineering, Xidian University, Xi′an, 710071, China)

**Abstract**: With the rapid development of information technology and social media, The travel information increases. The people′s travel demands and preferences are gradually diversified and customized, so tourism information service has become a hot topic among researchers. Moreover, communities and user-operated media in social media have tremendous tourism information resources which can be employed in travel information mining based on social media. Since travel information is used for smart travel applications, it promotes the fast growing of internet-based tourism informatization. In this paper, the background of internet-based tourism inforamatization is discussed. Secondly, the characteristics of travel information in social media are analyzed. Then, some hot research topics in tourism applications are presented and discussed. Finally, the challenges of social media based travel data mining and applications are summarized, and some further research directions are explored.

**Key words**: social media; travel heterogeneous information; travel information mining; travel information reranking; travel information recommendation

## 引　　言

为进一步解放和提高生产力,2015 年的总理政府工作报告中提出了"互联网＋"的经济战略布

**基金项目**:国家杰出青年科学基金(61125204)资助项目;中央高校基本科研业务费专项资金(BDZ021403)资助项目。
**收稿日期**:2015-12-05;**修订日期**:2016-01-06

局[1-3],即将互联网技术融合在传统行业中,以互联网为平台促进传统行业的快速发展。随着互联网技术在各个行业的不断渗透,旅游业也处于转型的分水岭,逐步从资源依赖型向技术驱动型转换。“互联网+旅游信息化”将促进信息技术与旅游业的融合创新,打造新的经济增长点,为智慧旅游业提供技术支撑。

“互联网+旅游信息化”是互联网平台与信息技术的结合,从旅游信息获取、挖掘、检索与推荐等都是以互联网为平台利用信息技术实现各个环节的旅游智能化。旅游数据是“互联网+旅游信息化”的基本保障,因此掌握庞大的数据信息才能实现旅游智能化。而目前人们在网络中获取旅游信息时,会遇到一些问题:旅游网站上提供的旅游信息都是静态的和局部的;不能根据用户需求进行精确信息查找;忽略用户个性化需求,不能结合用户情景上下文信息提供个性化、定制化旅游信息。而这些问题严重制约了获取景点信息,制定出行路线等用户旅游行为。因此,智能旅游服务不仅要掌握丰富的旅游信息资源,还需要对这些旅游信息进行专业化的处理、分析与挖掘。旅游数据挖掘是针对以智慧旅游为目的的特定应用,是一般数据挖掘的特殊情况,需要建立以数据为驱动的模型解决旅游应用问题。其难点就在于如何充分挖掘旅游数据的知识,并针对其特点高效地实现基于社会媒体的旅游数据挖掘与分析。

为了高效地分析数据建模,研究人员利用数据挖掘、机器学习以及统计学习等知识进行应用构建。尽管基于“互联网+旅游信息化”的智能旅游应用是一个新兴的研究课题,但是随着数据挖掘以及机器学习等基础研究领域的发展,基于社会媒体的旅游数据挖掘已经成为一个逐渐升温的研究课题。旅游数据挖掘是指从海量的社会媒体数据中提取旅游相关的知识、景点间的关系以及基于用户喜好的景点分类等,用于实现智能旅游信息服务。基于社会媒体的旅游数据挖掘应用的研究主要分为两类:(1)分析景点的描述信息并挖掘相关知识,如实现景点的三维立体重建,景点的摘要生成;(2)根据景点的信息关联,结合用户需求,实现旅游信息的检索与个性化推荐等。

由于社会媒体的旅游数据不仅具有网络数据的性质:冗余性和智慧性,还具有旅游数据特有的性质:多样性和异质性。传统的数据挖掘方法并不能完全解决基于社会媒体的旅游数据挖掘的问题,所以这引起了研究者的关注。回顾近期文章,主要关注于以下 3 个方面:(1)旅游数据挖掘;(2)旅游信息重排序;(3)旅游信息推荐。本文介绍了基于社会媒体的旅游数据的特点,近期关于基于社会媒体的旅游数据挖掘和应用研究目前基于社会媒体旅游数据的挖掘与应用的研究瓶颈以及未来的研究方向。

## 1 社会媒体旅游数据

### 1.1 社会媒体平台

互联网技术的发展使得各种媒体平台[4,5]生成大量信息数据,并且社会媒体平台提供信息发布、评论、下载和分享等功能,信息量开始呈现爆发式增长,因此,用户可随意发布信息的各种社会媒体平台已经成为人们获取各种所需信息的主流渠道。社会媒体平台按照信息传播方式主要可以分为以下两类。

用户分享性质的社区平台:该类社会媒体平台主要围绕某个特色的主题,例如蚂蜂窝是以旅游服务为主题,大众点评是以生活分享为主题的网站以及全民论坛网站——天涯等;在这些网站上用户通过发布主题帖子,其他用户围绕主题帖子进行点评。因此用户根据自己的兴趣爱好等共同的喜好参与进来,因此用户可以主动设立帖子并且与其他用户进行内容交流,所以这类平台是含有基于某些主题的多媒体信息内容。

自我表达性质的自媒体平台:个人信息传播成为某些平台的主要功能,例如新浪微博,网易博客,土豆优酷视频网站以及昵图网等图片网站等,用户通过此类平台发布消息,然后不确定的人群会获取用户发布的信息。因此,此类平台促进了自媒体的发展,用户在自媒体平台盛行的时代里,每个人都是信息发布平台的主体,每个人根据自己的喜好创造出自己感兴趣内容,所以这类自媒体平台有最新的信息,并且这些信息来源广泛,视角独特,推动了社会热点不断发展及更新。

通过与社会媒体进行对比,传统媒体在空间和时间上体现出局限性。然而社会媒体在时间上具有时效性高,内容上具有信息量大、来源广泛并且互动性强等特点,因此社会媒体平台中存在的大量个性化信息适合为用户提供完整的旅游信息和高效智能的旅游服务。

### 1.2 社会媒体信息特点

社会媒体的内容主要为海量的多媒体信息,例如:

文本:文本是社会媒体信息最主要的表达方式,不仅包含文字信息,同时还包括多媒体信息的文字标签以及网站上的众多评论等。互联网上存在大量的文本信息,是互联网使用者在碎片化的时间下积累的零散文本,因此很多文本信息呈现出碎片化特点。

图片:图片是社会媒体信息最丰富的表达形式,互联网上存在大规模海量的图片信息,用户可以通过各大社交媒体平台上传图片并对图片配上文字标签等信息,通过图片信息来更直观地表达用户当时的想法。但是往往存在图片信息和文字标签内容不匹配,因此造成了语义鸿沟问题。

视频:视频是社会媒体信息最生动的表达形式,互联网中很多内容分享平台允许用户上传并观看视频,视频同时伴随文字标签和文字评论。相比于文本和图像,视频文件的编辑操作更为复杂,因此视频发布的门槛较高,视频信息的数量较少。由于其格式复杂,表达内容更丰富,使得分析起来也有一定的困难。

数值:数值是社会媒体最直观的表达形式。互联网中到处分布着海量数值信息,其中包括数值评分、浏览统计数量和地理信息定位等,数值信息用来最直观地表达信息,往往因格式简单被忽略。

由于基于社会媒体的旅游数据挖掘与分析主要是利用图像、文本和数值信息,对于实现不同的应用需要的数据类型不同,研究者根据自己的应用需求自行建立数据库,进而研究如何在互联网海量的社会媒体数据中挖掘出有价值的旅游信息,并且深入研究基于"互联网+旅游信息化"的旅游应用,将是一个很有意义及前景的研究课题。根据社会媒体信息的特点和性质,在社会媒体平台中进行数据挖掘与旅游推荐面临以下难题:

(1)多样性:相同的主题下存在着语义相同但表达方式多种多样的信息,这些信息造成了语义鸿沟,使得用户无法快速检索到所需要的内容。

(2)冗余性:用户随意对社交媒体中的数据进行编辑操作,造成海量的冗余信息,使得用户无法判断这些冗余信息的溯源关系,因此无法迅速检索出所需要的信息。

(3)异质性:用户可以上传各种数据类型不同的信息,其中包括数值信息、图像以及文本等,用户获取到单一模态的信息容易造成理解的局限性和片面性,融合异质信息从更全面的角度为用户筛选出所需信息也是一个突破点。

(4)智慧性:集体智慧可以从用户上传的海量信息中挖掘出来,这样就可以通过社交媒体的信息汇集成一个高级知识库,在高级知识层面下,社交媒体的精华信息让用户快速检索和获取也是一个新的挑战。

## 2 基于社会媒体旅游数据挖掘与分析

目前,国内外学者对基于社会媒体的旅游数据挖掘及应用开展了大量而广泛的研究,主要涉及旅游数据挖掘及检索与推荐模型等多个课题,并且提出了一系列的解决方法。

### 2.1 智能旅游系统回顾

随着"互联网+旅游信息化"的发展,智能旅游[6,7]越来越成为旅游发展的主要方向,通过结合社会媒体旅游数据实现旅游检索和推荐系统。随着社会媒体平台的飞速发展,用户在网站上分享文字、图片等信息,因此这些社会媒体平台上包含了大量的旅游相关数据。在构建智能旅游系统时,从所利用的信

息资源角度看,大多数系统利用了丰富的旅游多媒体信息,包括了文本、图像以及数值等信息。Wikitravel 是较早期的旅游信息网站,它能提供可靠的旅游信息方便用户查询。随着 Web 2.0 时代的来临,用户生成的内容可以呈现在互联网上,因此涌现出一批典型的智能旅游系统。VirtualTour[8] 提供在线旅游服务,即从用户生成的高质量图片中挖掘景点信息,并且在其设计的用户交互界面中,提供搜索信息服务,且可以为用户规划路线。DiverseSearch[9] 通过分析 Flickr 中含有地理标记的图片,进行景点基于图像的多视角展示,满足景点视觉表达的多样性。PersonalizedMM[10] 利用多模态的景点信息,包括文本、地理标记图片以及视频生成景点的信息摘要,然后根据用户的查询为用户个性化地推荐景点。Photo2Trip[11] 利用社会媒体网站中的数千万张图片以及数十万游记,以获取景点信息并挖掘景点之间的路径信息,为用户进行行程规划。gTravel[12] 是含有社交功能的智能旅游系统,允许用户分享旅游信息并能帮助用户规划行程。Ji 等[13] 在手机上实现旅游搜索功能,将图片通过手机传递给远程服务端,用户就可以获取该景点的多视角图片。

由典型的智能旅游系统可知目前智能旅游技术的发展现状,而在构建智能旅游系统时,从利用的信息资源角度看,部分系统利用了单一模态信息,如 Jing 等[8] 提出了在线旅游系统,其从社会媒体网站中的海量照片中挖掘出高质量的景点视觉信息。Wu 等[10] 利用从社会媒体中获取的旅游异质信息生成景点摘要,再根据用户信息进行个性化推荐。Photo2Trip[11] 通过图片和游记挖掘景点内部和景点之间的路线,然后为用户规划旅游路线。Hao 等[14] 利用 Flickr 中的地理标记照片来表示景点,使景点实现视觉多样化。Xiao 等[15] 提出利用有标记的景点图片构建一个 3D 模型,然后在基于图像分类的框架下将未标记的照片进行景点识别和分类。Min 等[16] 的工作通过远程服务器处理了低分辨的查询图片,然后利用低质量的照片可以识别和搜索到景点。最后被识别的景点可以在照片集的基础上进行三维视角展示。然而,旅游系统开始趋向于利用旅游异质信息,包括了文本、图像以及数值等信息。W2go[17] 通过分析 Flickr 中带有附加信息的照片以及雅虎旅游网站中的信息,为用户推荐所在城市的景点,并将景点的信息展示给用户。

综上所述,智能旅游系统主要是根据所获取的旅游信息进行挖掘分析,实现某些功能以帮助用户提升旅游体验。图 1 所示为智能旅游相关的基本研究框架。首先以机器学习、数据挖掘以及统计学习等学科为基础,需要解决的主要基本研究问题为:信息优选,即获取高质量的社会媒体旅游信息;信息融合,即需要将旅游异质信息进行融合;景点检索,即实现对景点多媒体实体的基于异质信息的检索;个性推荐,即根据用户情景上下文与旅游异质信息实现用户个性化旅游信息推荐。然后根据基本研究问题设计算法模型,实现社会媒体旅游信息获取、旅游异质信息融合以及旅游信息检索与推荐等算法,以完成基于社会媒体旅游数据挖掘与应用的研究,最终达到智慧旅游的研究目标。


图 1 智能旅游的基本研究框架

Fig. 1 Basic researching framework of smart tourism

## 2.2 旅游信息挖掘

旅游数据挖掘的目的是通过分析社会媒体中的旅游信息,得到知识以实现智能旅游应用。旅游信息挖掘的方法主要分为全局分析[18-20] 和局部分析[21-22]。全局分析就是把景点看作一个整体,然后综合分析景点的异质信息。Lu 等[11] 为用户提供景点照片,并通过排序准则为用户推荐合适的旅游路线。Ren 等[23] 从社交媒体的照片中获取知识,再通过重排序景点的图片搜索结果来表示景点的视觉表示。DLMSearch 系统[24] 利用图片检索出未标记的景点图片生成景点的视觉表示,以保证景点视觉表示的多

样性。Hays 等[25]利用基于数据的场景匹配法,生成图片的地理位置信息,该方法能够从 6 百万张数据库图像中,提取出 120 幅与输入图像视觉相似性最强的结果,并附带图像的地理信息,因此用户可以根据检索图像的位置分布,得到输入图像的位置信息。Kalogerakis 等[26]通过估计旅行路线的先验分布,获取一组时序图像的地理位置。Laere 在文献[27]中首先将不同地理位置进行聚类,之后将其对应的文本标注编译为字典,最后利用位置信息与字典训练朴素贝叶斯分类器,并将输入文本信息进行分类,确定其对应的位置信息。

局部分析则从景点内信息角度进行分析挖掘。部分工作利用了景点的单一模态信息,例如视觉信息或文本信息进行挖掘。Kennedy[9]等从视觉角度进行景点信息挖掘,然后利用多景点视角的图像来表示景点,即可快速生成景点的可视化表示。也有 Gao 等[17]利用了多模态的景点信息,即从 Yahoo 上的评论、Filckr 上的图像以及标签信息得到景点的摘要信息。Zitouni 等[18]结合用户评论、图像以及 Yahoo 中的数值评分来综合分析并推荐景点。Jaffe 等[28]利用分层聚类方法,针对特定地理区域选择有代表性以及最相关的图像,减少了冗余信息。Crandall[29]利用局部视觉特征,基于图模型进行景点的图像间的关系表示,选择特定地理区域中有代表性的图像。Moxley 等[30]使用地理信息、视觉信息挖掘技术,通过标注信息间的相关性,生成图像缺失的标注信息。

地标信息指特定地理位置的显著视觉特征以及文化、自然和历史特征等信息的结合,随着互联网上地标信息的大量出现,许多计算机视觉、计算机绘图学以及多媒体检索等领域的研究开始转向对地标信息挖掘技术的研究。Kennedy 等[9]通过图像匹配以及聚类的方法,对于不同地理位置提取出具有代表性的地标信息,从而精简信息量、降低检索负担。Gao 等[17]利用地理标记信息、图片信息以及多种旅游推荐网站提供的用户先验知识,按照不同用户的需求,对地标信息进行识别与排序。Zheng 等[31-32]在论文中构建了全球范围的标注信息数据库,其中涵盖了国家、城市、经度和纬度等多种信息。Kretzschmar 等[33]利用含地理信息的图像估计地标方位以及相机姿态。文献[34～36]通过匹配的方法来识别图像中缺失的地标信息。作者首先提取出图像中的局部特征,之后利用随机贝努利过程计算出量化的匹配值,根据匹配值的高低将图像归入相应的类别中。Abbasi 等[37]利用地标信息进行旅游信息推荐,首先根据用户输入的城市名称搜索到与之相关的所有图像,并利用支撑向量机(Support vector machine, SVM)分类器根据地标信息将图像分成地标图像与非地标图像两类,针对地标图像采用 TF-IDF 对标注信息进行排序,最终将排序后的检索结果返回给用户。

### 2.3 旅游信息重排序

通常旅游信息重排序问题关注于智能旅游系统返回结果的准确性和多样性,且用户对于返回的列表通常只关注排在前面的,因此对于检索问题的本质就是希望排序在最前面的结果最相关,需要使用户满意度最大化以及信息负载最小化。对于旅游系统而言,它可以基于查询词并结合旅游异质信息为用户提供相关的旅游信息。因此,在旅游系统中的排序和重排序问题就是返回与查询或者用户喜好最相关的结果。

为获取满意的搜索结果,在顾及返回结果的准确性的同时需要兼顾多样性。部分研究者为检索到多样的旅游信息,利用社会媒体中旅游相关的图片、文本和数值等信息进行分析挖掘。Kennedy 等[38]利用聚类以及 TF-IDF 理论估计出具有代表性的标注,从而由非结构化的标注信息样本获得对景点概念性的描述,以返回多样性的景点信息。Hao 等[39]利用概率隐语义分析(Probabilistic latent semantic analysis, PLSA)模型,从游记中发掘主题信息并用于表示旅游景点,以主题的方式表示结果的多样性。Hao 等[40]利用概率主题模型对 PLSA 模型进行改进,着眼于旅行游记信息的挖掘,以视觉图像以及文字描述的形式生成旅游景点概述。除了单一模态信息,多模态的信息如图片或地理标记信息等,旅客上传的游记、博客等文本信息也逐渐得到关注。Ji 等[41]通过融合博客中的情景上下文、社会信息,挖掘城市地标信息,再根据用户上传的游记,自动检索出旅行城市、地标以及有代表性的图片。

由于智能旅游检索系统中都收集大量的异质信息,即文本、图像和数值等,因此,如何利用这些异质

信息进行重排序也是智能旅游检索的研究重点。目前多数的研究工作关注于利用视觉信息进行重排序。Ren 等[23]将含有典型景点视角的照片进行重排序以生成景点摘要。DLMSearch 系统[24]支持用户查询,并且考虑景点位置信息和图像的质量好坏,将基于用户查询返回的搜索图片结果进行重排序来保证景点搜索的多样性。还有一些系统利用了旅游多媒体信息,Kennedy 等[9]对景点生成代表性的视觉展示,其先利用基于文本的搜索进行初始查找图片,然后再结合视觉特征进行重排序。Zitouni 等[18]由网站评论和用户上传的图像标来判断用户喜好以及景点的受欢迎度,再基于这两点将合适的景点推荐给用户。Collins[19]建立的个性化的旅游推荐系统利用定位信息和地理标记照片来重排序景点。

因此,智能旅游系统主要关注返回结果的多样性以及相关性,进而考虑其重排序策略。考虑到如何平衡相关性和多样性的问题,很多排序模型包括基于分类的、基于聚类的以及基于图的方法等,都可以应用在智能旅游系统排序上。当前,研究者们开始关注多模态旅游信息,即考虑多样性、相关性等的时候,从异质信息融合的角度进行思考。对于异质信息融合主要分为两类:前融合和后融合。前融合则看作特征的融合[42],即对各模态信息进行特征提取进行特征级别的融合;后融合则是将单一模态信息的结果进行融合。

### 2.4 旅游信息推荐

旅游信息推荐可根据用户喜好及情景上下文信息为用户推荐其喜好的景点,从而帮助用户规划行程。随着旅游的智能化发展,旅游信息推荐成为了热门研究课题,它主要是对旅游信息和用户信息进行分析,所以主导的方法主要分为两类:(1)基于协同过滤的方法;(2)基于内容的方法。一方面,用户的旅游历史,用户的群组以及用户的定位关系等这些信息都可以被协同过滤方法采纳。Koutrika 等[43]利用基于开销的协同过滤方法,从旅游公司挖掘相应信息并将不同的旅游套餐结合开销因素建立隐因素模型进行推荐。Zheng 等[44-46]利用定位数据(Global position system, GPS)采用基于轨迹的方法,提取出大众旅行路线以及景点,由用户之间位置以及旅行历史的相似性,为用户进行个性化推荐。Yoon 等[47]提出了一种基于多用户 GPS 轨迹的方法,根据用户提供的起点、终点以及旅行时间,为用户提供有效的路径信息。Clements 等[48]根据用户输入的一张附带地理标注的图片,利用用户之间兴趣相似性以及其他用户的历史旅行信息,为用户智能化推荐旅行景点。Chen 等[49]根据用户特征(性别、年龄和种族)以及旅行记录,构建概率推荐模型,为用户个性化提供旅行景点信息。Kurashima 等[50]提出了一种摄像师行为模型,利用用户偏好数据以及位置信息,通过主题模型与马尔可夫模型,估计出摄像师选择旅行景点的概率,从而指示了不同景点的受欢迎程度。

另一方面,基于内容的推荐是通过挖掘旅游信息。对于基于内容的方法,一些工作利用了在社交网站中的地理信息标记的照片。Cao 等[51]根据地理定位信息将大量的地理标记的照片进行聚类,然后当用户用文本或者图片去查询景点,就推荐给用户和这些景点相似的一些景点。Jiang 等[52]挖掘照片的情景上下文信息,包括文本标记、地理位置和图片来寻找相似的用户,进而通过计算相似的用户来预测用户喜欢的景点,最后利用排序算法给出用户一个景点的排序列表。Memon 等[53]提出构建一个专业的旅游向导系统,不仅利用用户的旅游历史还考虑了结合旅游的时间信息和用户喜好来制定旅游计划。Angel 等[54]利用文本检索方法,通过关键词匹配获得互联网中关于城市、旅馆和航线等旅行相关信息的评分,为不同需求的用户提供多种出行信息。很明显,两种不同的分类在推荐中都有优缺点。一些方法结合协同过滤和基于内容的推荐以克服单一方法的不足。Cheng 等[55]提出了一种概率个性化旅游推荐模型,该模型综合考虑带有地理标记的照片信息以及用户喜好等,在贝叶斯学习框架中完成个性化的景点和路线推荐。Xie 等[56]利用多种推荐系统的评分信息,根据用户的时间、开销预算,为用户提供多种可供选择的旅行路线。Lu 等[57]提出了一种旅行数据挖掘模型,通过景点选择、时限评估和分值生成 3 个部分,根据旅行时间限制,为用户推荐最优路线。

由旅行数据的特点可知,旅行数据具有稀疏性和多样性。因为大多数人并不是经常出游,所以用户旅行历史不能大量获取,因此不能很好地执行协同过滤方法。为避免冷启动问题,采用基于内容的推荐

方法更加可行。与此同时,不同的城市有不同的风格,例如不同城市的两个公园从视觉上看是非常不同的,所以需要采用尽可能多的模态信息来表示景点信息,但是大多数的推荐方法只考虑了单一模态或者少数模态的旅游信息,而将其他模态信息作为情景上下文信息辅助推荐。尽管有极少数的工作考虑了多模态的融合,但是它们往往忽略了各模态之间的相关信息。因此,旅游信息推荐需要多关注多模态融合并且考虑模态间的相互关系来进行推荐。

## 3 研究的难点和未来展望

基于社会媒体的旅游信息挖掘的研究应结合传统的旅游信息服务功能,重点突出智能化检索和个性化推荐的应用建设,针对用户的需求给出个性化解决方案。然而,目前的智能旅游系统存在诸多问题,如用户定位不明确、忽略用户需求等,使得基于社会媒体的旅游数据挖掘的应用研究遇到极大的挑战。其中的难点体现在:(1) 网络信息难以有序地提取和处理。由于旅游数据种类繁多,具有图像、音频、文本和标记等多种形式,导致了同一信息会存在不同结构的表达方式,要对多媒体信息进行分析挖掘,首先需要对旅游数据进行去冗余处理,再进行结构化。因此旅游多模态异质信息的特征提取以及结构化处理是信息智能化分析的难点。(2) 数据受多种规则影响难以涵盖全面的信息。由于网络中的旅游信息涵盖面广、结构多样和内容相关性错综复杂,难以利用统一合理的模型来综合构建多种关系规则,从而制约了旅游信息分析推荐方法的研究。(3) 用户-景点关系复杂难以构建个性化模型。由于用户之间个人特性以及对不同景点偏好程度的差别,导致用户-景点关系网络非常繁琐,使得旅游信息智能推荐问题变得更加复杂。因此,为了研究基于社会媒体的旅游信息挖掘方法,首先需要将旅游异质信息进行结构化,再结合相关领域算法,抽象出新的数学模型,实现检索与推荐等智能旅游应用。上述问题涉及到机器学习、数据挖掘和信息检索推荐等一些基础问题,需要从景点表述的新方法、信息检索表示的新手段和个性化推荐的新思路出发,基于基础问题来研究新的模型和新的应用。新的研究成果不仅能够丰富和发展基于社会媒体的旅游数据挖掘的研究应用,还需要能指导数据挖掘、多媒体分析和机器学习学科的研究和发展。

随着互联网日益广泛蓬勃的发展,旅游信息经过爆炸式增长已跨入了大数据时代。此外,用户对网络信息的依赖,使该领域的应用研究向高效化、智能化发展。因此,鉴于实际应用的要求,智能旅游应用已经成为“互联网+旅游信息化”的一个必然方向。基于社会媒体的旅游数据挖掘的研究是建立在信息融合、联合优化和智能推荐等多种交叉学科的基础上,是为了解决网络大数据背景下旅游异质信息的分布散乱、难以合理表述和忽视用户需求等问题。此外,智能化旅游信息服务在真正意义上将用户作为重点关注对象,充分考虑个体的个性化需求,方便用户进行旅游相关决策,具有长远的应用价值。未来相关研究方向可概括如下:

(1)大数据旅游信息表述。大数据下的旅游信息主要来源于各类社会媒体平台,景点作为多媒体实体主要存在描述文本、评价文本、静态图片以及动态图像等大量异质的分散的非结构化数据,形成了基于大数据的多模态、多维度信息。因此,如何对此类多媒体实体进行特征提取,并且进行多模态融合以便于后续利用是研究的重点。而随着深度学习在机器学习领域的突破性进展,利用深度学习可避免特征的选择,进而利用其强大的信息表示能力进行建模,实现景点多模态异质信息描述,完成基于深度学习的智能旅游应用是研究的重点。

(2) 大数据旅游信息智能化检索。大数据时代网络中的旅游信息种类繁多、数量巨大,其中包含文本、图片、图形和视频等多种信息。对大规模的数据直接进行处理必将导致检索速率低下,难以满足用户实时性需求,此外由于用户知识及经验有限,往往难以准确表述出检索需求,因此如何兼顾信息间的多样性及相关性,并结合用户需求,从网络数据中提取出典型内容作为检索输出以提升检索效率、满足用户需求是研究的关键。

(3) 大数据旅游信息个性化推荐。随着互联网的飞速发展,网络用户日益增多,用户个性化需求日

益增加,为多媒体信息推荐技术提供了机遇与挑战。一方面,景点信息具有独特性,即不存在完全相同的景点信息,传统的机器学习方法需要大量训练数据,所以需要利用较少的标记数据训练模型使其具有迁移性,能服从不同的异质数据分布;另一方面,丰富的用户情景上下文的有效利用有利于满足个体需求,所以需要获取更多的情景上下文信息有助于构建更加合理的推荐模型。因此,如何利用较少训练数据并结合用户的个人偏好及情景上下文信息是设计个性化的推荐模型研究的核心。

## 4 结束语

在"互联网+旅游信息化"时代,基于社会媒体的旅游数据挖掘的研究是一个新兴的方向,这个方向在几个主要的研究领域被广泛关注,如计算机视觉、机器学习和多媒体分析等。本文综述了近期基于社会媒体的旅游数据挖掘的研究和应用,从已经取得的成果及目前面临的挑战来预测未来可能的研究方向。具体而言,综述了典型的智能旅游系统,并深入分析了旅游数据挖掘、旅游信息重排序和旅游信息推荐 3 个方面。基于目前的研究成果,总结了研究难点且提出研究挑战,最后展望了未来的发展方向,以期待出现更有吸引力的基于社会媒体的旅游数据挖掘应用研究的新成果。

**参考文献:**


[1] 国家发展和改革委员会. 国家及各地区国民经济和社会发展"十二五"规划纲要[M]. 北京:人民出版社,2011.
National Development and Reform Commission. The 12th five-year plan for countries, regions and national economic and social development program[M]. Beijing: People's Publishing House, 2011.

[2] 马化腾. 互联网+国家战略行动路线图[M]. 北京: 中信出版社集团,2015.
Ma Huateng. Internet+roudmap of national public strategy and plan[M]. Beijing: Citic Press Group, 2015.

[3] 王吉斌,彭盾. 互联网+:传统企业的自我颠覆、组织重构、管理进化与互联网转型[M]. 北京:机械工业出版社,2015.
Wang J, Peng D. Internet+: Traditional enterprise of cannibalism, refactoring, management evolution and the Internet transformation[M]. Beijing: Mechanical Industry Press, 2015.

[4] Correa T, Hinsley A W, De Zuniga H G. Who interacts on the Web? : The intersection of users' personality and social media use[J]. Computers in Human Behavior, 2010, 26(2): 247-253.

[5] Asur S, Huberman B. Predicting the future with social media[C]// IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Toronto, Canada : IEEE, 2010: 492-499.

[6] Werthner H. Intelligent systems in travel and tourism[C]//The International Joint Conference on Artificial Intelligence. Acapulco, Mexico: AAAI, 2003: 1620-1628.

[7] Majid A. 基于地理标签的社会媒体数据挖掘的智能旅游推荐研究[D]. 杭州: 浙江大学, 2012.
Majid A. Research on intelligent tour recommendation by mining geo-tagged social media data[D]. Hangzhou: Zhejiang University, 2012.

[8] Jing F, Zhang L, Maw Y. Virtual tour: An online travel assistant based on high quality images[C]//The ACM International Conference on Multimedia. Santa Barbara, CA: ACM, 22-28, 2006: 599-602.

[9] Kennedy L S, Naaman M. Generating diverse and representative image search results for landmarks[C]//The International Conference on World Wide Web. Beijing, China: ACM, 2008: 297-306.

[10] Wu X, Li J, Zhang Y, Tang S, et al. Personalized multimedia web summarizer for tourist[C]//The 17th International Conference on World Wide Web. Beijing, China: ACM, 2008: 1025-1026.

[11] Lu X, Wang C, Yang J-M, et al. Photo2trip: Generating travel routes from geo-tagged photos for trip planning[C]//The ACM International Conference on Multimedia. Firenze, Italy: ACM, 2010:143-152.

[12] Zhang R, Guo X, Sun H, et al. Travel: A global social travel system[C]//The ACM International Conference on Multimedia. Nara, Japan: ACM, 2012: 1291-1292.

[13] Ji R, Duan L-Y, Chen J, et al. Location discriminative vocabulary coding for mobile landmark search[J]. International Journal of Computer Vision, 2012, 96(3): 290-314.

[14] Hao Q, Cai R, Yang J-M, et al. Travelscope: Standing on the shoulders of dedicated travelers[C]//The ACM International Conference on Multimedia. Beijing, China: ACM, 2009: 1021-1022.

[15] Xiao X, Xu C, Wang J, et al. Enhanced 3-D modeling for landmark image classification[J]. IEEE Transactions on Multime-

dia, 2012, 14(4): 1246-1258.

[16] Min W, Xu C, Xu M, et al. Mobile landmark search with 3D models[J]. International Journal of Computer Vision, 2014, 16(3): 623-636.

[17] Gao Y, Tang J, Hong R, et al. W2Go: A travel guidance system by automatic landmark ranking[C]//The ACM International Conference on Multimedia. Firenze, Italy: ACM, 2010: 123-132.

[18] Zitouni H, Sevil S, Ozkan D, et al. Re-ranking of web image search results using a graph algorithm[C]//The International Conference on Pattern Recognition. Florida, USA: IEEE, 2008: 1-4.

[19] Collins M. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron[C]//The 40th Annual Meeting on Association for Computational Linguistics. USA: Association for Computational Linguistics, 2002 : 489-496.

[20] Tian X, Yang L, Wang J, et al. Bayesian video search reranking[C]//The ACM International Conference on Multimedia. Beijing, China: ACM, 2008: 131-140.

[21] Kennedy L S, Chang S F. A reranking approach for context-based concept fusion in video indexing and retrieval[C]//The ACM International Conference on Image and Video Retrieval. Amsterdam, the Netherlands: ACM, 2007:333-340.

[22] Hsu W H, Kennedy L S, Chang S-F. Video search reranking via information bottleneck principle[C]//The International Conference on Multimedia. Santa Barbara, CA: ACM, 2006: 35-44.

[23] Ren Y, Yu M, Wang X J, et al. Diversifying landmark image search results by learning interested views from community photos[C]//The International Conference on World Wide Web. NC, USA: ACM, 2010: 1289-1292.

[24] Ye J, Chen J, Chen Z, et al. DLMSearch: Diversified landmark search by photo[C]//The International Conference on Multimedia. Nara, Japan: ACM, 2012: 905-908.

[25] Hays J, Efros A. IM2GPS: Estimating geographic information from a single image[C]//The Conference on Computer Vision and Pattern Recognition. Anchorage, Alaska, USA:IEEE, 2008: 1-8.

[26] Kalogerakis E, Vesselova O, Hays J, et al. Image sequence geolocation with human travel priors[C]//The International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009: 253-260.

[27] Van Laere O, Schockaert S, Dhoedt B. Towards automated georeferencing of Flickr photos[C]//The 6th Workshop on Geographic Information Retrieval. Zurich, Switzerland: ACM, 2010: 5.

[28] Jaffe A, Naaman M, Tassa T, et al. Generating summaries and visualization for large collections of geo-referenced photographs[C]//The International Workshop on Multimedia Information Retrieval, Santa Barbara, California, USA: ACM, 2006: 89-98.

[29] Crandall D J, BackstromL, Huttenlocher D, et al. Mapping the world's photos[C]//The 18th international conference on World Wide Web. Madrid, Spain: ACM, 2009: 761-770.

[30] Moxley E, Kleban J, Manjunath B. Spirittagger: A geo-aware tag suggestion tool mined from Flickr[C]//The International Conference on Multimedia Information Retrieval. Vancouver, British Columbia, Canada: ACM, 2008: 24-30.

[31] Zheng Y T, Zhao M, Song Y, et al. Tour the world: Building a web-scale landmark recognition engine[C]//The International Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 1085-1092.

[32] Zheng Y T, Zhao M, Song Y, et al. Tour the world: A technical demonstration of a web-scale landmark recognition engine [C]//The ACM International Conference on Multimedia. Beijing, China: ACM, 2009: 961-962.

[33] Kretzschmar H, Stachniss C, Plagemann C, et. al. Estimating landmark locations from geo-referenced photographs[C]// The IEEE/RSJ International Conference on Intelligent Robots and Systems. Nice, France: IEEE, 2008: 2902-2907.

[34] Zheng Y T, Zhao M, Song Y, et al. Tour the world: Building a web-scale landmark recognition engine[C]// The International Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 1085-1092.

[35] Li X, Wu C, Zach C, Lazebnik S, et al. Modeling and recognition of landmark image collections using iconic scene graphs [C]//The European Conference on Computer Vision. Marseille, France: Springer, 2008: 427-440.

[36] Li Y, Crandall D J, Huttenlocher D P. Landmark classification in large-scale image collections[C]//The International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009:1957-1964.

[37] Abbasi R, Chernov S, Nejd W, et al. Exploiting Flickr tags and groups for finding landmark photos[C]//The 31th European Conference on IR Research on Advances in Information Retrieva. Toulouse, France: Springer, 2009:654-661.

[38] Kennedy L S, Naaman M, Ahern S, et al. How Flickr helps us make sense of the world: Context and content in community-contributed media collections[C]//The ACM International Conference on Multimedia. Augsburg, Germany: ACM, 2007: 631-640.

[39] Hao Q, Cai R, Wang C, et al. Equip tourists with knowledge mined from travelogues[C]//The International Conference on World Wide Web. NC, USA: ACM, 2010: 401-410.

[40] Hao Q, Cai R, Wang X J, et al. Generating location overviews with images and tags by mining user-generated travelogues [C]//The ACM International Conference on Multimedia. Beijing, China: ACM, 2009:801-804.
[41] Ji R, Xie X, Ma W Y. Mining city landmarks from blogs by graph modeling[C]//The ACM International Conference on Multimedia. Beijing, China: ACM, 2009: 105-114.
[42] 刘帅,李士进,冯钧. 多特征融合的遥感图像分类[J]. 数据采集与处理,2014,29(1):108-115.
Liu Shuai, Li Shijin, Feng Jun. Remote sensing image classification based on adaptive fusion of multiple features[J]. Journal of Data Acquisition and Processing, 2014, 29(1):108-115.
[43] Koutrika G, Bercovitz B, Garcia-Molina H. FlexRecs: Expressing and combining flexible recommendations[C]//The ACM SIGMOD International Conference on Management of Data. Providence, USA: ACM, 2009:745-758.
[44] Zheng Y, Xie X. Learning travel recommendations from user generated GPS traces[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(1):1-29.
[45] Zheng Y, Xie X, Ma WY. Mining interesting locations and travel sequences from GPS trajectories[C]//The 18th International Conference on World Wide Web. Madrid, Spain:ACM, 2009:791-800.
[46] Zheng Y, Zhang L, Ma Z, et al. Recommending friends and locations based on individual location history[J]. ACM Transactions on the Web, 2011, 5(1): 5.
[47] Yoon H, Zheng Y, Xie X, et al. Smart itinerary recommendation based on user-generated GPS trajectories[C]//The International Conference on Ubiquitous Intelligence and Computing. Brisbane, Australia: IEEE, 2009:19-34.
[48] Clements M, Serdyukov P, Vries A, et al. Using Flickr geotags to predict user travel behavior[C]//The 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval. Geneva, Switzerland: ACM, 2010:851-852.
[49] Chen Y Y, Cheng A J, Hsu W H. Travel recommendation by mining people attributes and travel group types from community-contributed photos[J]. IEEE Transactions on Multimedia, 2013, 15(6): 1283-1295.
[50] Kurashima T, Iwata T,Irie G, et al. Travel route recommendation using geotags in photo sharing sites[C]//The ACM International Conference on Information and Knowledge Management. Toronto, Canada: ACM, 2010:579-588.
[51] Cao L, Luo J, Gallagher A, et al. Aworldwide tourism recommendation system based on geotagged web photos[C]//The International Conference on Acoustics Speech and Signal Processing. Dallas, USA: IEEE, 2010:2274-2277.
[52] Jiang K, Wang P, Yu N. ContextRank: Personalized tourism recommendation by exploiting context information of geotagged web photos[C]//The International Conference on Image and Graphics (ICIG). Heifei, China: IEEE, 2011: 931-937.
[53] Memon I, Chen L, Majid A, et al. Travel recommendation using geo-tagged photos in social media for tourist[J]. Wireless Personal Communications, 2015, 80(4):1347-1362.
[54] Angel A, Chaudhuri S, Das G, et al. Ranking objects based on relationships and fixed associations[C]//The International Conference on Extending Database Technology: Advances in Database Technology. Saint Petersburg, Russia: ACM, 2009: 910-921.
[55] Cheng A J, Chen Y Y, Y. Huang T, et al. Personalized travel recommendation by mining people attributes from community-contributed photos[C]//The ACM International Conference on Multimedia. Scottsdale, AZ, USA: ACM, 2011:83-92.
[56] Xie M, Lakshmanan L, WoodP T. CompRec-trip: A composite recommendation system for travel planning[C]//The IEEE International Conference on Data Engineering. Hannover, Germany: IEEE, 2011:1352-1355.
[57] Lu E H, Lin C, Tseng V S. Trip-mine: An efficient trip planning approach with travel time constraints[C]//The IEEE International Conference on Mobile Data Management. Sweden: IEEE, 2011:152-161.

**作者简介:**

**高新波**(1972-),男,教授,博士生导师,研究方向:机器学习、模式识别和多媒体内容分析,E-mail:xbgao@mail.xidian.edu.cn。

**沈钧戈**(1987-),女,博士研究生,研究方向:机器学习、数据挖掘和多媒体内容分析。