# 基于微博数据的北京市热点区域意象感知

谢永俊[1],彭 霞[2*],黄 舟[1],刘 瑜[1]

(1. 北京大学遥感与地理信息系统研究所,北京 100871; 2. 北京联合大学旅游信息化协同创新中心,北京 100101)

**摘 要:**"城市意象"研究对城市文化感知、城市管理与规划、旅游资源开发等具有重要意义。近年来,随着智能移动终端和社交媒体的普及,产生了大量城市内包含有文本和地理位置等信息的社交媒体数据,涉及城市的各个区域,为开展城市意象的综合感知研究提供了新的途径。本文以2016年北京市带位置签到的新浪微博数据为例,在空间聚类发现热点区域的基础上,采用词频—逆文件频率(TF-IDF)与文档主题生成模型LDA两类典型的文本分析的方法,挖掘城市不同热点区域的主题,以感知北京市不同热点区域的社会文化功能和人群行为,并在此基础上通过对热点区域高频主题词进行共词聚类分析,深度挖掘北京市的总体意象。研究表明,运用文本挖掘及地理大数据分析的城市意象研究方法,能及时感知人群在城市不同场所的活动、态度、偏好,从而揭示城市的社会文化及功能特征,是对刻画城市物质形态的城市意象五要素模型的重要补充。此外,以北京市热点区域为例的实证研究结果对现实中的城市特色传承与空间品质塑造等有一定的启发意义。

**关 键 词:**地理空间数据;社交媒体;微博数据;文本分析;热点区域;城市意象

## 1 引言

"意象"(Image)是"抽象"的感觉和感知,是人们心中对某一区域或地方的心理图像,是人们的印象、看法、认知、评价、感情的综合体现(白凯, 赵安周, 2011)。对城市意象的研究始于20世纪60年代,其后一直是学者们研究的热点问题,对城市特色与空间品质塑造,城市规划与管理,以及城市旅游开发等具有重要意义(Tuan, 1997; 白凯等, 2008; 周尚意等, 2011)。城市空间是城市文化的主要载体,是城市文化复兴与创新的手段之一,对于城市的文化传承和特色展示具有极为重要的作用。如何高效、正确地感知城市意象,获得人们在不同城市空间中普遍的感受,并深入解析不同城市空间所承载的城市文化,是该项研究和应用的关键。

20世纪60年代,凯文·林奇(2001)最早提出了城市意象的研究方法,即通过视觉感知城市物质形态的理论,采用大量调查和绘制认知地图方法研究市民心目中的城市形象,并将城市意象概括为五要素:标志、节点、路径、边界和区域。凯文·林奇之后的学者普遍采用了与之相似的方法,即运用问卷调查、访谈和意向草图的社会学调查方法对城市意象的要素识别、空间分布与品质特征等进行研究。从凯文·林奇的认知地图五大要素出发,李郇等(1993)通过问卷调查研究了广州市民城市意象;顾朝林等(2001)通过照片辨认和认知地图调查对北京城市意象进行了研究,发现构成北京市意象的要素主要是道路、标志和节点。由于城市意象是人们对城市的

**收稿日期:**2017-06;**修订日期:**2017-08。

**基金项目:**国家自然科学基金项目(41501162);北京市社会科学基金项目(17JDGLB002);北京联合大学人才强校优选计划(BPHR2017DS08)[**Foundation:** National Natural Science Foundation of China, No.41501162; Beijing Philosophy and Social Science Foundation, No.17JDGLB002; Premium Funding Project for Academic Human Resources Development in Beijing Union University, No.BPHR2017DS08]。

**作者简介:**谢永俊(1993-),男,福建长汀人,本科生,主要研究方向为社交媒体地理数据挖掘、高性能地学计算,E-mail: afatpig@pku.edu.cn。

**通讯作者:**彭霞(1983-),女,江西樟树人,讲师,主要从事智慧旅游、地理大数据等方面研究,E-mail: ivy_px@163.com。

**引用格式:**谢永俊, 彭霞, 黄舟, 等. 2017. 基于微博数据的北京市热点区域意象感知[J]. 地理科学进展, 36(9): 1099-1110. [Xie Y J, Peng X, Huang Z, et al. 2017. Image perception of Beijing's regional hotspots based on microblog data[J]. Progress in Geography, 36(9): 1099-1110.]. DOI: 10.18306/dlkxjz.2017.09.006

主观感知，而城市的复杂性与动态性，以及个体的社会阶层、文化背景与生活经历等方面的差异导致了不同群体之间对城市意象的感知差异。之后的学者逐步在研究过程中引入观察者的社会文化背景因素，如Lee等(2010)对比了本地居民和外地游客对旅游地意象感知的差异；白凯, 张春晖等(2001)研究了不同文化群体(如不同国籍和不同语系群体)对中国旅游目的地意象感知的差异。还有一些学者专门开展了针对外来游客的城市意象研究，如田逢军等(2008)通过照片与意象草图调查相结合方法，对南昌市游客的旅游地意象空间进行了研究。

徐磊青(2012)通过系统地梳理和分析中国城市意象的研究文献，总结出中国城市意象研究的三大主题：基于凯文·林奇五要素模型的城市意象要素识别和城市空间结构研究，以揭示城市特色为目的的独特性城市意象要素研究，以提供城市空间品质为目的的意象元素评估性研究。然而，上述城市意象研究仍主要采用的是以问卷为主的社会学调查方法，在样本数量、数据代表性等方面具有局限性。近年来，随着智能移动终端和社交媒体的普及，人们可以随时随地在社交网络中表达自身的真实感受；并且，在社交媒体应用中，位置服务已成为主流，社交媒体数据既包含地理位置信息，也包含时间、文本、图片和多媒体信息等，提供了一个丰富、广阔、可探索的信息空间；此外，社交媒体数据还具有实时性强、数据量大等特点，从中可以感知信息发布者在发布信息时刻的真实想法、情绪及行为偏好。在城市中，每一个人都可视为一个能感受城市政治、经济、文化、历史、环境等等各种因素的“传感器”，他们将自己感受的内容以文字、图片、表情、标签等等形式通过社交媒体发布，而通过对这些数据的分析，可以提取归纳出对城市意象的普遍感知。因此，社交媒体数据为过去以调查为主要研究手段的城市意象研究提供了一个前所未有的“社会感知”新途径(刘瑜, 2016)。

此外，城市意象自提出至今，其研究内容与研究方法亦在不断发展。凯文·林奇的五要素理论构成了对城市意象物质空间结构研究的基础。随着后现代主义、女性主义等思潮的兴起，许多学者对凯文·林奇的结构型研究方法提出了质疑，批判其只关注构成城市的实体环境，而忽略了城市中影响意象形成的“社会意识、风土人情、历史变迁、城市功能”等因素(徐磊青, 2012)，并认为城市意象不仅包含看得见、摸得着的城市视觉形态感知，还应包含更为复杂的由市民活动所赋予的社会与文化意义(Gulick, 1963; Klein, 1967; Gordon, 1978; 沈益人, 2004; 徐磊青, 2012)。城市意象的研究已经从单纯的城市物质空间结构发展到以城市物质空间为主，并加入城市文化、社会行为等非物质要素。城市意象的研究方法，也从认知地图、调查问卷等传统社会科学研究方法发展到来自社交媒体的地理照片、网络开放数据等新数据方法；研究模式从单个城市的城市意象的描述性研究发展到多个城市特征识别的类型学研究和比较研究(曹越皓等, 2017)。其中，基于地理照片数据的城市意象感知研究正日益兴起，其主要包括两个方面；一是基于图片元数据(如拍照地点、时间)研究城市意象的要素构成与空间分布特征，从而对凯文·林奇的城市意象五要素进行实证、扩展和补充；二是通过对图片内容本身进行识别和分类，对不同城市的城市意象特色度、相似度及差异性进行研究。例如，Salesses等(2013)使用数千张地理照片，比较了美国的纽约、波士顿和奥地利的林茨、萨尔茨堡4个城市的安全性、社会阶层和独特性感知的差异，并发现纽约某街区安全性与社会阶层感知与自杀者数量之间的显著相关性。Liu等(2016)使用Panoramio和Flickr照片数据集，利用深度学习技术对照片进行意象分类，统计分析了全球7个典型城市的意象要素类型特征与空间分布特征，并探讨了各个城市在意象类型上的相关性和差异性；北京城市实验室(BCL)也就此开展了系列研究，龙瀛提出“图片城市主义”理论，综合分析Flicker照片的拍照点、标签信息和图片内容，开展中国24个城市的主导意象、城市意象特色度和相似度研究(曹越皓等, 2017; 龙瀛等, 2017)。这些研究反映了新数据环境下城市意象感知的新趋势。而微博数据作为社交媒体数据中的另一种重要数据源，直接反映了用户个体在不同地点、区域的活动、想法及行为。目前，亦出现了利用微博数据开展的城市意象研究。例如邓力凡等(2017)基于支持向量机(SVM)将微博用户划分为居民和游客两类，再通过可视化呈现，对比这两类人群对城市的感知区域与强度差异。而以上基于微博数据的城市意象研究，主要是利用微博数据中的签到地点、时间等元数据，而信息极为丰富的文本数据在当前城市意象研究中尚未得到足够重视。目前，微博文本数据主要用于研究舆论生成演

变机制、舆情监控与预测、情绪感知等领域，其中亦有一些与城市问题有关的典型研究。例如，Luo等(2017)以“占中”为例，研究了微博中次级危机传播机制，发现微博主题由“占中”这一政治事件转变为情绪更为负面的旅游抵制运动，并引发了中国香港与大陆人群的冲突对抗，从而对城市旅游造成重要影响。然而，将微博文本数据应用于城市意象的研究目前还比较鲜见。

事实上，通过对微博中的文本数据进行挖掘，能充分揭示不同区域、不同群体视角下社会文化多样性和功能特征，获得城市意象所包含个人情感、环境的社会文化意义及其被使用的频率等独特的非物质性要素，有效地扩展凯文·林奇的结构型意象研究内涵。因此，本文拟面向社交媒体大数据，在感知城市意象物质空间要素的基础上，研究和探索通过文本数据的主题分析提取城市意象非物质要素的方法，实现城市意象的一体化综合感知。

此外，文本主题挖掘方法也在不断发展，最典型的有用于关键词提取的词频—逆文件频率(Term Frequency- Inverse Document Frequency, TF-IDF)方法，用于文本主题分析的潜在狄利克雷分布(Latent Dirichlet allocation, LDA)模型(Blei et al, 2003)等。在文本分析的应用方面也有一定进展。比如在舆情分析领域，樊蕾(2013)开展了新浪微博“微话题”研究；宋蕾等(2014)基于微博数据，研究了基于LDA主题建模的舆情分析系统。同时，在心理学领域，一些学者基于微博数据，侧重于对个体或群体情绪的实时观察与分析，研究不同人格特征、社会文化背景人群在微博语言和行为上的差异，并在此基础上归纳出区分特定用户群体的特征(如人格个性、心理健康水平、生活满意度、主观幸福感等)。美国学者最早利用Twitter对人们的情绪变化进行实时分析，如美国东北大学与哈佛大学的学者开展了“国家的脉搏”(Paulse of the Nation)项目，实时分析了美国各州人们情绪的变化情况，之后基于Twitter数据分析了包括中国在内的84个国家不同文化背景用户的情绪变化规律(Golder et al, 2011)；汪静莹等(2016)分析了微博活跃用户在不同季节和时间下情绪的生物节律变化，还研究了不同生活满意度水平的微博用户在微博语言和行为上的差异，归纳出区分用户生活满意度高低的微博可识别特征。而在旅游领域，也有一些学者基于在线旅游社(Online Travel Agent, OTA)或旅游社交网站的游客评论数据，进行旅游目的地形象及品牌个性的感知研究。比如，钟栎娜(2015)利用携程网的用户评论文本，分析了感知旅游地的要素与结构；Wong等(2017)利用TripAdvior上2005年至2013年之间的游客评论数据，揭示了这一时间范围内澳门目的地形象的演进过程。由于旅游目的地形象与城市意象内涵上的相似性，因此，旅游领域的这些研究为从文本主题分析视角开展城市意象感知研究提供了新参考。

本文基于社交媒体大数据——带位置签到的北京市微博数据，尝试在城市意象物质空间结构提取的基础上，应用主题分析的手段提取附着于空间结构的非物质要素，力图以全新的视角、全新的方法综合感知当代北京的城市意象。重点研究了北京市热点公共空间的城市意象感知：通过空间聚类方法获得城市热点区域，再采用TF-IDF提取关键词、LDA建立主题模型等方法，提取北京市各热点区域内用户微博的关注主题，分析北京市各热点区域内的文化、功能和特性，深入挖掘人们对城市热点区域的普遍印象和认知。在此基础上进一步对比了各热点区域之间的差异，以洞察当代北京的不同空间所承载的城市文化元素的差异，力图通过大数据分析方法解析北京市热点区域之“城市意象”。

## 2 研究区域、数据与方法

### 2.1 研究区域

本文以北京市市区为研究区域，使用空间聚类方法预先将其划分为若干地区，如天安门、故宫、景山、天坛等等，每一个区域可以视为一个多边形，这些热点区域(人类活动密集区)相互不重叠，也不一定相互接触，即这些热点区域总体并不覆盖整个北京市市区的所有地块。北京市市区范围内人员密集，社交媒体普及率高，聚集了大量微博用户，可获得的数据密度高、数量大，地区特色鲜明、典型，因此在本次城市意向感知研究中具有足够的样本数据。

### 2.2 数据说明

相比手机移动数据等其他大数据来源，社交媒体签到数据被认为更适合于城市结构识别及人群行为研究，由于签到事件是人们有意识的行为，只有当人们在某个特定地点停留相对较长时间，并且认为有值得记录的事情时，才会在某地签到(Ka-

plan et al, 2010; Cai et al, 2017)。新浪微博是中国最流行的社交媒体平台之一,据统计截至2016年底,微博月活跃人数已突破3亿,其中移动端用户占比达90%。在本项研究中,使用微博签到数据用于代表北京市的人群活动。利用新浪微博开放API,编写程序抓取2016年北京市六环市区内的带位置签到的微博数据。记录总量约500万条,其中每条记录均包含用户信息、数据发布时的位置(经纬度坐标)、文本内容(微博文本)、发布时间等等。经剔除分类区外数据和不合格数据后剩余约70万条。

### 2.3 研究方法

本文使用北京市微博数据,以主题分析方法为核心技术手段,在数据预处理的基础上,形成以北京市热点区域为空间研究单元的文档集合,分别应用TF-IDF与LDA两类主题分析方法,对热点区域的关注点、主题分类等进行提取,在此基础上开展分析,形成对热点区域意象、以及区域之间差异的探讨。图1为本文研究的技术框架。具体处理步骤如下:

#### 2.3.1 数据预处理

在经空间聚类得到热点区域以后,将每个区域内的微博文本集视为一个文档,在使用文本分析算法分析数据之前,需要先对数据进行清洗和分词处理,使之具有较高的质量并能够满足算法要求。

微博文本通常是一些中文句子,去除其中的标签(通常以两个#符号标识,如:#北京#)、标点符号、数字、表情符号(在本例数据中通常以中括号括起,如:[大笑])、特殊符号、网址链接(如:http: //t.cn/RyE32Wc)、提及某人(如:@xxx)等等(如表1,略去了发布时间、用户、坐标)。每条数据都有其对应的发布位置,可用于分类和筛选数据。

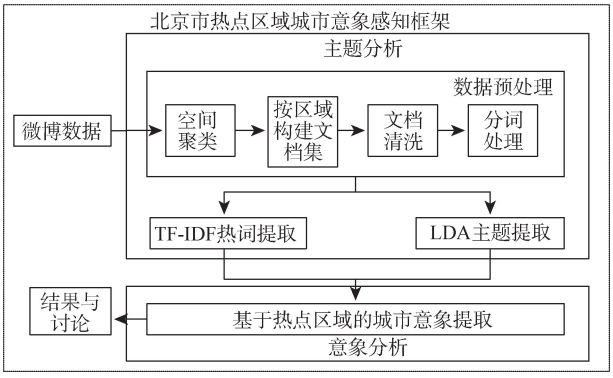


图1 北京市热点区域城市意象感知框架

Fig.1 Image perception framework of Beijing's regional hotspots

针对此类数据,本文采用以下数据预处理流程:

(1) 筛选和分类。根据数据的发布位置,按区域进行空间聚类处理,筛选和分类数据。将区域内的数据保留而去除区域外的数据。图2是部分区域(首都国际机场附近)数据的筛选示例,其中黑色的数据点将被去除。

(2) 构建文档集。将上述筛选并按各个地区分类后的数据,按各个地区组合成文档集,具体方法为:同一地区内的所有微博文本连接成一个长文本,这个长文本即本地区的文档,将所有地区的文档的集合作为文档集。

(3) 清洗文档。清洗每一个文档,去除其中不需要的字符,包括:标签、标点符号、数字、表情符号、特殊符号、网址链接等。具体地,使用正则表达式替换来完成上述处理。

(4) 分词并去除停用词。对经过清洗的每一个文档,使用分词系统(本文使用结巴分词系统)分词,

表1 微博数据示例

Tab.1 Examples of microblog data

| ID | 微博文本 |
|---|---|
| 1 | 新年快乐[礼物] http://t.cn/RyE32Wc |
| 2 | ♥Happy new year♥感谢一路有你们的陪伴! ♥ http://t.cn/z8A1inm |
| 3 | #一张照片再见2015#希望2016年更多快乐[心] http://t.cn/R2LTTrb |
| 4 | 2016[微笑] @小猪 http://t.cn/z8AckOe |

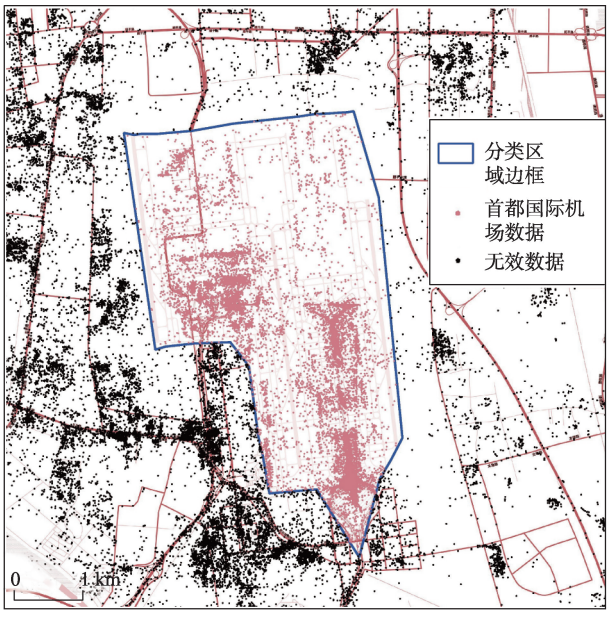


图2 首都国际机场附近地区数据筛选示意图

Fig.2 Spatial-clustering result near the Beijing Capital International Airport

并去除其中的停用词。使用一个完善的停用词表有助于提高结果的质量，也可以在得出结果后将结果中未去除的影响较大的停用词手动添加到停用词表中再重复计算。

经过这一步骤的处理后，得到由独立词语组成的文档集，可直接输入文本分析算法进行计算。

**2.3.2** 使用TF-IDF进行热词提取

TF-IDF是一种用于信息检索与数据挖掘的常用加权技术，它通过词的词频(Term Frequency, TF, 即某个词在文档中出现的频率)和逆向文件词频(Inverse Document Frequency，IDF，即使用该词的文档在文档集中的分布情况，使用该词语的文档越多，该词IDF值越小)评估单词在一篇文档中的重要性，具有较高词频和较低逆向文件词频的词被认为在文档中具有更重要的地位，可作为该文档的代表性标签。

TF-IDF权值的计算分为两部分，其公式为：

$$TFIDF_{i,j} = TF_{ij} \times IDF_i \quad (1)$$

式中：$TFIDF_{i,j}$为TF-IDF权值；$TF_{ij}$表示词语$i$在文档$j$中的词频，$IDF_i$表示词语$i$在所有文档中的逆向文件词频。将每个热点区域中的所有签到消息分词处理后得到一个文档，形成文档集合，进而使用TF-IDF算法对每个文档的每个词计算TF-IDF权值，并从大到小排列，由此可提取出微博数据中各热点区域对应的热词。

**2.3.3** 使用LDA法进行主题聚类

TF-IDF是分析文本数据的“显式”方法，简单直观，但其弊端是无法捕捉到隐藏在文本表面下的深层次信息。例如，两个同义词或两个具有强相关性的词，可能是由同一个主题产生的。一个文档虽然由很多词组成，而这些词背后的主题并不很多，而这些主题才是更为核心的信息。这种由词归类到主题的思路是各种文本主题模型(Topic Model)的共同思想。

LDA是一种著名的主题模型，同时它也是一种无监督学习算法，在训练时不需要对训练集进行手工标注，只需要文档集以及指定主题的数量$k$、迭代次数和狄利克雷参数即可。LDA的输入是由文档构成的集合(文档集)，其中的每一篇文档，则是由若干词语构成的，这些词语没有先后顺序；输出则为这些词语的主题聚类结果。

LDA模型中存在一个词袋(Bag of Words)，即主题—词对应关系。词袋描述了在某个主题中选词时，选中每个词语的概率。LDA模型中产生一篇文档的过程如下：首先确定该文档的主题分布，即该文档涉及的每个主题是什么以及每个主题所占的比重，然后按主题分布在确定的主题中选取一个主题，再在该主题下以词袋定义的词语概率分布选取一个词语，不断重复，从而生成一篇文档。LDA算法基于训练集，通过统计学原理对该过程进行逆向训练，可得到词的主题分布。

本文将2.3.1小节中数据预处理得到基于北京市热点区域的文档集后，经过统计得到文档—词矩阵，然后经LDA算法处理，为热点区域计算出相应的主题向量(包括各个主题的概率，以及主题下面对应的词语分布概率)，最终得出各热点区域的主题聚类结果。

## 3 结果分析

### 3.1 TF-IDF结果分析

**3.1.1** 总体评价

表2是TF-IDF算法结果的一部分，其中选取了几个典型区域排名前10的词语。

**表2　TF-IDF结果示例**

**Tab.2　Examples of the result of term frequency-inverse document frequency (TF-IDF) analysis**

| 名次 | 天安门 | 景山公园 | 故宫 | 鸟巢 | 三里屯 | 北海公园 | 圆明园 |
|---|---|---|---|---|---|---|---|
| 1 | 天安门 | 景山 | 故宫 | 鸟巢 | 三里屯 | 北海公园 | 圆明园 |
| 2 | 长安街 | 故宫 | 紫禁城 | 五月天 | 喜欢 | 荡起 | 遗址 |
| 3 | 升旗 | 紫禁城 | 天安门 | 演唱会 | 好吃 | 双桨 | 历史 |
| 4 | 永久 | 俯瞰 | 御花园 | 陈奕迅 | 优衣库 | 白塔 | 国耻 |
| 5 | 天安门广场 | 全景 | 角楼 | 现场 | 梦龙 | 小船 | 万园之园 |
| 6 | 升国旗 | 崇祯 | 红墙 | 阿信 | 终于 | 海面 | 勿忘 |
| 7 | 城楼 | 万春亭 | 午门 | 十年 | 开心 | 倒映 | 断壁残垣 |
| 8 | 祖国 | 日落 | 地方 | 建筑 | 太古 | 波浪 | 西洋楼 |
| 9 | 太阳升 | 中轴线 | 故宫博物院 | 一场 | 感觉 | 环绕着 | 荷花 |
| 10 | 故宫 | 北京城 | 皇帝 | 喜欢 | 生活 | 泛舟 | 大水法 |

从表2可以发现，TF-IDF分析的结果基本可以反映人们在不同区域的活动、话题和关注点，与区域的特点也较为相符，用其提取的关键词对区域的描述较为准确。在不同的热点区域，人们关注的话题主要可以分为以下几类：

(1) 城市标志性建筑、交通节点、场所。例如：①历史悠久或具有纪念、象征意义的建筑(群)，如故宫、天安门、人民大会堂等；②外形奇特的地标建筑，如中央电视台的大裤衩建筑、水立方等；③交通枢纽，如北京国际机场。以上对地标建筑、交通节点和场所的关注，可构建出人们心目中的北京市独特的城市意象空间。

(2) 城市活动。包括日常性城市活动、非常规大型活动。日常性城市活动如王府井的商业活动、故宫的游览活动、高校的学习活动、三里屯的休闲娱乐活动、在天安门观看升旗、在国家大剧院观看话剧、在798艺术区欣赏艺术作品、在北海公园划船等。非常规大型活动如在鸟巢举办的大型演唱会。由于大型活动往往伴随着大量突发流量，而常规性活动则细水长流，都能使相关话题在该区域的微博文本中占有相当比例。

(3) 地方文化。包括历史文化、民俗文化和当代文化。比如在北海公园，几十年前拍摄的少儿电影《祖国的花朵》的主题曲《让我们荡起双桨》的歌词被频繁提起；在雍和宫，上香与祈福是关键主题；而地坛最热门的词当属庙会。

(4) 重要事件和人物。在景山公园，人们会提到崇祯皇帝在此自缢；而在圆明园，人们会经常提起英法联军火烧圆明园的国耻；在三里屯，优衣库事件成为人们茶余饭后的热门谈资。

此外，还分析了热点词频在时间上的分布情况，发现人们在热点区域的话题取决于该区域本身的特点，关键词呈现大体稳定、随时间变化而波动的特征。其中，区域内的日常活动、地方产业、历史民俗文化、重要历史事件等话题保持相对稳定；而非常规活动与当代文化事件会随时间变化呈现出波动的特点，比如某场体育赛事、演唱会等事件主题的时间相关性就表现得相当明显。

**3.1.2** 区域间的相互联系

以上分析可以发现，在TF-IDF的结果中，在某些热点区域活动的人群经常会提及一些特定的地点，这意味着这些地点之间存在某种联系。通过共现分析可将热点区域与其他地点之间的联系挖掘出来。这就意味着这两个区域之间存在某种联系，使得大量的用户将其相提并论。概括而言，地点之间的联系可分为以下几种情况：

(1) 位置邻近：如故宫与天安门，北京大学旧地质馆与中国美术馆，什刹海与中国美术馆，等等。

(2) 主题相似：如都具有皇权色彩的故宫和颐和园；中国最好的两所大学：清华大学和北京大学。

(3) 中轴线：例如，景山公园可以俯瞰北京市的中轴线，因此景山公园的高频词会出现“中轴线”、“北京城”以及北京中轴线上的重要节点(如“故宫”、“天安门”等)，尤其是位置上也邻近的故宫更是占据了景山公园热词榜的第一位，可见在景山公园俯瞰故宫全景，也成为景山公园一大重要的旅游特色。

**3.1.3** 对现实的启示

TF-IDF计算结果在一定程度上反映了区域内关键词的情况，进而分析关注点、热点，可从人类认知角度获得大量城市意象相关信息，是热点区域意象感知的一种可行方案。

通过仔细分析热点区域的TF-IDF主题词分布，除可在实证层面揭示“人”对“地”的认知，也对现实世界的城市特色传承、空间品质塑造、旅游资源开发等具有较好的启发意义。从对鸟巢的关键词分析可以看出鸟巢在后续利用开发中所存在的问题，其基本上是用于旅游开发和举办大型演艺活动，而作为大众健身和职业体育赛事场馆的开发并不成功。从北海公园的标签《让我们荡起双桨》可见，创建景区文化品牌对于吸引游客具有很好效果，同时北海公园可以此为切入点改善自己的宣传策略。

**3.2 LDA结果分析**

**3.2.1** 总体评价

经过试验，LDA模型通过词的共现信息的处理，可以挖掘到文本表面下的很多有用信息。表3列举了一些重要的主题和各主题中重要的词。

由表3可见，LDA主题聚类结果较为清晰，具备一定的主题区别能力。如主题9是回家主题；主题18是旅游主题，体现了“暴走”、“逛”、“一日游”等旅游活动，其中风景和历史类景点受到游客的喜爱，还表现了“累”的情绪；主题31是互联网创业主题，涉及中关村附近的互联网企业，其中还提到“测试”、“加班”、“咖啡”等程序员工作状态，以及“直播”这个2016年互联网的最大风口；主题54是学习主题；主题2是圆明园主题，更明显地体现了“火烧圆明园”这一历史事件和“勿忘国耻”的爱国情绪，

表3 LDA结果主题示例

Tab.3 Examples of topics in the result of latent dirichlet allocation (LDA)

| 名次 | 主题9 | 主题18 | 主题31 | 主题54 | 主题2 | 主题7 | 主题21 | 主题46 |
|---|---|---|---|---|---|---|---|---|
| 1 | 回家 | 逛 | 创业 | 学习 | 圆明园 | 鸟巢 | 北大 | 王府井 |
| 2 | 再见 | 游 | 公司 | 老师 | 历史 | 演唱会 | 未名湖 | 烤鸭 |
| 3 | 火车 | 累 | 新浪 | 写 | 遗址 | 五月天 | 北京大学 | 全聚德 |
| 4 | 北京站 | 地方 | 中关村 | 学校 | 国耻 | Eason | 未名湖畔 | 小吃街 |
| 5 | 车 | 游客 | 测试 | 同学 | 勿忘 | 听 | 博雅 | 电影 |
| 6 | 火车站 | 一日游 | 微博 | 东西 | 断壁残垣 | 唱 | 燕园 | 大董 |
| 7 | 回 | 风景 | 大街 | 发现 | 之园 | 陈奕迅 | 校园 | 炸酱面 |
| 8 | 下次 | 逛逛 | 理想 | 努力 | 万园 | 十年 | 教授 | 小吃 |
| 9 | 晚点 | 脚 | 加班 | 电影 | 荷花 | 爱 | 听 | 长安街 |
| 10 | 坐 | 历史 | 大厦 | 事情 | 黑天鹅 | 一场 | 讲座 | 店 |
| 11 | 出发 | 旅行 | 再见 | 上课 | 皇家 | 青春 | 老师 | 王府井大街 |
| 12 | 车站 | 暴走 | 咖啡 | 早上 | 园子 | 第一次 | 塔 | 东来顺 |
| 13 | 小时 | 建筑 | 直播 | 未来 | 西洋楼 | 谢谢 | 同学 | 步行街 |
| 14 | 地铁 | 走走 | 互联网 | 几天 | 园林 | 三年 | 园子 | 外婆家 |
| 15 | 离开 | 一圈 | 下班 | 刷课 | 八国联军 | 嗨 | 食堂 | 酒店 |

也可看出西洋楼景区是人们游览的主要景点；主题7是鸟巢主题，可以看出鸟巢的后续利用主要是作为大型演唱会的举办场地，而2016年在鸟巢举办的演唱会中影响力最大的是陈奕迅和五月天的演唱会；主题21是北京大学主题，主要涉及北京大学校内的著名景点与大学学习生活相关的一些词汇；主题46是王府井主题，可以发现，在王府井逛街的人们最关注的是品尝北京特色美食，像全聚德、东来顺等老字号，以及大董烤鸭店和外婆家等餐厅都颇受欢迎。

**3.2.2** 分类比较

下面列举几种典型区域的LDA主题分布结果，并开展比较分析。主题分布结果以雷达图的形式展示，雷达图的径向表示概率的大小，越往外越大，雷达图的切向则表示主题编号。此外，雷达图的词语标注表示该编号主题下概率较高的词语，按概率从大到小排序。

(1) 高等院校

本文选取了几所高等院校：北京林业大学、清华大学、北京交通大学、北京大学(排名不分先后)的主题分布结果(图3)。

由图3可知，除北京交通大学以外，另外3所学校均有自己独立的主题，并且都比较符合高校的主题风格。表达“劳累”的主题在北京林业大学和北京交通大学所占的比例较高，而在清华大学和北京大学所占比例较低；这有可能是林大和交大的学生真的比清华北大的学生累，也有可能是他们更喜欢将这种累的情绪分享在社交媒体上。

表达学习的主题的概率在4所学校中几乎一致，但表达与期末、毕业、考试相关的主题概率在北京交通大学和北京林业大学中较高，而在清华大学和北京大学中较低，这可能有北大清华同时也是著名旅游景点的原因。但总体上而言，相对于平时的学习，北京交通大学、北京林业大学学生在微博中表现出来的对于考试主题的关注度更高。

(2) 皇家园林

本文选取了几座皇家园林：景山公园、颐和园、北海公园、圆明园(排名不分先后)的主题分布结果(图4)进行分析。可以看出，4处皇家园林均有符合自身风格的主题。颐和园、北海公园的主题合为一处，说明这两处园林具有很高的相似性，现实中也的确如此。从主题关键词看，颐和园、北海公园相对于其他两处园林的特色是“湖景”和“划船”，说明来到此处的人们更乐意享受泛舟湖中的惬意。同时，在前面TF-IDF中体现的北海公园的标志歌曲——《让我们荡起双桨》，在LDA主题中也有所体现。景山和故宫的密切关系在景山公园主题中体现得淋漓尽致；而圆明园因其过去的国耻事件，主题分布与其他各大园林风格迥异。此外，景山公园由于其重要的游览方式(即“俯瞰故宫全貌”)的影响，来这里的人们对于拍摄、纪念情有独钟；同时，人们多在景山俯瞰中轴线，视野易受雾霾影响，因

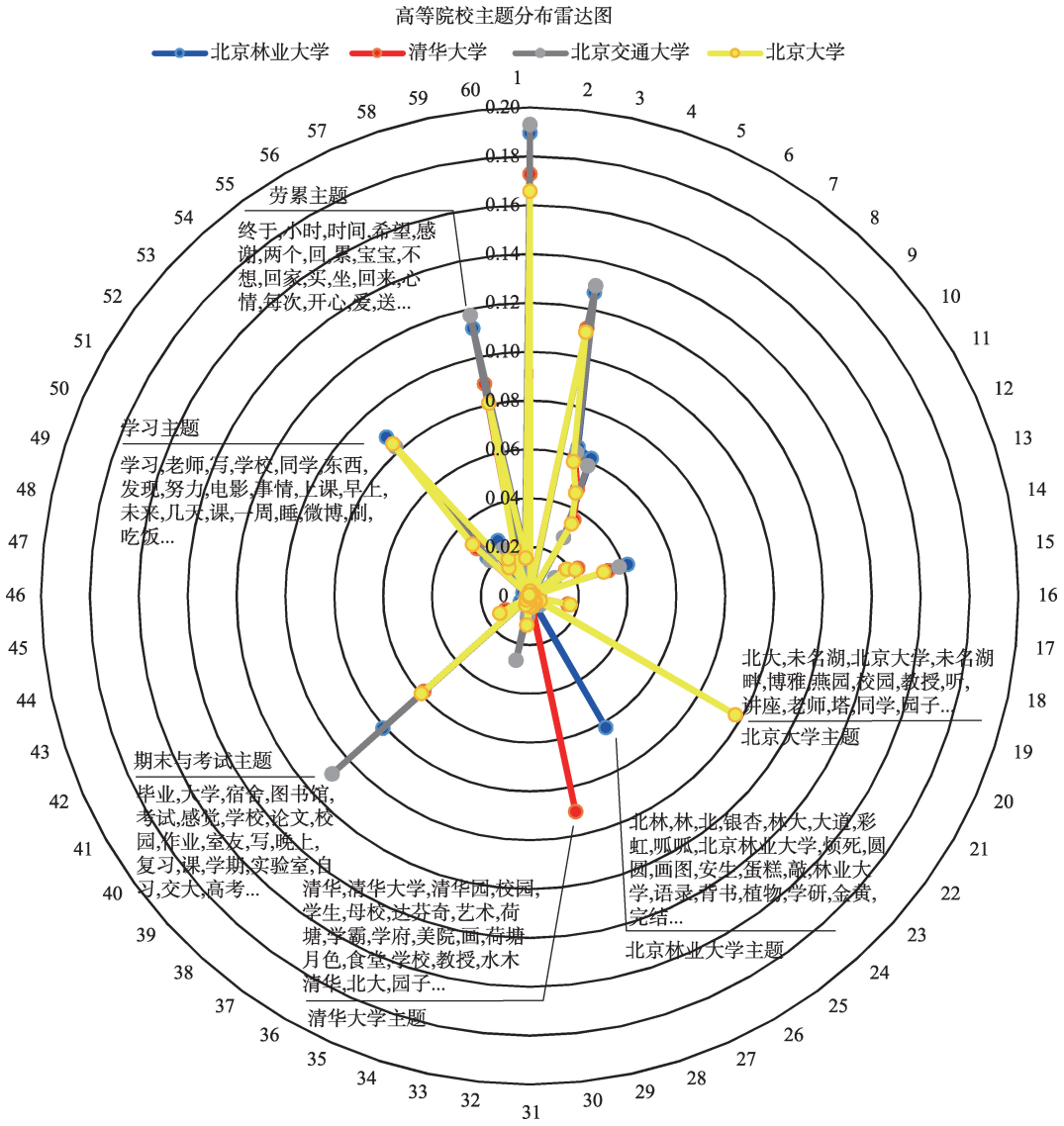


图3 高等院校主题分布

Fig.3 Topic distribution of universities

此在此处对雾霾的抱怨较多。

在游览皇家园林的时候,人们对“累”的表述呈现差异。其中,颐和园内微博用户“累”的比例最高,这应该与其面积大、内部景点多有关;圆明园次之,其实圆明园面积并不小,可能是其纪念性主题冲淡了对“累”的感知;而景山公园由于其面积较小,北海公园则是很多居民锻炼健身之所,这两处微博用户喊“累”的比例较低。

(3) 出入口

本文中选取北京市若干出入口进行主题的对比分析,包括:北京站(火车站)、北京南站(火车站)、首都国际机场(图5)。由图5可知,显示出入口可分为两类:机场和火车站。两个火车站主题类似,不过北京南站有较多去往天津、上海方向的旅客,并且较多乘高铁出行,这也与北京南站作为高铁站的现实相符。

此外可以发现,无论是机场还是火车站,“回家”和“旅行”总是不变的主题。不过相较于火车站的旅客,航班更易出现延误,所以机场旅客对晚点的抱怨也更多。

## 4 结论与讨论

本文以2016年北京市带位置签到的新浪微博数据为例,运用典型的文本分析方法,挖掘城市不同热点区域的主题,对北京市人群活动的热点区域

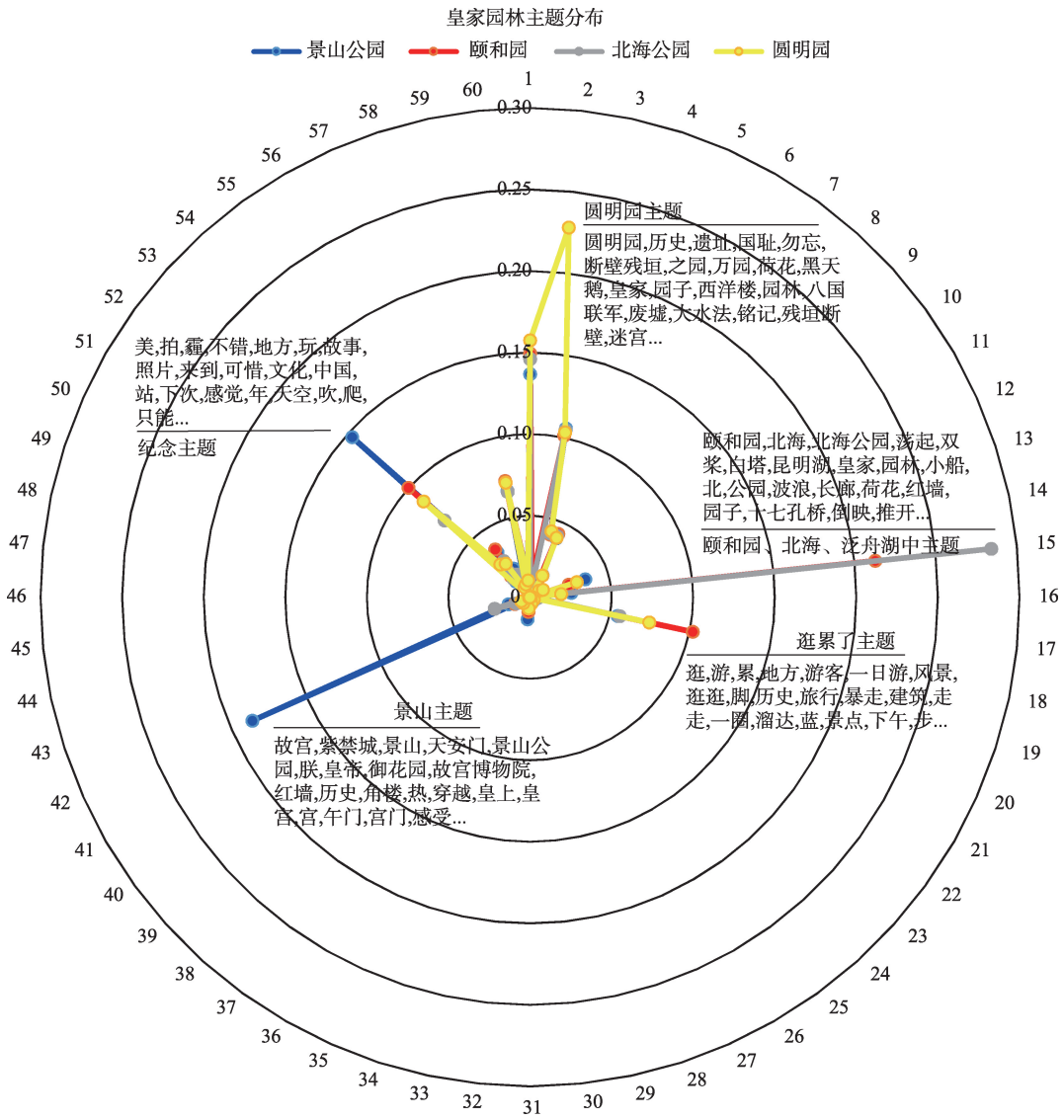


图4　皇家园林主题分布

Fig.4　Topic distribution of royal gardens

进行地理画像，以感知北京市不同热点区域的社会文化功能和人群行为，并在此基础上通过对热点区域高频主题词进行共词聚类分析，深度挖掘了北京市的总体意象。本文对于城市意象感知理论的主要贡献为：面向社交媒体大数据，在感知城市意象物质空间要素的基础上，探索了一种通过文本数据的主题分析提取城市意象非物质要素的路线，从而实现了城市意象的一体化综合感知。一方面可为同类研究提供城市意象感知的方法参考，另一方面以北京市热点区域为例的实证研究结果对现实中的城市特色传承与空间品质塑造等有一定的启发意义。同时，透过本文的实例研究发现，TF-IDF与LDA两种主题分析方法在城市意象的非物质要素感知方面各有所长：TF-IDF能非常直接地通过提取热词，有效反映单个区域的特色；LDA则可视为一种对词语的主题聚类方法，抽象层级更高，可更好地帮助研究人员对多个区域进行比较分析；应用两种方法能较好地实现对城市意象非物质要素的解构。

此外，需要注意的是，由于无法准确描述微博用户所代表的总体，所以微博签到数据在代表城市人群方面可能有偏差。据新浪微博公司统计，微博用户的主力群体是青年白领群体，30岁以下青年群体占比达到80%以上，拥有大学以上高等学历的用户占比高达77.8%。采用微博签到数据将会忽略一部分较少使用微博的社会群体，如儿童、老年人、贫

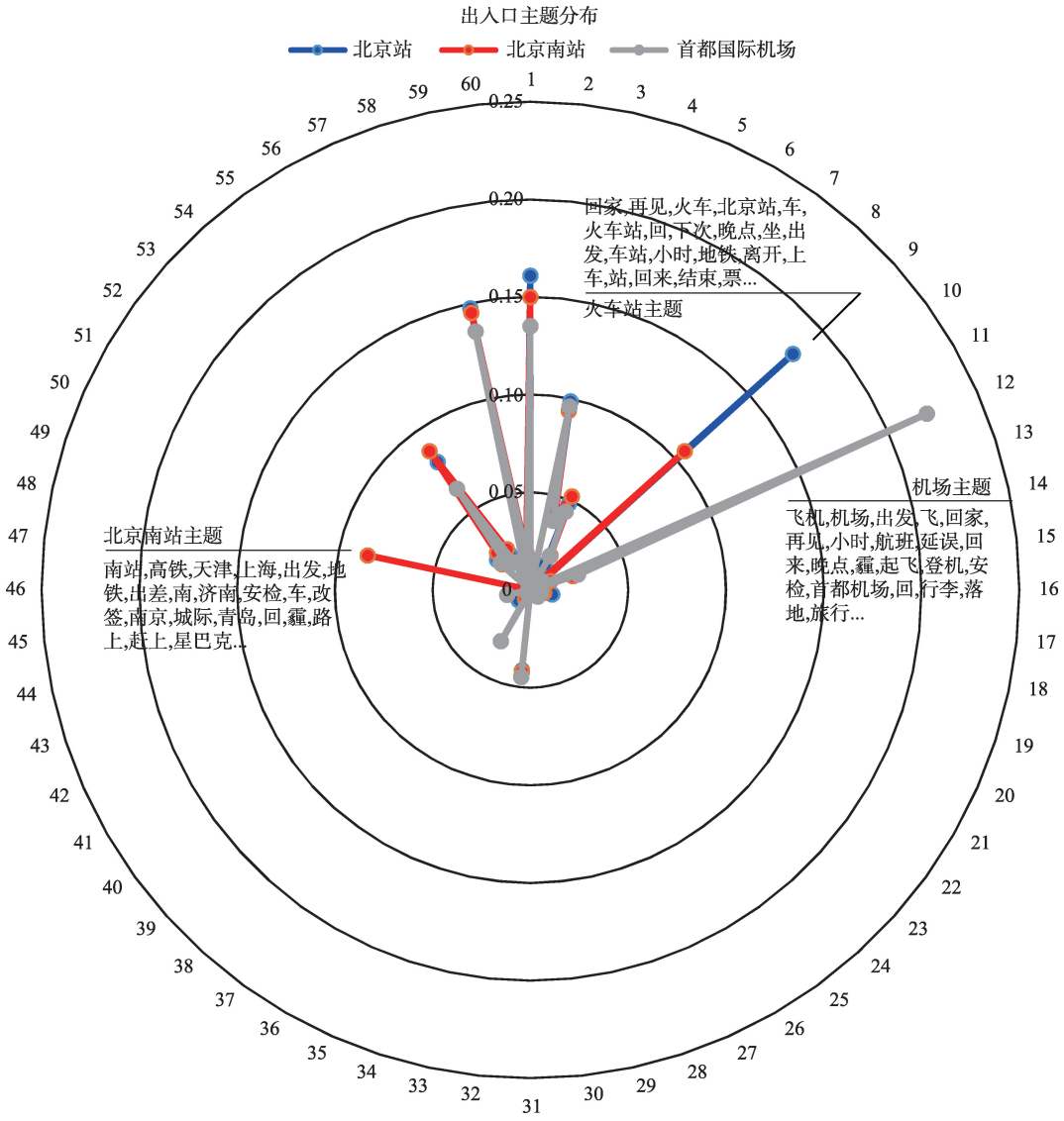


图5 出入口主题分布

Fig.5 Topic distribution of ports

困人群等。然而,即使如此,该数据依然能在一定程度上代表城市内部的人口及活动分布。随着未来关于社交媒体人群代表性研究的深入,该方法的效度和准确度也将进一步得到改进。

本文仍是在新数据环境下对城市意象综合感知的探索性研究,不足之处还表现在仅以北京市部分公共空间为例开展意象感知研究,城市意象的非物质要素感知未按用户群体细分,对意象感知结果缺乏深入的应用探讨等。因此,未来进一步的工作包括:

(1) 基于多种数据,力图构建更加完整的北京市"城市意象"。对各热点区域主题进行再聚类,对北京市区域的整体意象特征进行深度挖掘;引入更多数据,结合数据对图片内容进行分析;对城市中生活的人群观察者引入社会文化的考虑,按不同人群进行分类,对比不同群体对城市意向的感知差异;在时间维度上,由横向研究向纵向研究转变,由于微博产生的历史较短,未来可利用事件跨度更长的社交媒体数据,并结合历史书籍、图片、视频等资料,研究北京城市意向的演化机制。

(2) 优化意象感知的文本分析方法。在文本分析方法方面,TF-IDF算法可有效提取关键词,从而发现用户在某区域内谈论的焦点,并为进一步分析提供基础和依据;LDA主题提取效果显著,可较好地反映热点区域的主题、风格和特色,能提示很多在TF-IDF中并不显著的现象,以供进一步分析。

然而，也应看到上述两种方法还各有其优化空间，未来应当针对其不足，进行相应的优化或者选择更好的算法。比如，LDA算法，其不考虑词语在句子中的顺序，从而会遗失许多有用的语义信息，因此未来将在考虑词语顺序与语境的基础上，对主题分析模型进行进一步优化。

## 参考文献(References)


白凯, 马耀峰, 游旭群. 2008. 基于旅游者行为研究的旅游感知和旅游认知概念[J]. 旅游科学, 22(1): 22-28. [Bai K, Ma Y F, You X Q. 2008. Reflections on the conception of tourist perception and cognition based on researches of tourist behaviors[J]. Tourism Science, 22(1): 22-28.]

白凯, 张春晖, 郑荣娟, 等. 2011. 跨文化群体游客的中国旅游目的地意象色彩认知[J]. 地理科学进展, 30(2): 231-238. [Bai K, Zhang C H, Zheng R J, et al. 2011. Color cognition of China tourism destination image within the groups of cross-culture tourists[J]. Progress in Geography, 30(2): 231-238.

白凯, 赵安周. 2011. 城市意象与旅游目的地意象研究中的趋同与分野[J]. 地理科学进展, 30(10): 1312-1320. [Bai K, Zhao A Z. 2011. Studies on convergence and divergence of city image and destination image[J]. Progress in Geography, 30(10): 1312-1320.]

曹越皓, 龙瀛, 杨培峰. 2017. 基于网络照片数据的城市意象研究: 以中国24个主要城市为例[J]. 规划师, 33(2): 61-67. [Cao Y H, Long Y, Yang P F. 2017. City image study based on online pictures: 24 cities case[J]. Planners, 33(2): 61-67.]

邓力凡, 谭少华. 2017. 基于微博签到行为的城市感知研究: 以深港地区为例[J]. 建筑与文化, (1): 204-206. [Deng L F, Tan S H. 2017. Study of city perception based on microblog sign in behavior: A case study of Shenzhen and Hong Kong[J]. Architecture & Culture, (1): 204-206.]

樊蕾. 2013. 微博叙事中的社会图景: 基于新浪微博“微话题”的研究[D]. 上海: 华东师范大学. [Fan L. 2013. Social spectale in micro-blog narrative: Based on the micro-topics of Sina micro-blog[D]. Shanghai, China: East China Normal University.]

顾朝林, 宋国臣. 2001. 北京城市意象空间及构成要素研究[J]. 地理学报, 56(1): 64-74. [Gu C L, Song G C. 2001. Urban image space and main factors in Beijing[J]. Acta Geographica Sinica, 56(1): 64-74.]

凯文·林奇. 2001. 城市意象[M]. 方益萍, 何晓军, 译. 北京: 华夏出版社. [Kevin L. 2001. The image of the city[M]. Fang Y P, He X J, Trans.. Beijing, China: Huaxia Publishing House.]

李郇, 许学强. 1993. 广州市城市意象空间分析[J]. 人文地理, 8(3): 27-35. [Li X, Xu X Q. 1993. A spatial analysis of the image of Guangzhou City[J]. Human Geography, 8(3): 27-35.]

刘瑜. 2016. 社会感知视角下的若干人文地理学基本问题再思考[J]. 地理学报, 71(4): 564-575. [Liu Y. 2016. Revisiting several basic geographical concepts: A social sensing perspective[J]. Acta Geographica Sinica, 71(4): 564-575.]

龙瀛, 周垠. 2017. 图片城市主义: 人本尺度城市形态研究的新思路[J]. 规划师, 33(2): 54-60. [Long Y, Zhou Y. 2017. Pictorial urbanism: A new approach for human scale urban morphology study[J]. Planners, 33(2): 54-60.]

沈益人. 2004. 城市特色与城市意象[J]. 城市问题, (3): 8-11. [Shen Y R. 2004. Necessity of city image research from the aspect of city characteristic[J]. Urban Problems, (3): 8-11.]

宋蕾, 张培晶. 2014. 基于LDA主题建模的微博舆情分析系统研究[J]. 网络安全技术与应用, (4): 5-6. [Song L, Zhang P J. 2014. System design of micro-blog public opinion based on LDA topic modeling method[J]. Network Security Technology & Application, (4): 5-6.]

田逢军, 沙润. 2008. 城市旅游地意象空间分析: 以南昌市为例[J]. 旅游学刊, 23(7): 67-71. [Tian F J, Sha R. 2008. A spatial analysis of the image of urban tourist destinations: A case study on Nanchang[J]. Tourism Tribune, 23(7): 67-71.]

Tuan Y F. 1997. 经验透视中的空间和地方[M]. 潘桂成, 译. 中国台北: 台北编译馆. [Tuan Y F. 1997. Space and place: The perspective of experience[M]. Pan G C, Trans.. Taibei, China: Taipei Institute for Compilation and Translation.]

汪静莹, 朱廷劭, 郝碧波, 等. 2016. 微博用户生活满意度微博语言及行为特征分析[J]. 中国公共卫生, 32(2): 225-229. [Wang J Y, Zhu T S, Hao B B, et al. 2016. Life satisfaction among microblog users: An analysis on linguistic and behavior features[J]. Chinese Journal of Public Health, 32(2): 225-229.]

徐磊青. 2012. 城市意象研究的主题、范式与反思: 中国城市意象研究评述[J]. 新建筑, (1): 114-117. [Xu L Q. 2012. The rethinking of themes and paradigms: A review of urban image studies in China[J]. New Architecture, (1): 114-117.]

钟栎娜. 2015. 旅游地感知结构重构: 基于文本与复杂网络分析的研究[J]. 旅游学刊, 30(8): 88-95. [Zhong L N. 2015. A reconstruction of destinations' perception structure based on the context and complex network analysis[J]. Tourism Tribune, 30(8): 88-95.]

周尚意, 唐顺英, 戴俊骋. 2011. “地方”概念对人文地理学各

分支意义的辨识[J]. 人文地理, 26(6): 10-13, 9. [Zhou S Y, Tang S Y, Dai J C. 2011. Identification of the significance of the concept of place to branches under human geography[J]. Human Geography, 26(6): 10-13, 9.]

Blei D M, Ng A Y, Jordan M I. 2003. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 3(4-5): 993-1022.

Cai J X, Huang B, Song Y M. 2017. Using multi-source geospatial big data to identify the structure of polycentric cities [J]. Remote Sensing of Environment, doi: 10.1016/j.rse.2017.06.039. (in Press)

Golder S A, Macy M W. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures[J]. Science, 333: 1878-1881.

Gordon C. 1978. Human aspects of urban form: Towards a man: Environment approach to urban form and design: Amos Rapoport[J]. Urban Ecology, 3(4): 385-386.

Gulick J. 1963. Images of an Arab city[J]. Journal of the American Institute of Planners, 29(3): 179-198.

Kaplan A M, Haenlein M. 2010. Users of the world, unite! The challenges and opportunities of social media[J]. Business Horizons, 53(1): 59-68.

Klein H J. 1967. The delimitation of the town-centre in the image of its citizens[M]//Brill E J. Urban core and inner city. Netherlands: University of Leiden: 286-306.

Lee W, Gretzel U, Law R. 2010. Quasi- trial experiences through sensory information on destination web sites[J]. Journal of Travel Research, 49(3): 310-322.

Liu L, Zhou B L, Zhao J H, et al. 2016. C-IMAGE: City cognitive mapping through geo-tagged photos[J]. GeoJournal, 81 (6): 817-861.

Luo Q J, Zhai X T. 2017. "I will never go to Hong Kong again!" How the secondary crisis communication of "Occupy Central" on Weibo shifted to a tourism boycott[J]. Tourism Management, 62: 159-172.

Salesses P, Schechtner K, Hidalgo C A. 2013. The collaborative image of the city: Mapping the inequality of urban perception[J]. PLoS One, 8(7): e68400.

Wong C U I, Qi S S. 2017. Tracking the evolution of a destination's image by text- mining online reviews: The case of Macau[J]. Tourism Management Perspectives, 23: 19-29.

# Image perception of Beijing´s regional hotspots based on microblog data

XIE Yongjun[1], PENG Xia[2*], HUANG Zhou[1], LIU Yu[1]

(1. Institute of Remote Sensing and Geographical Information System, Peking University, Beijing 100871, China;
2. Collaborative Innovation Center of Tourism, Beijing Union University, Beijing 100101, China)

**Abstract:** Research on "city image" can facilitate urban culture perception, urban management and planning, and tourism resource development. In recent years, as intelligent mobile terminals and social media apps became increasingly popular, a large number of social media geo-tagged data containing text and location information have been generated, providing a new solution for city image perception studies. This article uses the social media geo-tagged data (Sina weibo check-in data in Beijing, 2016) to explore regional hotspots through spatial clustering, and mining the topics of different hotspots through two typical methods— term frequency-inverse document frequency (TF-IDF) and latent Dirichlet allocation (LDA). The results reflect the topics that users were concerned about and discussed in different places, revealing the culture, functions, and characteristics of diverse places of Beijing in great depth. The proposed city image abstraction approach by integrating text mining and spatiotemporal big data analysis can promptly expose the differences on themes of activities, attitudes, and preferences in different places in Beijing, thus reveal the social and cultural characteristics of the city. Our method is an important complement to the five-element model of city image, which focuses on the urban material form. In addition, the case study results of Beijing regional hotspots facilitate the preservation of city characteristics and shaping of space quality.

**Key words:** geospatial data; social media; microblog data; text mining; regional hotspot; city image