中国经济文库·应用经济学精品系列（二）

黎　嘤◎著

旅游大数据研究

Research on Tourism Big Data

中国经济出版社
CHINA ECONOMIC PUBLISHING HOUSE

中国经济文库 · 应用经济学精品系列（二）

黎　嶸◎著

旅游大数据研究

Research on Tourism Big Data

中国经济出版社
CHINA ECONOMIC PUBLISHING HOUSE
北京

图书在版编目(CIP)数据

旅游大数据研究 / 黎嶷著.
—北京:中国经济出版社,2018.7
ISBN 978-7-5136-3372-7
Ⅰ.①旅… Ⅱ.①黎… Ⅲ.①旅游业—数据处理—研究 Ⅳ.①F59-39
中国版本图书馆 CIP 数据核字(2018)第 151958 号

责任编辑　王　建
责任印制　巢新强
封面设计　华子设计

出版发行　中国经济出版社
印 刷 者　北京富泰印刷有限责任公司
经 销 者　各地新华书店
开　　本　710mm×1000mm　1/16
印　　张　17.25
字　　数　271 千字
版　　次　2018 年 7 月第 1 版
印　　次　2018 年 7 月第 1 次
定　　价　56.00 元
广告经营许可证　京西工商广字第 8179 号

中国经济出版社 **网址** www.economyph.com **社址** 北京市西城区百万庄北街 3 号 **邮编** 100037
本版图书如存在印装质量问题,请与本社发行中心联系调换(联系电话:010-68330607)

前　言

随着计算机技术的飞速发展，大数据和人工智能的时代已经到来。如何充分利用技术发展为人类带来的巨大便利，是旅游行业研究者和从业者均需审视的问题。

大数据目前已经在物理学、天文学、生物学、社会学等众多学科领域得到了广泛应用。旅游研究是研究旅游活动及其规律的领域，大数据技术为旅游研究的数据获取与旅游活动分析的数字化表达提供了坚实的技术基础。这些数据化表达能够实现对旅游活动长期、实时、动态化的记录，在此基础上可以开展更广泛的旅游研究。如何将大数据技术有效地应用于旅游研究及旅游业并促进其发展，是旅游研究目前面临的一个重要挑战。

我和团队自2014年开展对旅游大数据的相关研究，从各种可以获得的旅游大数据入手，比如互联网用户生成内容、游客手机信令数据、搜索引擎指数，开展对游客情感极性分析、游客轨迹分析、旅游统计、旅游客流预测、旅游大数据相关标准等方面的研究，力图在大数据背景下对现有旅游研究方法和研究对象进行补充，拓宽旅游研究范式、方法及范畴。同时，针对不同的应用场景开展研究成果的推广工作，以期发挥研究成果的实际应用价值。

本书的主要内容基于作者及其团队已经取得的旅游大数据研究成果，试图从多方面向读者展示旅游大数据研究的不同方法和研究实例。感谢研究团队成员长期的坚持和努力，迎难而上，不计得失，在新方向上不断探索。感谢我的硕士研究生谢宗彦、周纯洁；感谢郝志成高级工程师、朱伟博士、文玲高级经济师、桂婕副研究员、张斌儒博士。感谢对旅游大数据研究一直全力支持的黄先开教授和对研究团队进行耐心指

导的 Rob Law 教授和向征博士。感谢中国经济出版社的王建编辑，他对书稿认真负责的态度值得我和团队在今后的工作中学习。最后，感谢家人，没有家人的鼎力支持，我无法完成在 2018 年的春天交付此书的任务。

黎嘤

2018 年 5 月

目录 Contents

绪论 Introduction

第一节　旅游大数据研究的意义

一、大数据研究的意义

自20世纪90年代以来，电子传感器、通信、计算及存储技术的快速发展产生并迅速集聚了大量数据，大数据的概念应运而生。大数据中隐含了对商业、科学、政府及社会有价值的信息，逐渐引起学界、业界及政府的关注。

在学界，2008年，《自然》(*Nature*)推出了《大数据》专刊(*Big Data*)，探索大数据对当代科学的意义[①]。同年，计算机社区联盟(Computing Community Consortium)发表了非常有影响力的报告——《大数据计算：创造商业、科学与社会的革命性突破》(*Big Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science, and Society*)，阐述了在数据驱动的研究背景下，解决大数据问题所需的技术，以及面临的一些挑战[②]。2011年，《科学》(*Science*)推出专刊——《处理数据》(*Dealing with Data*)，围绕科学研究中大数据的问题开展讨论，阐述大数据对科学研究的重要性。

在业界，自2010年开始，易安信(EMC)、国际商用机器公司(IBM)、惠普、微软等全球知名公司开始对大数据相关厂商进行收购，实现大数据技术整合。2011年，EMC公司在主题为“云计算相遇大数据”的World 2011大会中阐述了云计算与大数据的理念与技术趋势。同年，麦肯锡公司在《大数据：创新、竞争和生产力的下一个前沿》(*Big Data: The Next Frontier for Innovation, Competition, and Productivity*)报告中称：“数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于大数据的挖掘和运

① 网址为：https://www.nature.com/news/2008/080903/full/455008a.html。

② 网址为：http://www.cra.org/ccc/docs/init/Big_Data.pdf。

用,预示着新一波生产力增长和消费盈余浪潮的到来。”①维克托·迈尔-舍恩伯格和肯尼斯·库克耶所著《大数据时代》指出:大数据具有促进人们生活、工作与思维变革的价值与意义[1]。徐子沛所著《大数据:正在到来的科技革命》指出:大数据将成为人们下一个观察人类自身社会行为的“显微镜”和监测大自然的“仪表盘”[2]。我国工程院院士李国杰对大数据的出现、发展到对经济与社会的影响逐渐加大这一现象给出了总结:大数据与人工智能是信息时代的一个新阶段。

我国政府与行业主管非常重视大数据的发展与应用。2015 年国务院印发《促进大数据发展行动纲要》,系统部署大数据发展工作。2016 年《中华人民共和国国民经济和社会发展第十三个五年规划纲要》提出“实施国家大数据战略”,强调把大数据作为基础性战略资源,加快政府数据开放共享、促进大数据产业健康发展。2017 年 1 月,工信部发布《大数据产业发展规划(2016—2020 年)》全面部署“十三五”时期大数据产业发展工作;5 月,水利部发布《关于推进水利大数据发展的指导意见》;6 月,最高检发布《检察大数据行动指南(2017—2020)》。截至 2017 年 9 月底,国务院、国家发改委、原环保部、原国土资源部、原国家林业局、交通运输部、原农业部、工信部、水利部、公安部等部委均发布了大数据战略文件;国家信息中心发布了《中国大数据发展报告》,指出了中国旅游大数据的发展现状及未来趋势;国家自然科学基金委员会发布了《大数据驱动的管理与决策研究重大研究计划 2017 年度项目指南》以指导大数据领域的重要研究计划(见表 0-1)。随着大数据产业政策继续加速出台,政府数据开放共享取得突破,大数据建设将不断深入,创新体系将不断完善。

表 0-1　国家与部分行业大数据政策

政策名称	发布日期	发文单位
关于运用大数据加强对市场主体服务和监管的若干意见	2015 年 7 月	国务院办公厅
促进大数据发展行动纲要	2015 年 8 月	国务院

① 网址为:http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation。

续表

政策名称	发布日期	发文单位
关于组织实施促进大数据发展重大工程的通知	2016 年 1 月	国家发改委
生态环境大数据建设总体方案	2016 年 3 月	原环保部
关于印发促进国土资源大数据应用发展实施意见	2016 年 7 月	原国土资源部
关于加快中国林业大数据发展的指导意见	2016 年 7 月	原国家林业局
关于推进交通运输行业数据资源开放共享的实施意见	2016 年 8 月	交通运输部
农业农村大数据试点方案	2016 年 10 月	原农业部
大数据产业发展规划(2016-2020 年)	2017 年 1 月	工信部
中国大数据发展报告(2017)	2017 年 2 月	国家信息中心
关于推进水利大数据发展的指导意见	2017 年 5 月	水利部
大数据驱动的管理与决策研究重大研究计划 2017 年度项目指南	2017 年 7 月	国家自然科学基金委员会
智慧城市时空大数据与云平台建设技术大纲(2017 版)	2017 年 9 月	原国家测绘地理信息局
关于深入开展“大数据 + 网上督察”工作的意见	2017 年 9 月	公安部

二、旅游大数据研究的意义

旅游是信息密集型产业,旅游活动涉及面极为广泛,旅游产业具有综合性、依托性、关联性强的性质和特征,旅游消费和经营服务、组织管理高度依赖信息资源。信息化对旅游活动的各个方面都产生深刻影响,旅游与信息化的融合是全方位、深层次和十分紧密的,信息技术必然是打造中国旅游升级版的主要技术支撑力量,旅游信息化是将旅游业培育成为国民经济战略性支柱产业和人民群众更加满意的现代服务业的主要科技途径。

目前,大数据技术的旅游应用是旅游信息化的一个重要里程碑(见图 0-1)。旅游业涵盖了“食、住、行、游、购、娱”等多个方面,影响、带动和促进的相关行业达 110 多个,涉及数据类型多样、要素繁多、数量庞杂。同时,旅游又是一系列游客持续移动的行为过程,随着信息技术的发展,这一系列过程的所有痕迹都能够以数据的形式被实时记录。同时,旅游业大量涌现出区别于传统线下服务的新型服务业态——在线旅游服务,无论是在旅游前、旅游中,还是旅游后,游客越来越多地选择通过在线旅游服务进行消费。游客的移动行为及其与旅游在线服务企业的交互活动实时产生了巨量数据,旅游大数据产生。

这些大数据的分析及利用，对于旅游公共管理与服务水平、旅游企业竞争力、游客满意度的提升，以及旅游产业升级改造具有重要意义。《“十三五”全国旅游信息化规划》的发布，为旅游大数据的发展指引了方向，将“推进旅游大数据运用，引领新驱动”作为“十三五”期间旅游信息化的主攻方向之一，提出了4项具体要求：一是运用大数据对游客数量、结构特征、兴趣爱好、消费习惯等信息进行收集，为旅游市场的细分、精准营销，以及旅游战略的制定提供依据；二是运用大数据对旅游消费信用等信息进行收集分析，增强对旅游市场主体服务和监管的意识；三是运用大数据对游客信息进行关联分析，进一步优化旅游公共服务资源配置；四是运用大数据对旅游景区信息进行关联分析，为景区流量控制及安全预警提供数据支持①。

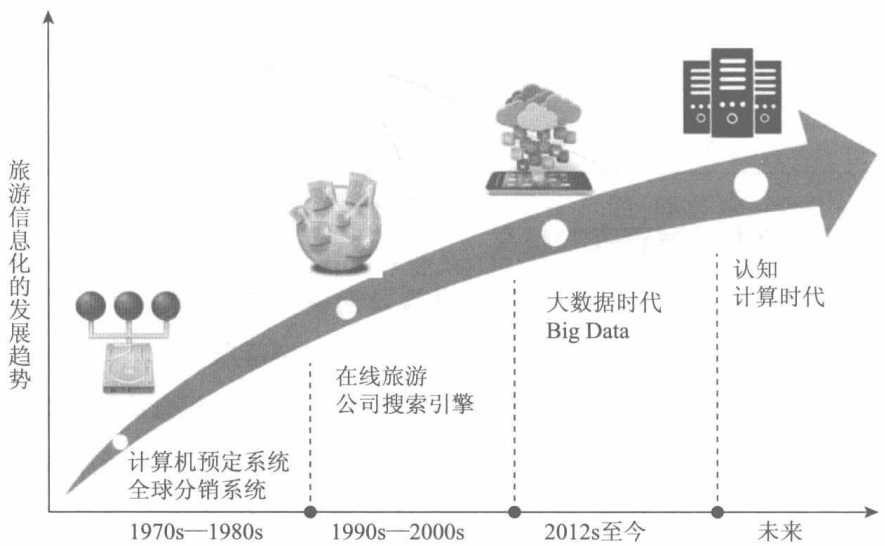

图0-1　旅游大数据：旅游信息化发展的里程碑

然而，目前我国旅游业界对大数据的应用尝试仍停留在对大数据概念与技术的浅层理解上，对旅游问题，支撑旅游业大数据应用的基础性、技术性，以及创新应用研究目前还处于空白阶段。事实上，旅游大数据具有量大、分散、多模态、跨媒体、跨行业等特征，传统的数据处理技术难以进行感

① 中华人民共和国文化与旅游部．国家旅游局办公室关于印发“十三五”全国旅游信息化规划的通知（旅办发〔2016〕346号）[EB/OL]．（2017-03-07）[2018-05-10]．http://www.cnta.gov.cn/zwgk/tzggnew/gztz/201703/t20170307_817856.shtml.

知、分析及应用。在旅游大数据的智能感知方面，与旅游相关的群智感知理论尚处于起步阶段，还不能为云计算环境下的层次化可扩展的旅游大数据平台系统的构建与应用提供支撑。

在旅游大数据的智能分析方面，现有的技术不能对跨媒体、多模态旅游大数据的自然属性、社会属性，以及交互行为等上下文之间存在的复杂关系进行建模提供有力支持，难以处理旅游大数据的格式及内容理解的主观性与多义性，不能从分布的、超大规模的、超高维度的、不完全的、有噪声的、模糊的、随机的旅游大数据中挖掘出可理解的知识模式。

在旅游大数据的应用方面，以基于旅游大数据的可信服务提供为研究对象，对基于旅游大数据的旅游服务质量评价方法、品牌与精准营销、面向游客的服务及其实现机制的研究尚属空白。目前，缺乏针对旅游问题域的大数据关键技术及其旅游应用的系统性研究。因此，尽管大数据技术还存在维护成本、处理原则、技术壁垒、回报周期、隐私侵犯，甚至包括人性伦理影响等方面的问题，甚至社会对大数据概念有意无意地过度消费，但旅游大数据无论对旅游领域还是大数据技术而言，都具有不容置疑的研究价值。

第二节　本书的内容结构

本书的研究目标为构建旅游大数据研究的体系结构，并就其中的重点方向和内容开展深入研究，为目前国内旅游大数据研究的体系化、规范化、工具化及方法构建等提供参考。

本书首先界定旅游大数据的概念，以此作为本书开展相关研究的概念基础。其次，就旅游文本大数据、游客移动大数据、游客搜索大数据、旅游大数据标准开展深入研究，内容包含研究框架、方法、案例。最后，对旅游大数据研究的主流工具加以介绍并讲解其使用方法。

第一章
Chapter 1

旅游大数据的概念

第一节　旅游大数据的定义与特征

大数据的提出并不是一个突发事件，而是伴随着人类社会的发展与技术进步自然产生的。大数据的产生与人类社会生活网络结构的复杂化、生产活动的数字化、科学研究的信息化相关，其意义和价值在于可帮助人们解释复杂的社会行为和结构，提高生产力，进而丰富人们发现自然规律的手段。随着人类社会进入21世纪，科技发展日新月异，信息技术迅速普及，特别是互联网和移动终端技术的发展，使得人与人之间的联系日益密切，社会结构日趋复杂，社会生产力与人们的生活水平得到极大提升，并且人的创造性活力在技术的大力发展与支撑下得以充分释放，与之相应的数据规模和处理系统也发生了巨大改变，从而使得大数据成为学界和业界的热点[3]。

尽管在银行、电信、零售、医疗等领域，大数据的研究与应用已经得到了人们的普遍关注与深入讨论；但在旅游领域，旅游大数据还缺乏科学且清晰的界定，大数据在旅游中的应用还停留在探讨阶段，存在大量简单将传统统计分析称为大数据分析的相互混淆的表述。旅游大数据的概念界定无论对于学界和业界，都是十分有必要的。

一、旅游大数据的定义

（一）大数据的定义

迄今为止，大数据还没有一个公认的定义。“大数据”一词作为术语被使用，最早可以追溯到20世纪90年代。1997年，迈克尔·考克斯与戴维·埃尔斯沃思在美国电子电器工程师协会（IEEE）第八届可视化会议中发布的

论文——《为外存模型可视化而应用控制程序请求页面调度》中使用了“大数据”:“可视化为计算机提出了一个有趣的挑战:通常情况下数据集相当大,耗尽了主存储器、本地磁盘,甚至远程磁盘的存储容量。我们将整个问题称为大数据。”同年,绍洛姆·韦斯和霓庭·因杜尔亚在《预测性文本挖掘基础》中指出:“收集的大量数据可在数据仓库中汇编,并使用强大的算法来全面研究数据。‘大数据’能帮助数据挖掘应用得出更为优质的结果。”1999年,史蒂夫·布莱森、大卫·肯怀特、迈克尔·考克斯和戴维·埃尔斯沃思在《美国计算机协会通讯》上发表的《千兆字节数据集的实时性可视化探索》一文中,提出了“大数据的科学可视化”。可以看出,这一时期的“大数据”指数据规模大。

然而,大数据如果仅仅是指数据规模大,则使用海量数据、超大规模数据这些术语也能够表达相同的含义,“大数据”这一术语的提出就失去了科学意义。即“数据规模大”不足以清晰描述大数据。随着大数据不断涌现,大数据应用需求的增长,大数据技术的进一步发展,大数据的定义得到了更为清晰的表述。一般认为,大数据是指利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集。例如,中国科学院院士李国杰认为:大数据是指无法在可容忍的时间内用传统信息技术和硬件工具对其进行感知、获取、管理、处理和服务的数据集合[4];麦肯锡公司认为:大数据是指其大小超出了典型数据库软件的采集、存储、管理和分析等能力的数据集(《大数据:下一个具有创新、竞争和生产力的前沿领域》);美国国家标准技术研究所(NIST)在《大数据:定义和分类》中提出大数据是指那些用传统数据架构无法有效处理的新数据集;Gartner 公司认为:大数据超出了常用硬件环境和软件工具在可接受的时间内为其用户收集、管理和处理数据的能力。

在大数据的定义得到广泛引用和认可的同时,人们往往会进一步提出问题:到底什么情况是“常用软件耗时超过可容忍时间”呢?要理解这一点,需要区分作为常用软件工具的传统数据库和大数据的区别。孟小峰和慈祥(2013)从数据规模、数据类型、模式和数据的关系、处理对象,以及处理工具5个方面区分了传统数据库和大数据的本质差别(见表1-1)[5]。据此可以看出,大数据超出了传统数据库的处理能力,在数据来源、数据处理方式和数据思维等方面产生了革命性的变化。

表 1-1　传统数据库与大数据的区别[5]

	传统数据库	大数据
数据规模	通常以 MB 为基本单位	通常以 GB，甚至 TB、PB 为单位
数据类型	类型单一，通常仅有一种或少数几种，以结构化数据为主	类型繁多，数以千计，数据包含结构化、半结构化及非结构化数据，其半结构化和非结构化数据所占比例越来越大
模式和数据的关系	先有模式，后有数据	无预定模式，模式只有在数据出现后才能确定，模式随着数据量的增长处于不断地演变之中
处理对象	数据作为处理对象	数据作为一种资源来辅助解决其他诸多领域的问题
处理工具	一种或少数几种工具就能处理所有数据	不存在一种工具能够处理所有数据的情况

在非技术领域，对大数据还存在着不同视角的理解。《大数据时代：生活、工作与思维的大变革》所提出的大数据概念在人文社会学领域，以及一些与其交叉的自然科学领域如人文地理学等具有较大影响。该书指出：大数据是指不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法；即大数据是“全数据”，“样本＝总体”，大数据中的“大”是相对“所有数据”的大而不是绝对意义上的大[6]。由于“全数据”并不一定在技术上需要采用区别于传统数据库技术的大数据方法，这种“全数据”的大数据理解显然与技术上对大数据的定义并不完全吻合。当“全数据”规模之大或者之复杂，难以用传统数据处理工具在可以容忍的时间内进行处理时，两种定义完全吻合；而当“全数据”规模小，根本无须以“GB”“TB”来计，且传统数据处理工具完全能够处理时，或者所处理的数据难以用传统软件处理，但却不是“总体”或者距离“总体”有较大距离时，两种定义存在相互矛盾，一种定义难以覆盖另一种定义。

对于矛盾的第一种情况，即当“全数据”规模小，根本无须以“GB”“TB”来计，且传统数据处理工具完全能够处理时，可以借用信息技术领域中对“大数据中的小数据”来进行界定。2013 年，美国康奈尔大学 Estrin 教授在“新神经信息处理系统国际会议（NIPS 2013）”上指出，从用户上网和使用各种移动设备过程中所产生的大量用户行为轨迹数据中提取出的个体数据，

为揭示人类行为模式的规律提供了可能，这些个体数据被称为大数据时代的小数据[7,8]。Estrin 教授所提出的大数据中的小数据，可以是个体的“全数据”，无论是否需要大数据技术去处理。当个体数量开始积聚达到一定程度时，需要以“TB”或更大数量级来计时，这时就需要大数据技术来处理[8]。

第二种矛盾情况是目前大数据在人文社会科学领域饱受诟病的主要原因之一。目前，由于互联网数据的丰富和可获得性，人文社会科学领域涌现出大量基于互联网用户行为数据而开展的相关研究，而互联网用户被认为不是全部人口，无法代表全样本[9-11]。分析起来，造成这种情况的原因实际上并不在于互联网数据本身，无论从技术角度，还是从人文社会学角度，对网络用户行为及基于互联网生成的大量数据开展研究都是非常有意义的，但研究者需要对所采用的数据有深入的理解，采用科学的研究方法，才能产生有说服力的研究结论。目前，随着大数据的热潮逐渐退去，冷静下来的学者开始研究如何弥补互联网数据的缺陷，探索大数据方法与传统以问卷调查为主要代表的抽样调查方法相互补充、相互验证及佐证的研究方法体系[9]。

（二）旅游大数据的定义

旅游研究根植于旅游现象自身的规律，同时汲取其他相关学科丰富的营养。基于上述大数据的定义与理解，以及社会与技术发展的阶段性特征，旅游大数据可以定义为：旅游大数据既指旅游领域中那些“样本＝总体”的全数据集，又指那些利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集。

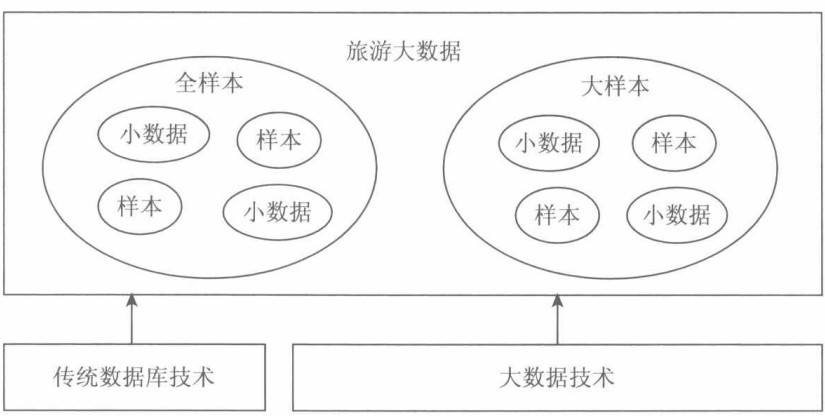

图 1-1 旅游大数据的定义

在该旅游大数据定义下，旅游大数据的范畴既包含旅游行业企业、部门、单位运行直接产生的数据（包括各类景区、酒店、旅行社、旅游主管部门、行业协会等信息系统产生的数据），旅游相关行业和领域的数据（交通、气象、环境、人口、规划等涉旅数据），游客行为数据（GPS 轨迹、移动通信手机信令、互联网浏览、点击、查询等行为数据），以及来自公共与社交媒体的旅游舆情数据（微博、微信、论坛、广播电台等提供的文字、图片、音视频等数据）等。

二、旅游大数据的特征

（一）大数据的特征

目前，学界和业界对大数据特征的界定没有形成统一的认识。许多研究者与企业都提出了不同的观点，从而构成了目前对大数据特征的不同认识，其中有代表性的是大数据的 5V 特征（见表 1-2）。

表 1-2　大数据的特征

特征	提出者	含义
规模性（Volume）	道格·莱尼《3D 数据管理：控制数据数量、速度及种类》	数据量大
多样性（Variety）		数据类型多样
高速性（Velocity）		数据处理速度快
真实性（Veracity）	IBM 公司	数据反映客观事实
价值性（Value）	IDC 公司	价值密度低但总价值高

（二）旅游大数据的一般特征

旅游大数据具有明显的大数据 5V 特征，这 5 个特征能够用来区分旅游大数据与其他旅游数据的不同。

1. 规模性（Volume）

旅游大数据具有巨大的数据量。据世界旅游理事会（WTTC）测算，早在 20 世纪 90 年代初，旅游业就已经超过汽车和石油等传统产业，成为世界经济中的第一大产业。现代旅游业综合性强、关联度高、产业链长，已经明显突破了传统旅游业的范围，广泛涉及并交叉渗透到 29 个相关经济部门，直接