基于大数据的文化遗产认知分析方法
——以北京旧城中轴线为例

杨微石[1,2,3],郭旦怀[4,5*],逯燕玲[6],王德强[4,5],朱映秋[4,5],张宝秀[6]

(1. 中山大学地理科学与规划学院 广东省城市化与地理环境空间模拟重点实验室,广州 510275; 2. 中国科学院地理科学与资源研究所,北京 100101; 3. 中国科学院陆地表层格局与模拟重点实验室,北京 100101; 4. 中国科学院计算机网络信息中心,北京 100190; 5. 中国科学院大学, 北京 100049; 6. 北京联合大学, 北京 100101)

摘 要:以北京旧城中轴文化遗产为例,利用2012、2015年的相关微博、报刊新闻、学术文献数据,通过提取关键词,抽取词频、tf-idf权重、互信息、后验概率等特征,从群体、时间、空间多个维度分析文化遗产的认知。在人群维度上,通过具有特征性人群的传媒信息,发现不同人群对文化遗产的认识存在异同:对于中轴文化遗产核心单元故宫、天安门、天坛的认知相对一致,而对于钟楼鼓楼、太庙、地安门的认识,官方偏向于行政管理,学者偏向于历史价值,大众则偏向于生活化。在时间维度上,提取文化遗产关注程度和认知变化。如相对于2015年,大众对故宫、天安门的关注程度相对提高,对太庙的历史价值认识更为丰富。大众相对于官方和学者对文化遗产的认知更容易发生变化,且对热点事件敏感。在空间维度上,挖掘文化遗产单元之间的认知转移和关联模式,一方面,空间上相连的天安门—正阳门—正阳门大街具有较高的双向认知;另一方面,中轴文化遗产中,故宫、天安门、天坛的后验概率较高,表现出跨空间的认知汇聚模式。基于大数据的认知分析方法,是问卷调查、文献调研、访谈分析等传统方法的重要补充方式,能够降低数据收集者的主观影响,增加分析维度和效率,有助于发现隐含的知识和模式。本文结论可为文化遗产价值挖掘、保护提供决策支持。

关 键 词:大数据分析;数据挖掘;文化遗产感知;tf-idf权重;北京中轴线

1 引言

文化遗产的认知是文化遗产现状、文化遗产的保护与改造、相关街区的规划与发展等研究中的一个重要的方向(刘祎绯, 2015)。文化遗产认知涉及遗产本身的科学、美学、历史文化价值(李乾夫, 2009),以及多样性的文化和地理环境特征等要素(刘丽华等, 2009; 王军, 2016)。目前,对文化遗产认知的分析多使用问卷调查、访谈调研等传统社会调查方法(沈加锋, 1989; Babbie et al, 1998; 李路珂, 2003; 陆建松, 2010; 包书月等, 2011)。

传统文化遗产认知调查方法,是在投入较多人力和时间条件下,从宏观的角度采取定量的手段、依据客观的验证来认识和说明社会现象的调查研究方式,包括抽样、统计分析、问卷调查或访谈。抽样是决定调研的代表性,问卷或访谈是进行变量测量和资料收集的工具,统计分析则是通过对问卷和访谈数据的处理和统计,分析总体的调研结果。问卷调查与访谈调研方法存在以下不足:①数据更新周期相对较长;②数据的获取容易受到调研者主观和能力的影响;③样本量相对较小;④数据系统性和可分析性较弱(王俊芳等, 2004)。

大数据方法的文化遗产认知,就是利用网络舆情数据,如推特、微博等社交媒体数据来描述社会

收稿日期:2017-06;**修订日期**:2017-09。

基金项目:国家自然科学基金项目(41371158, 41371386);北京市自然科学基金项目(9172023)[**Foundation**: National Natural Science Foundation of China, No.41371158, No.41371386; Beijing Natural Science Foundation, No.9172023]。

作者简介:杨微石(1985-),男,江西赣州人,博士生,主要研究方向为土地利用变化及其生态效应,E-mail: 43106363@qq.com。

通讯作者:郭旦怀(1973-),男,江西南康人,博士,副教授,主要从事高性能地理计算时空大数据研究,E-mail: guodanhuai@cnic.cn。

引用格式:杨微石, 郭旦怀, 逯燕玲, 等. 2017. 基于大数据的文化遗产认知分析方法: 以北京旧城中轴线为例[J]. 地理科学进展, 36(9): 1111-1118. [Yang W S, Guo D H, Lu Y L, et al. 2017. Analyzing perception of cultural heritage sites based on big data: A case study of Beijing Central Axis[J]. Progress in Geography, 36(9): 1111-1118.]. DOI: 10.18306/dlkxjz.2017.09.007

认知，在海量的网络文本数据中，通过提取关键词来过滤信息，浓缩有用的观点和知识，借助社交媒体中话题的热度有效地刻画大众对特定问题的关注度和意识。基于大数据方法的文化遗产认知研究已经有了很多成功的应用(计维斌等, 2013; Förster et al, 2015; Yin et al. 2015)。

相对于传统的调研的方法，大数据可以通过海量数据囊括社会各个层面的人群，覆盖不同时间段，具有更大的样本量和更短的更新周期，以及更好的数据可获得性、系统性、多维性和可分析性，更易于展开人群、时间、空间等维度的多维分析，便于全面、迅速地掌握文化遗产认知的变化。但大数据方法也存在数据有偏性、过拟合等不足(杨振山等, 2015)。

为了解社会公众对文化遗产的看法、观点，本文构建关键词，并基于语义的本体网络扩充关键词。设置爬虫，自动抓取相关信息，采集大量网络文本数据；进行数据清洗，然后提取特征关键词，基于信息源差异，了解不同人群维度对文化遗产认知的异同。通过时间的对比，分析社会认识的时间维度变化。同时，基于词频数据和文化遗产地理空间上的分布，挖掘文化遗产社会认知的空间维度信息。尝试将基于大数据分析的方法应用于对文化遗产的多维认知分析。

2 研究对象

北京旧城中轴线对北京市的城市及文化发展格局有着重要的意义(李琛, 2015)。中轴线是一条南北走向的轴线，从南向北，由永定门、正阳门、天安门、端门、午门、故宫，景山、地安门、钟楼鼓楼构成(李晨, 2011)(图1)。历史可以追溯到元朝，经历700余年的岁月，积淀了厚重的历史文化价值(李路珂, 2003)。北京的中轴被誉为北京独有的壮美秩序的源头(梁思成, 1951)，是一种历史性城市空间组织与设计的思想和方法，对城市形态和城市肌理发展起到重要影响(成亮, 2009)，为城市线状文化扩张的典型代表。

北京旧城中轴线，越来越受到广泛的关注和重视。“十二五”以来，北京旧城中轴线各段被纳入历史文化名城建设规划的保护体系。《北京城市总体规划(2004-2020年)》进一步强调了北京中轴作为城市发展的重要性，指出中轴线以文化功能为主。2008年奥运公园和奥林匹克森林公园成为中轴的北段延长，进一步体现了中轴线作为文化和城市格局发展轴线的重要作用。北京中轴线的研究对旅游线路的设计、城市及文化的发展有着重要意义。

本文以北京旧城中轴线文化遗产为研究对象，结合中轴线地理空间结构，从历史作用、官方认知、大众认知3个维度研究大众对中轴线文化遗产的认知变迁。试图进一步理解北京中轴线对城市空间布局、社会人文环境等方面的作用，挖掘中轴线文化遗产的文化价值和社会功能。

3 研究方法

本文采用大数据分析方法，采集、清洗、提取、归一化数据，通过具有不同人群指针的数据来源信息分析各人群的认识。从微博、报刊新闻、学术文献3个角度，分析大众、学者、官方3类群体对历史文化遗产认识的异同。对比2012年和2015年两年数据，了解公众对于中轴线历史文化遗产认知的变化；利用关联分析，了解历史文化遗产在空间上的联系规律。从人群、时间、空间3个维度研究公众对于历史文化遗产的认识(图2)。

图1 北京城市中轴线空间位置关系图

Fig.1 Spatial relation of cultural heritage sites along the central axis of Beijing

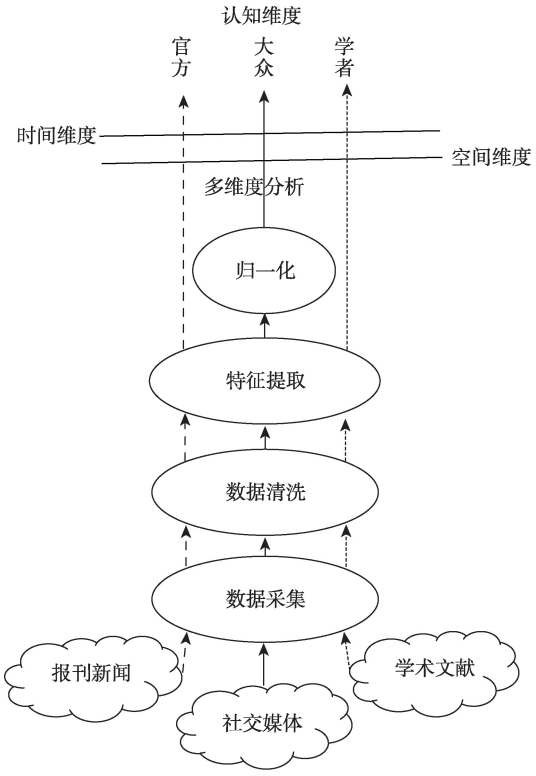

图2　基于大数据的文化遗产认知分析框架

Fig.2　Framework of big data-driven perception analysis on the cultural heritage sites along the central axis of Beijing

3.1 数据采集

微博等社交媒体、报刊新闻、学术文献等是文化社会认知的重要数据来源。社交媒体数据来源于新浪微博；报刊新闻、科研文献来源于中国知网(http://www.cnki.net)。本文收集2012年、2015年两年数据，数据量在清洗前达到651.20万条；清洗后为132.30万条。

数据具体获取方法为：通过部署分布式的数据爬虫，爬取新浪微博、主要期刊、在线报纸、新闻中与北京城市中轴线相关的文本。首先设置与北京旧城中轴线有关的关键词，如太庙、故宫、天安门等，然后基于语义的本体网络(Sangpachatanaruk et al, 2004)扩充关键词。例如，将“太庙”与太庙的另一个名称“劳动人民文化宫”关联。最后将文本中出现的高频相关词语也加入关键词中。通过设置的这些关键词，自动爬取相关文本。

3.2 数据清洗

数据清洗的目的在于消除原始文本数据中存在的不完整、数据重复、语义歧义和低价值、指向性不明确的信息，例如：为发现北京中轴线上的天桥地区相关信息，在搜索关键词“天桥”时，会返回很多无关信息，需要删除。数据清洗通常使用空值、错误值、不一致数据、不完整数据、错误数据类型、干扰与异常数据、重复记录等清洗方法(Dallachiesa et al. 2013; Gueta et al, 2016)。

在以上方法的基础上，还使用了以下清洗数据的方法：①增加限定语，如“北京天桥”等；②根据发信息者的IP地址，去除非北京地区发送的信息；③根据微博的前后时空关联的延续性，剔除一些只发表非北京地区信息的微博；④根据内容的相关性，剔除与社会认知无关的评论、历史事件等；⑤剔除非人群特征性信息，如除微博中的非大众认知信息，如转发的报纸、学术期刊、官方信息等。

3.3 特征提取

扫描清洗后的文本，将本体网络中关键词词频，作为描述相应对象的关注程度。另外，文本使用tf-idf权重来确定关键词的重要程度(Zhang et al, 2011)，衡量文化遗产和相关关键词之间的关联性。tf-idf权重在基于大数据分析的应急舆情管理等领域已经有了成功的应用(Wu et al, 2008)，分析方法已经较为成熟(Fabret et al, 2001; Wu et al, 2008; Rao et al, 2012)。通过设置阈值和排序，筛选权重较高的关键词。tf-idf值计算公式为(Aizawa, 2003)：

$$tf(x)-idf(x)=(1+tf(x))\cdot idf(x) \tag{1}$$

$$idf(x)=\log\frac{N}{1+|\{d\in D:x\in d\}|} \tag{2}$$

$$tf(x)=\frac{f_x}{N_d} \tag{3}$$

式中：x 表示文档 D 中的关键词；$tf(x)$ 为某一特征值在文档中出现的频数；$idf(x)$ 为某一特征值在整个文档集合的分布情况；N 为 D 文档数据集中文本总数，D 为某个文化遗产的描述数据集；$\{d\in D:x\in d\}$ 为文档 D 中包含关键词 x 的文本数；f_x 为关键词出现的次数；N_d 为词的总数。

3.4 归一化

归一化主要针对微博数据。获取微博文本的数量变化，但总数据量的减少与大众对文化遗产的关注和认知并无直接关联，因此为了对比不同时间段、不同主体的数据，必须数据归一化。本文将特征数据归一化到[0,1]区间。使用下列公式归一化：

$$n_t(y)=\frac{tf_t(y)}{\sum_{x\in S}tf_t(y)} \tag{4}$$

式中：$tf_t(\cdot)$ 表示词语在时间 t 的词频；$n_t(\cdot)$ 表示归一化后的特征；S 表示关键词的集合。

3.5 数据分析

3.5.1 关联度分析

本文采用互信息(Peng et al, 2005)作为衡量地理对象之间的关联程度，表示两个文化遗产要素在文档中同时出现的概率大小。计算公式为：

$$I(x,y)=\sum_{x}\sum_{y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (5)$$

式中：$I(x,y)$为互信息；x和y表示不同的文化遗产；$p(\cdot)$表示词语在整个数据样本中的分布密度。

3.5.2 空间向量关联分析

基于文本数据认知分析的基础，引入空间维度。探求文化遗产单元在地理空间上的关联关系，挖掘空间维度上的向量关联模式。

北京旧城中轴线及覆盖的文化遗产单元具有明显的南北串联地理相关分布模式，在关联度分析中，各个文化遗产单元之间的互信息，是无向的，无法表达两个单元之间的相对关系。本文利用后验概率，来描述文化遗产单元之间的有向联系。给定文化遗产单元y，文化遗产单元x的后验概率为：

$$P(x|y)=\frac{p(x,y)}{p(y)}=\frac{c(x,y)}{c(y)} \quad (6)$$

式中：$c(y)$表示搜索y得到的信息数目；$c(x,y)$表示搜索y返回的微博中出现x的信息数目。在这种场景下，后验概率可直观理解为：当大众提到文化遗产y时，同时提到x的概率。它表达出单元之间的相对联系。如：去景山的游客会在景山上眺望故宫，因而P(“故宫”|“景山”)较高，反映出游客的认知中存在从景山到故宫的注意力转移，而反过来从故宫到景山的转移较少，P(“故宫”|“景山”)较低。

空间向量关联维度可以用矩阵表示，将后验概率映射到空间分布上，输出热力图，便于直观地发现空间—语义模式。如式(7)所示。矩阵元素$[P]_{ij}$表示给定第j个单元，第i个单元在微博中出现的后验概率$P(u_i|u_j)$。

$$P=\begin{bmatrix}1 & P(u_2|u_1) & \cdots & P(u_n|u_1)\\ P(u_1|u_2) & 1 & \cdots & P(u_n|u_2)\\ \vdots & \vdots & \ddots & \vdots\\ P(u_1|u_n) & P(u_2|u_n) & \cdots & 1\end{bmatrix} \quad (7)$$

计算出该矩阵后，再根据后验概率的值将矩阵绘制成热力图，图上每个方块的颜色深浅表示对应单元的后验概率。由此得出的热力图既能通过方块颜色反映向量关联相对关系，又具有中轴线空间结构，能够直观地发现空间向量关联模式。

4 结果与分析

4.1 人群维度分析

不同的传媒渠道，具有表达特征人群信息的功能。微博更倾向于表达普通大众的认识，报刊新闻是官方意识的表达渠道，而学术期刊则是学术界表达信息的平台。因此我们通过具有特定人群信息功能的信息来源平台，了解不同人群对文化遗产的认识。大体可归纳为以下两种情况。

(1) 官方、大众和学术界认识差异较大。如地安门、钟楼鼓楼在报纸新闻中，建筑、旧城、改造、交通、整治、补偿、工程、文化城、规划、木质、艺术、旅游、开发、历史文化保护等特色关键词出现了较高的tf-idf权重。说明官方认为地安门、钟楼鼓楼是旧城改造的重要组成部分，是历史文化保护区和旅游开发区，总体是一个行政管理的重要区域。而在微博上，小吃、美食、味道、开心、风味、文玩、佛殿、表演、心情、记忆、新年等特色词出现了较高的权重，说明在大众认知里，地安门、钟楼鼓楼是一个休闲娱乐场所。在学术期刊中，出现一些如建筑群、设计、风貌、名人故居、梁思成、国家文物局等高频特色词。可以看出地安门在学者认知中更多的是历史事件和文物古迹。

群体之间认知也出现了完全相反的格局。如太庙在建国后被改造为劳动人民文化宫，其功能发生了巨大转变，功能的转变对其社会认知产生深远的影响。在微博中，婚纱、婚纱照、音乐、展览等关于婚纱摄影和作为劳动文化宫功能的特色关键词权重之和高达51.69%；而关于祭祀、奉先殿、祖宗、大典等与祭祀有关的特色关键词词频权重之和仅为21.93%。不难看出，其作为劳动人民文化宫、婚纱照拍摄点在大众认知形象中，远大于历史上作为祭祀祭祖的功能形象；而在学术期刊则正好出现相反的状况，其关于祭祀的特色词频权重之和高达62.37%，关于婚纱摄影、劳动文化宫则仅为12.58%。

(2) 官方、大众、学者具有较一致的认识。故宫、天坛、天安门的关键词中，历史、文物、古迹、祈年、保护、皇帝、博物馆、中国、开会等词汇在3种传媒方式中具有较高的权重。可以看不同人群对这些中轴线上最核心的历史文化遗产单元具有较为一致的认识。

可以看出，中轴核心文化遗产故宫、天安门、天坛往往也是国家文化的代表，具有一定的权威性、正统性、受商业影响较小，官方、学者、大众3个人群

的认知相对一致。而太庙、钟楼鼓楼、地安门，政府则偏向于改造、保护和管理，学者偏向于历史价值，而大众的认知则更为偏向生活化。

4.2 时间维度分析

本文计算了上述关键词的词频、tf-idf权重，从中抽出较为显著的关键词来作为中轴线文化遗产单元的社会认知分析的参考。

对比2012年、2015年北京旧城中轴文化遗产的相关微博比例，可以发现社会大众对中轴线相关文化遗产景点的关注变化。图3显示北京旧城中轴线从北至南各个主要文化遗产相关微博数量的比例。可以看出，北京旧城中轴线内部各单元受关注程度的分布是非常不均匀的。从2012年到2015年，中轴线上最核心的区域——故宫、天安门受关注程度保持在较高的水平，并持续增加，而钟鼓楼、地安门、正阳门等则相对有所下降，中轴线文化遗产单元关注度的分布不均进一步加剧。

时间维度的分析，也能显现单个文化遗产在社会认知上的转变，可从侧面反映出当时社会公众关注的热点变化和宣传引导的作用。由于2012年开始清宫戏愈发火热，随着许多书法大赛、国学活动在太庙举办，太庙所蕴含的历史文化积淀得到越来越多的关注，关于其建筑艺术、祭祀传统的关键词权重有所增加。2015年太庙举办的跨年活动，则引起了跨年关键词权重的增加(图4)。

另一个典型例子是社稷坛：从2012年到2015年，“孙中山”、“祭祀”、“五色土”、“文化”、“历史”等词权重和比例明显增加，反映出随着对其历史文化资源的利用与开发加大，公众对其历史价值认识更丰富，社稷坛历史文化教育功能渐渐强化。

在官方、学者方面，通过对比2012、2015年差异，可以发现各关键词权重变化较小，说明官方、学者对文化遗产的认知更稳定，受热点事件的敏感度相对较小。

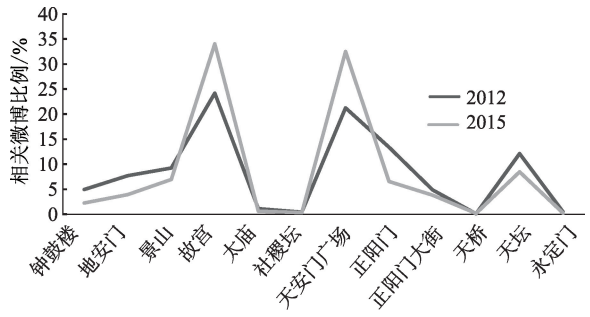

图3　2012、2015北京城市中轴线文化遗产相关微博比例

Fig.3　Proportions of microblogs related to the cultural heritage sites along the central axis of Beijing, 2012 and 2015

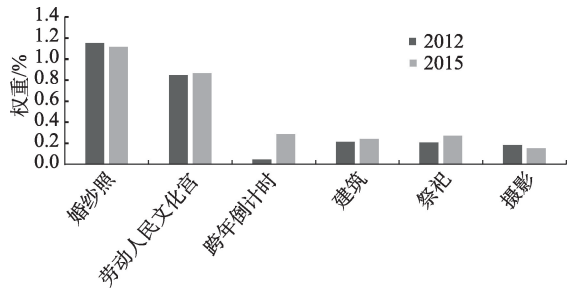

图4　太庙相关微博中特色关键词权重

Fig.4　Weights of keywords describing characteristics of the Imperial Ancestral Temple in relevant microblogs

4.3 空间维度分析

4.3.1 中轴线文化遗产单元认知转移分析

本文综合基于词频的后验概率和空间分布，用空间—语义分析方法，将后验概率映射到地理空间分布上的热力图，热力图表示相对于中轴线文化遗产单元来说，横轴地理单元的后验概率。颜色深浅表示后验概率的高低(图5)。

通过对比后验概率热力图，可以发现两种模式：

(1) 空间维度和语义维度都高度关联的模式。如图中红色实线方框区域所示，一些局部密切关联的区域如：景山—故宫、天安门—正阳门—正阳门大街。这些单元在邻接矩阵和热力图上的关联关系相一致，颜色较深，具有较强的双向认知转移能力，反映在这些地理上邻接的单元在大众的认知中

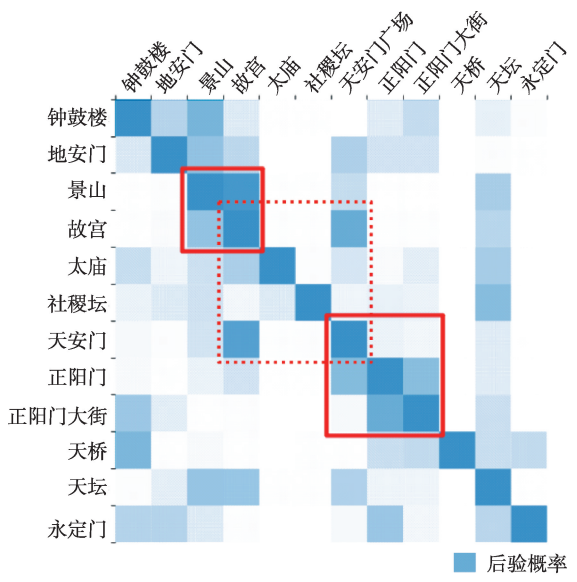

图5　北京城市中轴线各单元间后验概率热力图

Fig.5　Heat map of posterior probabilities among the cultural heritage sites along the central axis of Beijing

也紧密相连。

(2) 跨空间的认知汇聚模式。如红色虚线方框所示,在故宫—社稷坛—太庙—天安门这个空间上邻接的小区域,故宫、天安门相互之间的后验概率显著,而社稷坛、太庙的后验概率则非常低。由此在这一区域出现忽略社稷坛、太庙的故宫—天安门的关联模式,表明社会认知跨越地理空间的关联,是由于对这些单元的社会认知远远强于它们的地理邻接关系,说明认知转向的模式具有中心性。同时可以看出,故宫、天安门、天坛3个单元对应的纵轴整体颜色都较深,说明其后验概率普遍更高,认知容易转向中轴线上最核心的文化遗产单元。

4.3.2 中轴线内单元关联分析

通过文化遗产地名之间的互信息,描述北京旧城中轴线文化遗产要素之间的关联程度。以中轴线北段的地安门为例,图6显示与地安门互信息比较显著的若干中轴文化遗产地名。其中,后海、钟鼓楼、南锣鼓巷和地安门关联程度显著高于其他地名,而它们在地理空间上也距离较近,形成一个地域的"小圈"。

5 结论与展望

本文以北京城市中轴线文化遗产认知分析作为例,将大数据分析方法应用于对文化遗产的社会认知研究,构建了一个数据收集和预处理的大数据方法,建立了基于大数据的人群、时间、空间多维度分析框架,为文化遗产认识研究引入了新的方法。

本文研究框架充分利用海量舆情数据,展示较为客观可靠的分析结果。相比于问卷调查、文献调研等传统方法,跳出了传统方法小样本的局限,能快速地收集大量数据,展开人群维度、时间维度、空间维度的文化遗产认知分析,在分析效率和分析维度上都有较大的提升,有助于新知识、新模式的发现。

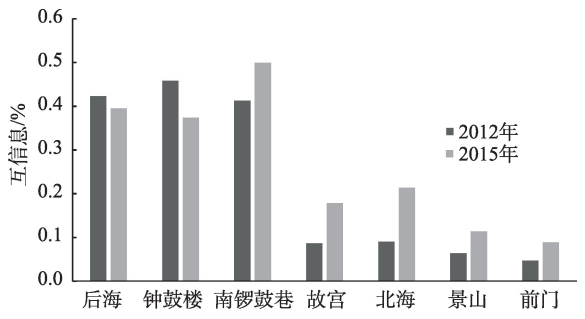

图6 北京城市中轴线其他地名与地安门之间的互信息

Fig.6 Mutual information between Di'anmen and other place names of the central axis of Beijing

本文主要结论为:①大数据的方法可以较好的从人群、时间、空间3个维度展开文化遗产认知分析:利用具有人群特征的数据源,可以有效分析群体差异对于文化遗产的认识异同;利用词频和时间比对的方法,可以有效地了解公众对于北京旧城中轴文化遗产单元的关注变迁;利用后延概率的方法,可以有效地了解文化遗产的空间内聚方式和认知转移方向。②不同群体对北京中轴部分文化遗产单元认知是相异的:官方偏向于行政管理,学者偏向于历史价值,大众偏向于生活化。如官方认知中的钟楼鼓楼、地安门偏向于是一个历史文化保护和旧城改造的行政区域;大众偏向于休闲娱乐场所,比如认知中太庙是婚纱摄影和展览场所;而学者则偏向于其历史事件和文物古迹。但各群体对北京中轴文化遗产最核心的单元故宫、天安门、天坛认知相对一致。③各群体对不同遗产单元关注程度和认知的分布会随着时间发生变化,如相对于2012年,2015年大众对于北京中轴文化遗产单元的关注更集中于故宫、天安门,同时对于太庙的历史价值认知有所提高。而相对于官方、学者,大众对文化遗产的认知变化较为显著,对热点事件的敏感度较高。④北京中轴文化遗产的公众关注度分布不均匀,其中故宫、天安门、天坛关注度最高。空间上相连的景山—故宫、天安门—正阳门—正阳门具有较高的认知双向关联。故宫、天安门、天坛,北京中轴的核心文化遗产单元具有较强跨空间的关联模式,认知易从中轴其他文化单元转向故宫、天安门、天坛核心文化遗产单元。

尽管大数据的文化遗产认知方法应用取得了较好的效果,但是影响文化遗产的认知是多元的,要完全了解各类人群对于文化遗产的认知,具有很大的挑战性,如大数据方法难以获取不使用网络的老年人的相关数据。在今后的研究中还有许多需要改进之处,例如,如何将GIS的空间分析方法、更精准的问卷调查方法融入于大数据文化遗产认知研究,将更多的社会经济、地理环境等大数据加以充分利用,优化文化遗产的大数据认知方法。

参考文献(References)

包书月, 张宝秀. 2011. 北京城中轴线发展历程及其对城市空间结构的影响[J]. 北京联合大学学报: 人文社会科学版, 9(3): 39-44. [Bao S Y, Zhang B X. 2011. Developmen-

tal history of Beijing city's axis and its influence on urban structure[J]. Journal of Beijing Union University: Humanities and Social Sciences, 9(3): 39-44.]

成亮. 2009. 浅析城市轴线在城市规划中的运用[J]. 现代城市研究, (1): 35-42. [Cheng L. 2009. Preliminary analysis of applying of urban axis to urban planning[J]. Modern Urban Research, (1): 35-42.]

计维斌, 谢珍君. 2013. 微博互动对网络购买态度的影响研究[J]. 甘肃社会科学, (3): 107-110. [Ji W B, Xie Z J. 2013. Weibo hudong dui wangluo goumai taidu de yingxiang yanjiu[J]. Gansu Social Sciences, (3): 107-110.]

李琛. 2015. 历史文化街区居民旅游影响的感知研究: 以北京什刹海地区为例[J]. 北京联合大学学报, 29(4): 36-44. [Li C. 2015. The influence of historical and cultural blocks residents tourism perception research: In Beijing the Shichahai lake area as an example[J]. Journal of Beijing Union University, 29(4): 36-44.]

李晨. 2011. "历史文化街区"相关概念的生成、解读与辨析[J]. 规划师, 27(4): 100-103. [Li C. 2011. Generating, reading and discrimination of "urban historic conservation areas" and relevant concepts[J]. Planners, 27(4): 100-103.]

李路珂. 2003. 北京城市中轴线的历史研究[J]. 城市规划, (4): 37-44, 51. [Li L K. 2003. A study on the history of the central axis in Beijing[J]. City Planning Review, (4): 37-44, 51.]

李乾夫. 2009. 论非物质文化遗产的认知、自觉与保护[J]. 大理学院学报, 8(11): 25-28. [Li Q F. 2009. Cognition, awareness and protection of the intangible culture heritage [J]. Journal of Dali University, 8(11): 25-28.]

梁思成. 1951. 我国伟大的建筑传统与遗产[J]. 文物参考资料, (2): 6-19. [Liang S C. 1951. Woguo weida de jianzhu chuantong yu yichan[J]. Journal of Heritage Reference, (2): 6-19.]

刘丽华, 何军. 2009. 国内民众的非物质文化遗产认知度实证研究: 以沈阳市民的辽宁省非物质文化遗产认知为例[J]. 旅游论坛, 2(4): 611-615. [Liu L H, He J. 2009. Empirical study on the recognition degree for Intangible cultural heritage of the masses: A case study on the recognition of Liaoning's intangible cultural heritage of Shenyang citizens[J]. Tourism Forum, 2(4): 611-615.]

刘祎绯. 2015. 我国文化遗产认知的时间扩展历程[J]. 建筑与文化, (6): 132-133. [Liu Y F. 2015. The expanding process of time in cultural heritage perceiving in China[J]. Architecture & Culture, (6): 132-133.]

陆建松. 2010. 中国文化遗产保护管理的政策思考[J]. 东南文化, (4): 22-29. [Lu J S. 2010. Thinking about the administration policies of Chinese cultural heritage conservation [J]. Southeast Culture, (4): 22-29.]

沈加锋. 1989. 从一个调查看北京中轴线的印象[J]. 城市规划, (4): 25-27. [Shen J F. 1989. Cong yige diaocha kan Beijing zhongzhouxian de yinxiang[J]. City Planning Review, (4): 25-27.]

王军. 2016. 探索城市历史文化价值认知的方法体系: 以历史性手工业名城浙江龙泉为例[J]. 城市发展研究, 23(2): 30-38. [Wang J. 2016. A exploration of method system on historical cultural value of these historical Cities: A case study of Longquan, which is a historical handicraft industry city in Zhejiang[J]. Urban Development Studies, 23(2): 30-38.]

王俊芳, 时俊卿. 2004. 问卷调查的类别、优缺点及实施[J]. 教育科学研究, (9): 58-59. [Wang J F, Shi J Q. 2004. Wenjuan diaocha de leibie, youquedian ji shishi[J]. Educational Science Research, (9): 58-59.]

杨振山, 龙瀛, Douay N. 2015. 大数据对人文—经济地理学研究的促进与局限[J]. 地理科学进展, 34(4): 410-417. [Yang Z S, Long Y, Douay N. 2015. Opportunities and limitations of big data applications to human and economic geography: The state of the art[J]. Progress in Geography, 34 (4): 410-417.]

Aizawa A. 2003. An information-theoretic perspective of tf-idf measures[J]. Information Processing & Management, 39 (1): 45-65.

Babbie E. 1998. The practice of social research[M]. Belmont, CA: Wadsworth Publishing Company.

Dallachiesa M, Ebaid A, Eldawy A, et al. 2013. NADEEF: A commodity data cleaning system[C]//Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York, NY: ACM: 541-552.

Fabret F, Jacobsen H A, Llirbat F, et al. 2001. Filtering algorithms and implementation for very fast publish/subscribe systems[J]. ACM Sigmod Record, 30(2): 115-126.

Förster T, Mainka A. 2015. Metropolises in the twittersphere: An informetric investigation of informational flows and networks[J]. ISPRS International Journal of Geo-Information, 4(4): 1894-1912.

Gueta T, Carmel Y. 2016. Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models[J]. Ecological Informatics, 34: 139-145.

Peng H C, Long F H, Ding C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8): 1226-1238.

Rao W X, Chen L, Hui P, et al. 2012. Move: A large scale keyword-based content filtering and dissemination system[C]// Proceedings of the 32nd International Conference on Distributed Computing Systems. Macau, China: IEEE: 445-454.

Sangpachatanaruk C, Znati T. 2004. Semantic Driven Hashing (SDH): An ontology-based search scheme for the Semantic Aware Network (SA Net)[C]//Proceedings of the Fourth International Conference on Peer-to-Peer Computing. Washington, DC: IEEE: 270-271.

Wu H C, Luk R W P, Wong K F, et al. 2008. Interpreting TF-IDF term weights as making relevance decisions[J]. ACM Transactions on Information Systems, 26(3): 131-37.

Yin J, Lampert A, Cameron M, et al. 2012. Using social media to enhance emergency situation awareness[J]. IEEE Intelligent Systems, 27(6): 52-59.

Zhang W, Yoshida T, Tang X. 2011. A comparative study of TF* IDF, LSI and multi-words for text classification[J]. Expert Systems with Applications, 38(3): 2758-2765.

Analyzing perception of cultural heritage sites based on big data: A case study of Beijing Central Axis

YANG Weishi[1,2,3], GUO Danhuai[4,5*], LU Yanling[6], WANG Deqiang[4,5], ZHU Yinqiu[4,5], ZHANG Baoxiu[6]

(1. Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-Sen University, Guangzhou 510275, China; 2. Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China; 3. Key Laboratory of Land Surface Pattern and Simulation, CAS, Beijing 100101, China; 4. Computer Network Information Center, CAS, Beijing 100190, China; 5. University of Chinese Academy of Sciences, Beijing 100049, China; 6. Beijing Union University, Beijing 100101, China)

Abstract: This article analyzes perceptions concerning cultural heritage sites along the central axis of Beijing from community, temporal, and spatial perspectives by extracting keywords, word frequency, term frequency-inverse document frequency (TF-IDF) weight, mutual information, posterior probability, and other features in microblogs, newspapers and magazines, and academic publications in 2012 and 2015. On the community dimension, through media information of characteristic groups, we found that different groups have different understanding of cultural heritage sites. The core sites of Beijing Central Axis cultural heritage, such as the Imperial Palace, Tiananmen, and Temple of Heaven are perceived relatively consistently by different communities. But the perceptions of the Bell and Drum Towers, Imperial Ancestral Temple, and Di'anmen are varied: officials are concerned with their administrative aspects, scholars are concerned with their historical values, and the public are concerned with their leisure and entertainment qualities. On the temporal dimension, changes of level of attention and perception on these cultural heritage sites are also observed. In 2015, the public paid more attention to the Forbidden City, Tiananmen, the Temple of Heaven, and the Imperial Ancestral Temple for their historical values as compared to 2012. Public perception, compared with that of officials and scholars, is more likely to change and more sensitive to significant events. On the spatial dimension, this research has examined the transfer of perception and correlation between cultural heritage sites. First, Tiananmen, Zhengyang Gate, and Zhengyang Avenue, which are connected in space, show higher two-way perceptions. Second, the posterior probability of the Imperial Palace, Tiananmen, and the Temple of Heaven is higher among the central axis cultural heritage sites, showing a cross space perception convergence model. Thus the analytical framework for perception of cultural heritage based on big data is an important supplement for traditional methods such as questionnaires, literature research, and interview analysis, as it increases the dimension and efficiency of analysis and aids to discover hidden knowledge and patterns. The conclusion of this study can provide important support for policy making in the rediscovery and protection of cultural heritage values.

Key words: big data analysis; data mining; perception of cultural heritage; term frequency-inverse document frequency(TF-IDF) weight; Beijing Central Axis