遥感学报
NATIONAL REMOTE SENSING BULLETIN
EI CSCD
Scopus CNKI

地理大数据聚合的内涵、分类与框架

Big geodata aggregation: Connotation, classification, and framework

裴韬;黄强;王席;陈晓;刘亚溪;宋辞;陈洁;周成虎

引用本文:
裴韬，黄强，王席，陈晓，刘亚溪，宋辞，陈洁，周成虎.2021.地理大数据聚合的内涵、分类与框架.遥感学报，25（11）： 2153-2162 DOI: 10.11834/jrs.20210480.
Pei T, Huang Q, Wang X, Chen X, Liu Y X, Song C, Chen J and Zhou C H. 2021. Big geodata aggregation: Connotation, classification, and framework. National Remote Sensing Bulletin, 25（11）: 2153-2162 DOI: 10.11834/jrs.20210480.

在线阅读 View online: http://www.ygxb.ac.cn/thesisDetails?columnId=19404547

您可能感兴趣的其他文章
Articles you may be interested in

基于激光雷达的自然资源三维动态监测现状与展望
Current status and prospect of three-dimensional dynamic monitoring of natural resources based on LiDAR
遥感学报.2021,25(1):381-402.

面向中国深空探测任务的行星数据系统的设计与实现
Design and implementation of a planetary data system for Chinese deep space exploration
遥感学报.2021,25(2):599-613.

数据驱动的蒸散发遥感反演方法及产品研究进展
Progress of data-driven remotely sensed retrieval methods and products on land surface evapotranspiration
遥感学报.2021,25(8):1517-1537.

高分五号高光谱数据融合方法比较
Comparison of fusion methods on GF-5 hyperspectral data
遥感学报.2022,26(4):632-645.

融合Landsat 8与Sentinel-2数据的红树林物候信息提取与分类
Fusion of Landsat 8 and Sentinel-2 data for mangrove phenology information extraction and classification
遥感学报.2022,26(6):1121-1142.

地理大数据聚合的内涵、分类与框架

裴韬[1,2,3]，黄强[1,2]，王席[1,2]，陈晓[1,2]，刘亚溪[1,2]，宋辞[1,2]，陈洁[1,2]，周成虎[1,2]

1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室, 北京 100101;
2. 中国科学院大学, 北京 100491;
3. 江苏省地理信息资源开发与利用协同创新中心, 南京 210023

摘 要：地理大数据是地理对象所产生的“足迹数据”，而地理大数据挖掘就是通过反演分析地理对象的“足迹数据”，揭示其中蕴含的人地关系及其时空模式。近年关于地理大数据的重大研究进展显示，其研究结论的取得大多需要借助多种大数据的信息综合。为此，本文提出研究地理大数据聚合的必要性：代表新的研究范式、产生新的研究视角、提升研究的全面性，并将地理大数据聚合定义为：不同地理大数据之间通过转换形成面向研究对象的多维数据集合。在阐述地理大数据分类和特点的基础上，将地理大数据聚合分为：时空聚合、面向对象聚合、面向主题聚合以及面向模型聚合四种类型。地理大数据聚合作为地理大数据挖掘的重要组成部分，其过程大致可以分为：确定内核—信息溯源—反演汇聚3个步骤，而其中的关键科学问题包括：统一时空框架和基准、统一数据表达与存储、匹配同一观测对象的多数据源、统一多源数据的时空范围、调和不同对象之间不同的变化速率、解决数据共享与隐私保护、大数据与小数据的聚合等。地理大数据聚合方法与技术的发展，将会拓展地方的感知、土地功能的识别、对人群的观测与理解、功能与流动之间关系、地表复杂巨系统等新领域的研究。

关键词：地理大数据思维，城市功能区，人群活动，地理信息科学，数据融合

引用格式：裴韬,黄强,王席,陈晓,刘亚溪,宋辞,陈洁,周成虎.2021.地理大数据聚合的内涵、分类与框架.遥感学报,25(11):2153-2162

Pei T, Huang Q, Wang X, Chen X, Liu Y X, Song C, Chen J and Zhou C H. 2021. Big geodata aggregation: Connotation, classification, and framework. National Remote Sensing Bulletin, 25(11):2153-2162[DOI:10.11834/jrs.20210480]

1 引 言

大数据时代的到来表面上看是以数据的类型和数量的爆发为主要标志，但实际上，大数据的本质特征是其产生和记录的方式。与传统目的性采样的“小”数据相比，大数据的产生大多并非带有明确的目的性，通常是对象在运动和变化过程中“不自觉”留下的“足迹”信息。在地理学的研究中，被广泛应用于城市功能和人群活动感知的手机数据，就是用户在使用手机时“不自觉”产生的数据，而非直接面向城市研究而获取的。近年来，与地理大数据相关的研究产出了若干标志性的成果，如综合多源地学数据对全球树的数目进行估计（Crowther等，2015），利用DEM、NDVI，以及考古和历史文献记载的牧民营地等数据对古代丝绸之路的重建（Frachetti等，2017），综合多源遥感、水文观测以及河流测量等数据对全球河流面积的估算（Allen和Pavelsky，2018），融合搜索引擎关键词与CDC报告门诊量数据对流感的预测（Ginsberg等，2009），使用已发表的气候与冲突暴力关系的60个案例以及全球气温数据量化气候对人类冲突的影响（Hsiang等，2013），透过手机数据感知城市土地功能（Pei等，2014；Jia等，2018；Cai等，2017；Niu等，2017）等。上述研究中看似与结果“风马牛不相及”的大数据最终能够产生颠覆性的认知，原因就在于以上

收稿日期：2020-10-30；**预印本**：2021-08-10

基金项目：国家自然科学基金(编号:41525004,42071436)

第一作者简介：裴韬,1972年生,男,研究员,研究方向为地理大数据挖掘。E-mail：peit@lreis.ac.cn

研究充分利用了大数据思维，即从丰富的“足迹”大数据中通过关联、综合与反演的思路揭示了研究对象的分布、变化及机理。

在上述成果中，从数据的产生到规律的发现可以大致归纳为从“正演”到“反演”的过程。具体讲，地理对象“不自觉”产生的各种信息被不同传感器记录下来形成大数据的过程可以视为“正演”，而反演则是基于万物互联的思想通过大数据反推地理对象的特征、机理与演化的过程。以基于多源大数据进行城市功能区划分的研究为例（Jia等，2018；Cai等，2017；Niu等，2017），所谓正演，即一方面，城市的地表要素及其格局变化可通过遥感数据记录下来；另一方面，城市人群的活动将会产生手机信令、签到评论和出租车轨迹等大数据。而应用多源地理大数据推断城市土地功能区则是“反演”的过程，即将遥感数据以及人群活动信息（如手机信令、签到和出租车轨迹等大数据）汇聚在一起，首先通过遥感数据反演城市功能分区，再利用“不同功能区”将产生不同的“行为模式”这一关系，通过社交与行为大数据反演出不同分区的城市土地功能，最终实现城市分区空间范围及其功能的确定（Pei等，2014；Jia等，2018；Cai等，2017；Niu等，2017）。

需要指出的是，在以上地理大数据反演的过程中，城市土地功能综合了“地”和“人”的信息，因此需要将遥感数据与行为数据汇聚到研究对象上。不仅如此，上文提到的若干研究均使用了多种数据，即通过多源地理大数据的综合实现研究目的，这一过程称为聚合。具体地，地理大数据聚合可以定义为：多源地理大数据通过转换形成聚焦研究对象的多元数据集合的过程。经过聚合的地理大数据，都是地理对象直接或间接正演得到的，并可通过反演建立与研究目标之间的关系。

地理大数据聚合作为地理大数据研究的前提和基础，尤其是在面对较为复杂的地理学问题时，需要综合来自不同侧面的多源信息，才能实现研究目标。地理大数据聚合的研究意义可概括为以下3个方面。（1）地理大数据的聚合是数据驱动研究范式的重要表现形式，即不同来源的信息发生综合与碰撞会迸发出新的科学问题，拓展地理信息科学的研究的范围。例如，街景数据与道路数据的碰撞，就可以产生街道绿视率研究（Li等，2015）；（2）大数据聚合会创造出新的研究视角，有可能产生原创性的结论，例如：手机数据价值的发现成为感知土地功能的新视角（Soto和Frías-Martínez，2011；Pei等，2014）；（3）地理大数据的聚合会使得问题的研究更为全面和系统，例如，遥感数据与社交行为数据的综合使用，可以使得对土地功能的识别更加准确（Jia等，2018）。地理大数据聚合这种跨格式、含义、尺度的碰撞成为有别于传统研究的重要特征，已经成为为孕育新问题和新方法的重要动力源。

地理大数据聚合之于地理大数据挖掘虽然重要，也时常作为一个术语出现在文献中，但由于地理大数据的产生不是根据目的“设计”出来的，因此，其种类不仅繁多，而且形式上也缺乏标准，故地理大数据的聚合需要克服内涵、形式上的不统一，目前仍然是地理大数据研究中较为薄弱的环节。

“聚合”（Aggregation）一词最初源于化学领域，是指高分子化学中单体小分子通过相互连接成为新的高分子化合物的过程（杜晖，2013）；在信息领域，是指具有多来源的信息单元的融合和重组，其实质是多源信息单元的合并。目前针对大数据聚合的进展大致可分为两个方面：总体方法论研究以及针对不同数据类型的具体融合方法研究。针对大数据聚合理论，孟小峰和杜治娟（2016）提出了大数据融合的过程以及若干共性关键技术；秦文斐（2018）提出采用张量理论进行大数据的融合；姚富光等（2018）提出粒计算的大数据融合方法。针对地学大数据的特殊性，陈国强（2014）提出基于相关性的大数据融合思路；朱月琴等（2015）针对地质大数据的融合策略，提出采用HDFS技术实现多源异构地质文件的存储与查询；申占恒（2020）提出采用深度学习进行城市大数据融合的思路。除此之外，还有不少关于专业领域大数据聚合的成果，如暴雨（吴先华等，2017）、旅游（王嘉茵 等，2020）、公共交通（李传明，2020）、公共安全（颜培超，2019）、物联网（陈慧，2020）、地缘关系（李萌 等，2021）、疫情大数据（裴韬 等，2021）等数据。综合而言，已有研究针对大数据的聚合从技术原理和应用领域两个不同维度进行了探讨，但对其中一些基础性问题的认识仍然有待于深入。虽然遥

感数据的融合已经成为遥感科学重要的分支（翁永玲和田庆久，2003；张继贤，2011；张立福 等，2019；张良培和沈焕锋，2016；李树涛，2021），但其研究仅局限于遥感数据内部时—空—谱信息之间的融合，其他类型的地理大数据涉及不多，因而可以视为地理大数据聚合的特例。总体上，地理大数据聚合研究仍处于初级阶段，主要与地理大数据的多样性、对地理大数据认知的歧义、地理大数据挖掘思维的特殊性等因素有关。

鉴于此，本文针对地理大数据所具有的多粒度、多模态、多尺度等特征，从地理大数据的特点出发，分别就地理大数据聚合的本质、聚合的种类、聚合的关键问题以及对地理学若干领域的推动作用等几个方面进行探讨。

2 地理大数据的分类及特点

前文已阐述了地理大数据的内涵，其外延按照所观测的对象的不同又可分为对地观测大数据以及社交与行为大数据。对地观测大数据以遥感数据及对地观测台网数据为主，主要记录地表要素的分布及其变化；而社交与行为大数据种类较多，包括：手机信令及通话、出租车轨迹、社交媒体、公交刷卡等数据，主要记录人的行为模式。两类数据关注的对象不同，而产生方式、表达模型、时空粒度等也不一样。地理大数据的快速增长，尤其是社交和行为大数据的爆发，以及由此导致的对地观测大数据与社交行为大数据的聚合，为地理学中人—地关系的研究提供了前所未有的机遇，其碰撞将会产生观察人—地关系的新视角（裴韬 等，2019b）。

地理大数据除了根据关注对象的不同划分为两种类型之外，还可以按照获取方式、表达形式、单元粒度、来源等进行区分。按照获取方式可以分为：传感器观测、台站记录、“志愿者”上传的大数据等；按照表达形式可以分为：数据、文本、影像、视频等大数据；按照空间承载的类型可以分为栅格（格网）和矢量大数据；按照来源可以分为科学观测、社交媒体、互联网、统计部门、运营商大数据等。

地理大数据可以视为近似全量样本的数据，与传统的目的性采样数据相比，具有密度高、粒度细、广度宽、偏度重、精度差等五度特征（裴韬 等，2019b）。对于科学研究，地理大数据在表达、结构、来源等方面的多样性及其5“度”特征是一把双刃剑。在地理大数据的聚合过程中，一方面，地理大数据的多源性，以及在广度、密度和粒度等方面的优势使之成为地理学研究新的信息源；另一方面，数据之间巨大的差异，以及在偏度和精度方面的劣势，也使之成为研究中不确定性的重要来源。因此，如何利用好地理大数据就成为数据聚合的重要目标。

3 地理大数据聚合的本质与类型

引言中已讨论了地理大数据产生的机理，由于地理大数据的产生不是根据目的采样得到的，因此，地理大数据的聚合需要面对数据本身的诸多的天然属性，相比目的性采样的“小数据”融合，影响因素更多，并具有更多的类型。为此，本节将进一步阐述地理大数据聚合的本质，并试图对其进行分类。地理大数据聚合的本质以及类型划分将分别是对其中关键问题的归纳以及发展方向预测的依据。

3.1 地理大数据聚合的本质

地理大数据聚合的动力是其中蕴藏的万物关联关系，并集中体现了地理学研究的综合性及人—地关系耦合的特征。因此，地理大数据聚合的本质就是不同类型大数据的信息互补、增强与催生。为了更加明晰地理大数据聚合的本质，将其与影像数据融合进行对比。Pohl 和 van Genderen（1998）认为，影像融合就是利用某种算法，将两幅或多幅影像组合成一幅新影像的技术。Alparone 等（2015）认为遥感数据融合的目标是针对某种调查现象，协同组合两个或更多影像数据，以获取比单一影像更多的知识。张良培和沈焕锋（2016）将遥感数据融合定义为对同一场景并具有互补信息的多幅遥感数据或其他观测数据，通过对其综合处理以获取更高质量数据、更优化特征、更可靠知识。由此可见，遥感数据融合面对的是同一观测对象，综合不同角度的多源遥感数据，结果是产生具有统一格式与内涵的新数据集。遥感数据融合的目的是提供更准确、精确、详细的数据，并有可能产生新的信息或知识。遥感数据的融合作为地理大数据聚合的特殊形式，因其融合之前在数据来源、表达方式以及观测对象之间都存在一定相似性，可以看成是数据之间的“近

亲繁殖”。与之不同，地理大数据聚合可视为不同来源信息面向对象的重组和整合的过程，因此在数据来源、聚合机制、聚合目的以及采用方法等方面均有所不同，可以看成是数据之间的“转基因”操作。二者的差别主要有以下4点。

（1）二者参与的数据源不同。遥感数据融合的对象主要是对地观测数据，且以遥感影像数据为主；而地理大数据的聚合则可能包含两类大数据，对地观测数据之间、社交与行为大数据之间，以及对地观测大数据与社交行为大数据之间都可以进行聚合。

（2）聚合的机制不同。遥感数据融合的机制依赖于不同遥感数据的成像原理；而地理大数据聚合的基础则是各种数据与研究对象的时空关系，而这些时空关系包括：相同对象从不同角度观测得到的信息（如：同一城市的遥感数据与街景数据；同一地区的台站记录气温数据与红外遥感反演的温度产品）、同一主题下不同对象的信息（如：自然灾害的受灾信息以及与此相关的微博记录）、记录同一对象的不同部分的信息（如同一用户轨迹的室外部分与室内部分）。

（3）聚合目的不同。遥感数据融合的目标以提升遥感数据的质量和信息量为主，而地理大数据聚合的目的则是以发现新的关系与知识为主。地理大数据的聚合就是通过“反演”的思路找到与研究对象相关的不同数据进行汇聚，产生相关数据集，并藉此通过一系列挖掘方法产生新的知识。

（4）所采用的方法不同。遥感数据融合采用的方法主要基于时—空—谱原理的机理模型和统计模型方法，而地理大数据聚合所采用的方法则更为广泛，主要为基于地理大数据之间各种时空关系的处理方法，包括时空拓扑、空间聚类，时空关联、时空回归等分析方法。

由此可见，地理大数据聚合无论是从数据来源、聚合的机理、聚合的目的以及所采用的方法，都较遥感数据融合方法有了实质性的扩展。

3.2 地理大数据聚合的种类

地理大数据聚合的最终目的是将不同类型的数据通过时空关系映射到同一个研究对象上，根据映射手段的不同可以将地理大数据聚合分为时空聚合、面向对象聚合、面向主题聚合以及面向模型聚合4种类型。

（1）时空聚合。时空聚合的思路是将满足某种时空关系的大数据进行聚合。时空聚合中较为常见的方式就是基于位置的聚合。位置聚合是以位置为基准，将不同地理大数据聚合在一起。例如，对于规划领域的“一张图”工程，实际上就是利用位置作为线索，将不同的数据聚合在一起，实现各种不同目的规划的兼容和优化（张恒 等，2019）。除了位置聚合之外，时空聚合还包括时空拓扑（如：通过POI点与遥感分类信息的组合判断城市功能）（Song等，2018）、时空相关（搜索引擎关键词的空间频率与手足口病例空间分布之间的聚合（Huang和Wang，2018）、时空关联（自然灾害与微博信息的聚合（Wang等，2019）以及街景图像与出租车OD点分布的聚合（Zhang等，2019））等关系的聚合。

时空聚合所需的关键技术包括：栅格计算、尺度转换、位置关联、拓扑关联。栅格计算是指将相关数据转化为栅格图层，进行对应的时空叠加。尺度转换是将数据中不同的时空承载单元进行统一（Chen等，2019）。位置关联是以位置为支撑，将不同的对象通过相同的位置联系在一起。拓扑关联则以拓扑关系为纽带，将不同的对象通过拓扑关系结合在一起。

（2）面向对象聚合。面向对象聚合是指将同一个对象在不同时空范围内所产生的大数据聚合在一起的方式。例如，一个人在室外的活动轨迹可以通过其手机产生的信令数据得到，而在室内活动的轨迹可通过其在室内使用WiFi上网产生的数据得到，两类轨迹数据由于属于共同的对象，故可以可通过手机的MAC标识进行聚合。然而，在实际应用中，室、内外信息还难以真正的打通，其原因在于数据共享的鸿沟以及隐私保护需要。除了上述这种时空互补型的聚合，对象在时空上的重叠也常常是聚合的任务，如一辆出租车的GPS轨迹与其手机信令轨迹之间的匹配就属于此情况。这种数据匹配的解决方案有两种，其一是通过关联手机用户与司机身份进行数据聚合；其二是通过两类轨迹的时空相似性进行数据聚合。前者需要克服数据共享的障碍，而后者需要构建数据匹配的方法。

面向对象的聚合依赖于时空匹配技术。时空匹配技术又可分为两种。其一为数据库中字段的

匹配，即不同的对象具有唯一的标识，通过字段关联将同一对象进行聚合。上述的出租车轨迹与手机信令轨迹的聚合中，出租车司机的身份即为二者的共同的唯一标识。其二为空间匹配，即通过两类数据中出现的共同位置信息将不同的对象进行匹配。例如，出租车轨迹与手机信令轨迹之间的匹配就可以通过寻找二者共同出现的位置及其数量进行匹配。

（3）面向主题聚合。该类方法的思路是通过同一个主题将不同类型的大数据聚合在一起。这种聚合的典型例子就是互联网中按照某一主题汇聚各种信息的主题搜索。裴韬等（2019a）应用文本挖掘和知识图谱构建了互联网事件信息聚合框架，通过某种主题（事件）可将互联网中不同来源的信息聚合在一起。例如，关于某次地震事件的信息，就可以将互联网中关于该地震的震前异常、震害发生、震后救援、后期安置等信息通过事件的本体框架聚合起来，形成研究整个地震事件从自然灾害到社会影响全过程的数据专题。

面向主题的聚合涉及的数据多为文本数据以及带有标签数据等，主题关键词与对象标签是联系不同数据的枢纽，因而这种类型的聚合所依赖关键技术是文本匹配。在上文例子中，某地震名就是面向主题聚合的关键词，聚合的过程就需要通过文本匹配将包含某地震名的数据搜索出来，进而根据语义判断它们之间是否具有关联性。

（4）面向模型聚合。面向模型聚合是指将同一个地理过程中所涉及到的数据进行聚合，而聚合的数据均属于同一个地理过程，并可通过专业模型关联起来。这种聚合的典型例子就是数据同化。在数据同化中，遥感数据可以作为陆面或水文模型的输入参数或边界条件，与其他观测数据最终形成某种数据产品（李新和黄春林，2004）。在地理大数据研究中，该类聚合就表现为某类数据中的一些特征恰为某模型的输入，例如，在针对城市风貌的研究中，可以利用街景图像中的某些特征，通过训练建立判别模型，从而对城市建筑风格进行识别（Zhang等，2020），这其中，街景图像与地图数据的聚合就产生了城市风貌的空间分布。

面向模型聚合的核心纽带就是共同的模型。由于模型大多具有多重不同类型的输入数据。这其中的关键技术则与具体的专业知识有关。例如，模拟水文过程需要水文过程的知识，识别复杂地物的深度学习模型则需要建立地物与输入信息之间的相关关系。

4 地理大数据聚合的框架与其中的科学问题

地理大数据的聚合作为地理大数据挖掘的前提和基础，是有据可循的。同时，地理大数据聚合也存在诸多理论难点，是目前地理信息科学亟需解决的理论问题。

4.1 地理大数据聚合的框架

地理大数据挖掘的目标是识别地理对象之间的异同及关系，并进一步发现新的模式或新的知识（裴韬 等，2019b）。总体上，地理大数据挖掘的整个过程可分为：问题确定、数据获取、数据聚合、数据分析、结果验证、机理解释等步骤。地理大数据聚合作为地理大数据挖掘的一个阶段，大致可以分为3个步骤：确定内核—信息溯源—反演汇聚。确定内核的含义即为确定所要聚合的地理对象；信息溯源是根据内核回溯，找到其所产生的数据；而反演汇聚则是利用不同的技术方法将不同类型的地理大数据以内核为中心汇聚成数据集。以应用手机数据进行城市功能识别为例，其中的内核为城市功能分区，而信息溯源是指从城市功能出发，找出哪些信息可以用来进行功能区的判断，如手机信令数据和遥感数据等，而反演汇聚则是指如何将手机信令数据与遥感数据汇聚在统一的框架下。

4.2 地理大数据聚合的科学问题

由于地理大数据的多源、多粒度和多模态特性，其聚合需要解决更多的难题，可归纳为如下的7个科学问题。

（1）如何构建统一的时空框架和基准。由于地理大数据的来源不同，可能导致其位置信息中的时空基准存在差异，因此必须建立统一的时空框架和基准才能保证不同类型大数据之间的分析与计算。统一的时空框架和基准是实现地理大数据时空聚合的基础，为此，需要实现的内容包括：统一的大地基准面、统一的坐标系、统一的投影方式等。

（2）如何解决聚合中数据表达与存储的差异。大数据的表达方式的不同最终导致其数据结构不

同，具体地，遥感大数据多为半结构化的栅格形式；而社交与行为大数据主要表现为矢量格式的结构化数据，如签到数据、出租车轨迹以及手机信令数据等；而互联网中的网页多为文本数据。如此多类别的大数据实现聚合，必须通过合适的转换形成统一的数据格式。其中，文本数据必须经过结构化处理之后才能进行计算，而文本数据结构化需要借助知识图谱的理论（裴韬 等，2019a）；矢量数据之间、栅格数据之间以及矢量与栅格数据之间的转换，其实质是不同数据单元之间的转换，或可视为数据承载单元之间的转换。例如，手机数据（包括通话数据和信令数据）是以基站产生的泰森多边形为承载，在与栅格数据（如遥感数据）、多边形数据（如交通小区）进行聚合时，由于承载单元的不一致往往会造成误差。即便如此，不少关于手机大数据的研究仍直接通过简单的拓扑关系进行聚合。实际上，数据承载单元之间的聚合不仅需要考虑空间拓扑关系，同时还应顾及空间相关性以及空间异质性等因素。

（3）如何匹配同一观测对象的不同数据源。对于同一个地理对象，不同的传感器会记录到其不同侧面的信息，在研究中如何将其匹配是从不同角度分析对象的关键。例如，在城市人群活动的研究中，出租车车载导航轨迹可视为出租车司机一天的行程，而其手机数据则反映其一天轨迹，如果匹配成功，就可以研究其工作时间之外的日常活动规律，及其与载客习惯之间的关系。而在室内商场顾客行为的研究中，顾客的室内轨迹与其消费记录之间存在天然的鸿沟，如果二者实现贯通，则可研究室内活动与购买行为之间的关系（Liu等，2020）。

（4）如何解决大数据因时空范围不同而导致的鸿沟。由于数据获取手段的限制，研究对象的信息有可能因为不同情景而被割裂。例如，在遥感数据处理中，由于研究区范围较大而需要多幅影像拼接就属于此种情形（张良培和沈焕锋，2016）；在个体行为的研究中，同一用户室内外的活动信息分别来自不同的定位手段，因而其轨迹由于数据来源不同而被割裂。前者需要通过变换消除不同影像之间的差异，而后者则需要通过技术手段关联被观测对象的标识。社交与行为大数据中，如何聚合不同情景下的数据，形成统一而完整的个体出行轨迹仍然是当前地理大数据聚合的瓶颈。

（5）如何调和数据中不同对象之间不同的变化速率。对地观测数据所记录的对象是变化相对缓慢的地表要素，而社交与行为大数据所记录的是移动性相对较快的人群。在将二者进行综合的研究中，经常会遇到对象变化速率不一致的情景。例如，在研究气候变化对人的行为的影响时，则需要将气候数据与人的行为数据进行聚合。如何通过人的活动数据反演相对缓慢的气候变化所带来的影响，则成为研究的关键。针对此类问题，可以通过情景外推的方法进行解决（Wang等，2020）。

（6）聚合中数据共享与隐私保护的问题。除了技术瓶颈之外，存在于地理大数据聚合中的还有制度和政策的壁垒。这种非技术障碍一方面是来自不同数据采集单位之间实现数据共享的阻力，其实质是数据主体之间的利益冲突；另一方面则是来自于隐私保护的要求。对于前者，共享政策与法规的制定和执行是关键；对于后者，隐私保护政策的制定与实施，以及隐私保护策略的研究是解决问题的关键。

（7）大数据与小数据的聚合。地理大数据是非目的性的“足迹”数据，虽然数据量大、更新快、粒度细、密度高、范围广，但价值密度低，不确定性大，而目的性采样数据虽然数据量小、更新慢、范围小、密度低，但相比大数据，价值密度和质量较高。两类数据的特点基本互补，因此，如何充分发挥各自的优势取长补短，即通过小数据校正大数据，利用大数据拓展小数据，是目前大数据聚合中的重要课题。

5 地理大数据聚合驱动下新的研究领域与方向

传统的地理学研究是通过目的性的采样获得研究对象的信息，故观察角度和信息获取都存在一定限制。而在大数据时代下，多源信息的聚合，将对一些领域的研究产生巨大的推动作用。以下就以人—地关系为核心，阐述地理大数据聚合下可能会产生的新的研究领域或新的研究视角。

（1）地方的感知。关于地方（place，也有研究称为场所）的探讨一直是地理学研究的重点之一。地方与空间（space）的区别就在于地方是基于人对其的认同（Tuan，1977），例如：某城市行

政边界内的区域属于空间的概念，而其中的文化以及历史则成为地方不可或缺的组成部分。地理大数据的聚合使得对地方的研究可以进一步深入。例如，对地方的感知可以从不同渠道的大数据中获得：应用遥感数据可获得其空间结构，通过手机数据可掌握人群活动，通过微博数据可了解人群情感，利用互联网数据可再现历史事件等。不同信息的汇聚将重塑对地方的印象，为进一步系统研究、恢复地方的意境提供素材。

（2）土地功能的识别。土地功能的识别是地理研究的重要内容。传统的土地功能研究通常是以调查或者遥感数据为基础（黄波 等，2021），这种思路借助外在的形态信息对土地功能进行划分，忽视了人“使用”土地的本质，难以形成对土地功能本质的认识，并有可能忽略土地功能（尤其是城市土地功能）的混合性以及时变性。如果将土地功能作为大数据聚合的内核，则可以关联在一起的信息不仅有外在的结构（通过不同分辨率的遥感数据获得），还有其上的人的活动（通过手机数据、出租车数据获得），以及不同功能区中人的情感（通过微博数据获得）。这些数据聚合在一起就可以“反演”出传统方法所无法得到的认识：从使用者的角度观察土地的功能，甚至可以得到更细粒度（如建筑物尺度）的功能（Niu 等，2017），发掘城市功能的混合性（Xing 等，2018），甚至其随时间变化的规律。

（3）对人群的观测与理解。“人”作为地理学的核心“人—地”关系中的一方，对其观测也逐渐成为地理学重要的发展方向（刘瑜，2016）。人群活动的规律体现在多个方面，包括：人群的分布、行为、情感等，刻画这些活动的数据包括：手机通话、微博文本、签到评价、公交刷卡、共享单车数据等；同时也离不开人活动的环境和背景数据，包括街景图像、POI、监控视频等。通过对上述大数据的聚合，不仅可以研究人群的活动特征及其影响因素，探索人群活动的动力学机制，还可以实现对人群活动的预测（许珺 等，2020）。关于人的研究在地理大数据聚合的助推下，将成为人文地理乃至地理学潜在的广阔领域。

（4）功能与流动性之间关系的解析。功能分布和流动性是一对矛盾（刘瑜 等，2020）。以城市研究为例，一方面，城市功能的形成受城市多源流的影响，而另一方面，城市中人、物、信息的流动又受到城市功能区的约束。地理大数据的多样性，使得城市功能与流动性的研究更加深入，可以让我们从格局与流之间关系的视角重新审视传统的空间交互模型、城市社区结构以及距离衰减效应等，即一方面，可透过城市功能分析多源流的成因，另一方面，从城市流动性的角度理解功能区的形成与演化。功能与流动性之间关系的解析不仅适用于城市内部，而且适用于城市之间。

（5）地表复杂巨系统研究。地球表面包括土壤、大气、水、生物等多个圈层，不同圈层之间相互作用与影响形成复杂的巨系统。由于全球对地观测网的建立以及监测技术的不断发展，全球遥感及监测数据不断完善，成为全球变化、全球可持续发展、碳排放与碳中和等研究的重要推动力，并为其提供了重要的数据支撑（Guo 等，2021）。上述研究的开展需要在全球尺度聚合多源大数据，并通过复杂性理论进行分析与模拟才能揭示其中的复杂性特征与机制。

6 结 语

本文诠释了地理大数据聚合的内涵，并将地理大数据聚合定义为不同地理大数据汇聚到研究内核的过程。地理大数据聚合是传统数据融合的扩展，是数据类型、时空尺度、时空单元、研究对象等的重组。地理大数据聚合是地理大数据挖掘的重要组成部分，并大致可以分为确定内核—进行溯源—反演汇聚3个步骤，集中反映了地理大数据思维。在此基础上，本文将地理大数据聚合分为4种类型：时空聚合，面向对象聚合，面向主题聚合和面向模型聚合，地理大数据聚合之所以会有多种类型主要是源于地理大数据表达、内涵与结构的多样性。

要实现地理大数据聚合，必须解决其中存在的若干关键问题，即统一时空框架和基准、统一数据表达与存储、匹配同一观测对象的多数据源、统一多源数据的时空范围、调和不同对象之间不同的变化速率、实现数据共享与隐私保护、协调研究问题与聚合方式之间关系、大数据与小数据的聚合等。地理大数据聚合的发展还将促进对地方的理解，城市功能区的定量化，人群活动特征的识别，以及城市功能与流动性之间关系的解析等新领域的拓展。

未来随着地理大数据获取和处理技术的逐步

成熟，将推动数字地球、深时数字地球、地理虚拟环境、孪生数字空间等工程与技术的快速成长，并引领地理科学、遥感科学、测绘科学及相关学科与技术的发展。其中的关键技术之一就是如何将类型众多的地理大数据聚合在一起，而地理大数据聚合的方法与技术无疑将发挥重要的基础和支撑作用，同时它本身也必将成为地理信息科学重要的发展方向。

参考文献(References)

Allen G H and Pavelsky T M. 2018. Global extent of rivers and streams. Science, 361(6402): 585-588 [DOI: 10.1126/science.aat0636]

Alparone L, Aiazzi B, Baronti S and Garzelli A. 2015. Remote Sensing Image Fusion. Boca Raton, Florida, USA: CRC Press

Cai J X, Huang B and Song Y M. 2017. Using multi-source geospatial big data to identify the structure of polycentric cities. Remote Sensing of Environment, 202: 210-221 [DOI: 10.1016/j.rse.2017.06.039]

Chen G Q. 2014. A big data fusion model based on generalized correlation and its application in mineral prediction//Proceedings of the 13th National Symposium on Mathematical Geology and Geomatics Information. Xuzhou: Geological Society of China: 210 (陈国强. 2014. 基于广义相关性的大数据融合模型及在矿产预测中的应用//第十三届全国数学地质与地学信息学术研讨会. 徐州: 中国地质学会: 210)

Chen H. 2020. Video and IoT big data fusion analysis application platform. Digital Technology and Application, 38(8): 114-115 (陈慧. 2020. 视频与物联网大数据融合分析应用平台. 数字技术与应用, 38(8): 114-115) [DOI: 10.19695/j.cnki.cn12-1369.2020.08.44]

Chen Y H, Zhang R J, Ge Y, Jin Y and Xia Z L. 2019. Downscaling census data for gridded population mapping with geographically weighted area-to-point regression kriging. IEEE Access, 7: 149132-149141 [DOI: 10.1109/ACCESS.2019.2945000]

Crowther T W, Glick H B, Covey K R, Bettigole C, Maynard D S, Thomas S M, Smith J R, Hintler G, Duguid M C, Amatulli G, Tuanmu M N, Jetz W, Salas C, Stam C, Piotto D, Tavani R, Green S, Bruce G, Williams S J, Wiser S K, Huber M O, Hengeveld G M, Nabuurs G J, Tikhonova E, Borchardt P, Li C F, Powrie L W, Fischer M, Hemp A, Homeier J, Cho P, Vibrans A C, Umunay P M, Piao S L, Rowe C W, Ashton M S, Crane P R and Bradford M A. 2015. Mapping tree density at a global scale. Nature, 525(7568): 201-205 [DOI: 10.1038/nature14967]

Du H. 2013. Research on In-Depth Aggregation of Academic Information Resource on the Basis of Coupling Relationships. Wuhan: Wuhan University: 43-45 (杜晖. 2013. 基于耦合关系的学术信息资源深度聚合研究. 武汉: 武汉大学: 43-45)

Frachetti M D, Smith C E, Traub C M and Williams T. 2017. Nomadic ecology shaped the highland geography of Asia's Silk Roads. Nature, 543(7644): 193-198 [DOI: 10.1038/nature21696]

Ginsberg J, Mohebbi M H, Patel R S, Brammer L, Smolinski M S and Brilliant L. 2009. Detecting influenza epidemics using search engine query data. Nature, 457(7232): 1012-1014 [DOI: 10.1038/nature07634]

Guo H D, Chen F, Sun Z C, Liu J and Liang D. 2021. Big Earth Data: a practice of sustainability science to achieve the Sustainable Development Goals. Science Bulletin, 66(11): 1050-1053 [DOI: 10.1016/j.scib.2021.01.012]

Hsiang S M, Burke M and Miguel E. 2013. Quantifying the influence of climate on human conflict. Science, 341(6151): 1235367 [DOI: 10.1126/science.1235367]

Huang B and Jiang X L. 2021. An enhanced unmixing model for spatiotemporal image fusion. National Remote Sensing Bulletin, 25(1): 241-250 (黄波, 姜晓璐. 2021. 增强型空间像元分解时空遥感影像融合算法. 遥感学报, 25(1): 241-250) [DOI: 10.11834/jrs.20210459]

Huang D C and Wang J F. 2018. Monitoring hand, foot and mouth disease by combining search engine query data and meteorological factors. Science of the Total Environment, 612: 1293-1299 [DOI: 10.1016/j.scitotenv.2017.09.017]

Jia Y X, Ge Y, Ling F, Guo X, Wang J H, Wang L, Chen Y H and Li X D. 2018. Urban land use mapping by combining remote sensing imagery and mobile phone positioning data. Remote Sensing, 10 (3): 446 [DOI: 10.3390/rs10030446]

Li C M. 2020. Evaluation of public transport service index in multi-source big data fusion. Journal of Xiamen University of Technology, 28(3): 77-83 (李传明. 2020. 多源大数据融合的公交服务指数评价. 厦门理工学院学报, 28(3): 77-83) [DOI: 10.19697 /j.cnki.1673-4432.202003013]

Li M, Yuan W, Yuan W, Niu F Q, Li H Q and Hu D M. 2021. Big data analysis on geographical relationship of the Arctic based on news reports. Acta Geographica Sinica, 76(5): 1090-1104 (李萌, 袁文, 袁武, 牛方曲, 李汉青, 胡段牧. 2021. 基于新闻大数据的北极地区地缘关系研究. 地理学报, 76(5): 1090-1104) [DOI: 10.11821/dlxb202105004]

Li S T, Li C Y and Kang X D. 2021. Development status and future prospects of multi-source remote sensing image fusion. National Remote Sensing Bulletin, 25(1): 148-166 (李树涛, 李聪妤, 康旭东. 2021. 多源遥感图像融合发展现状与未来展望. 遥感学报, 25(1): 148-166) [DOI: 10.11834/jrs.20210259]

Li X and Huang C L. 2004. Data assimilation: a new means for multi-source geospatial data integration. Science and Technology Review, (12): 13-16 (李新, 黄春林. 2004. 数据同化——一种集成多源地理空间数据的新思路. 科技导报, (12): 13-16) [DOI: 10.3321/j.issn:1000-7857.2004.12.004]

Li X J, Zhang C R, Li W D, Ricard R, Meng Q Y and Zhang W X. 2015. Assessing street-level urban greenery using Google Street View and a modified green view index. Urban Forestry and Urban Greening, 14(3): 675-685 [DOI: 10.1016/j.ufug.2015.06.006]

Liu Y. 2016. Revisiting several basic geographical concepts: a social sensing perspective. Acta Geographica Sinica, 71(4): 564-575 (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. 地理学报, 71(4): 564-575) [DOI: 10.11821/dlxb201604003]

Liu Y, Yao X, Gong Y X, Kang C G, Shi X, Wang F H, Wang J E, Zhang Y, Zhao P F, Zhu D and Zhu X Y. 2020. Analytical methods and applications of spatial interactions in the era of big data. Acta Geographica Sinica, 75(7): 1523-1538 (刘瑜, 姚欣, 龚咏喜, 康朝贵, 施迅, 王法辉, 王姣娥, 张毅, 赵鹏飞, 朱递, 朱欣焰.

2020. 大数据时代的空间交互分析方法和应用再论. 地理学报, 75(7): 1523-1538) [DOI: 10.11821/dlxb202007014]

Liu Y X, Cheng D Y, Pei T, Shu H, Ge X H, Ma T, Du Y Y, Ou Y, Wang M and Xu L M. 2020. Inferring gender and age of customers in shopping malls via indoor positioning data. Environment and Planning B: Urban Analytics and City Science, 47(9): 1672-1689 [DOI: 10.1177/2399808319841910]

Meng X F and Du Z J. 2016. Research on the big data fusion: issues and challenges. Journal of Computer Research and Development, 53(2): 231-246 (孟小峰, 杜治娟. 2016. 大数据融合研究: 问题与挑战. 计算机研究与发展, 53(2): 231-246) [DOI: 10.7544/issn1000-1239.2016.20150874]

Niu N, Liu X P, Jin H, Ye X Y, Liu Y, Li X, Chen Y M and Li S Y. 2017. Integrating multi-source big data to infer building functions. International Journal of Geographical Information Science, 31(9): 1871-1890 [DOI: 10.1080/13658816.2017.1325489]

Pei T, Guo S H, Yuan Y C, Zhang X Y, Yuan W, Gao A, Zhao Z Y and Xue C J. 2019a. Public security event themed web text structuring. Journal of Geo-Information Science, 21(1): 2-13 (裴韬, 郭思慧, 袁烨城, 张雪英, 袁文, 高昂, 赵志远, 薛存金. 2019a. 面向公共安全事件的网络文本大数据结构化研究. 地球信息科学学报, 21(1): 2-13) [DOI: 10.12082/dqxxkx.2019.180680]

Pei T, Liu Y X, Guo S H, Shu H, Du Y Y, Ma T and Zhou C H. 2019b. Principle of big geodata mining. Acta Geographica Sinica, 74(3): 586-598 (裴韬, 刘亚溪, 郭思慧, 舒华, 杜云艳, 马廷, 周成虎. 2019b. 地理大数据挖掘的本质. 地理学报, 74(3): 586-598) [DOI: 10.11821/dlxb201903014]

Pei T, Sobolevsky S, Ratti C, Shaw S L, Li T and Zhou C H. 2014. A new insight into land use classification based on aggregated mobile phone data. International Journal of Geographical Information Science, 28(9): 1988-2007 [DOI: 10.1080/13658816.2014.913794]

Pei T, Wang X, Song C, Liu Y X, Huang Q, Shu H, Chen X, Guo S H and Zhou C H. 2021. Review on spatiotemporal analysis and modeling of COVID-19 pandemic. Journal of Geo-Information Science, 23(2): 188-210 (裴韬, 王席, 宋辞, 刘亚溪, 黄强, 舒华, 陈晓, 郭思慧, 周成虎. 2021. COVID-19疫情时空分析与建模研究进展. 地球信息科学学报, 23(2): 188-210) [DOI: 10.12082/dqxxkx.2021.200434]

Pohl C and Van Genderen J L. 1998. Review article multisensor image fusion in remote sensing: concepts, methods and applications. International Journal of Remote Sensing, 19(5): 823-854 [DOI: 10.1080/0143116982157481].

Qin W F. 2018. Research on Big Data Fusion Model based on Tensor. Jilin: Northeast Electric Power University (秦文斐. 2018. 基于张量的大数据融合模型研究. 吉林: 东北电力大学)

Shen Z H. 2020. An analysis of the applications and challenges of deep learning in urban big data fusion. China CIO News, (3): 164, 166 (申占恒. 2020. 深度学习在城市大数据融合中的应用与挑战分析. 信息系统工程, (3): 164, 166) [DOI: 10.3969/j.issn.1001-2362.2020.03.075]

Song J C, Lin T, Li X H and Prishchepov A V. 2018. Mapping urban functional zones by integrating very high spatial resolution remote sensing imagery and points of interest: a case study of Xiamen, China. Remote Sensing, 10(11): 1737 [DOI: 10.3390/rs10111737]

Soto V and Frías-Martínez E. 2011. Automated land use identification using cell-phone records//Proceedings of the 3rd ACM international workshop on MobiArch. Bethesda, Maryland, USA: Association for Computing Machinery: 17-22 [DOI: 10.1145/2000172.2000179]

Tuan Y F. 1977. Space and Place: The Perspective of Experience. Minneapolis: University of Minnesota Press

Wang J H, Obradovich N and Zheng S Q. 2020. A 43-million-person investigation into weather and expressed sentiment in a changing climate. One Earth, 2(6): 568-577 [DOI: 10.1016/j.oneear.2020.05.016]

Wang J Y, Su Y X, Ling G M, Chen C and Chen W Z. 2020. Designing a tourism flow research methodology based on multi-dimensional big data fusion analysis - a case study in Guangzhou City, China. Scientific and Technological Innovation Information, (5): 59-60 (王嘉茵, 苏宇新, 凌国明, 陈忱, 陈伟钊. 2020. 基于多维度大数据融合分析的旅游流研究方法设计——以广州市为例. 科学技术创新, (5): 59-60) [DOI: 10.3969/j.issn.1673-1328.2020.05.036]

Wang N, Du Y Y, Liang F Y, Yi J W and Wang H M. 2019. Spatiotemporal changes of urban rainstorm-related micro-blogging activities in response to rainstorms: a case study in Beijing, China. Applied Sciences, 9(21): 4629 [DOI: 10.3390/app9214629]

Weng Y L and Tian Q J. 2003. Analysis and evaluation of method on remote sensing data fusion. Remote Sensing Information, (3): 49-54 (翁永玲, 田庆久. 2003. 遥感数据融合方法分析与评价综述. 遥感信息, (3): 49-54) [DOI: 10.3969/j.issn.1000-3177.2003.03.015]

Wu X H, Xiao Y, Li L S and Wang G J. 2017. Review and prospect of the emergency management of urban rainstorm waterlogging based on big data fusion. Chinese Science Bulletin, 62(9): 920-927 (吴先华, 肖杨, 李廉水, 王国杰. 2017. 大数据融合的城市暴雨内涝灾害应急管理述评. 科学通报, 62(9): 920-927) [DOI: 10.1360/N972016-01080]

Xing H F, Meng Y and Shi Y. 2018. A dynamic human activity-driven model for mixed land use evaluation using social media data. Transactions in GIS, 22(5): 1130-1151 [DOI: 10.1111/tgis.12447]

Xu J, Xu Y, Hu L and Wang Z B. 2020. Discovering spatio-temporal patterns of human activity on the Qinghai-Tibet Plateau based on crowdsourcing positioning data. Acta Geographica Sinica, 75(7): 1406-1417 (许珺, 徐阳, 胡蕾, 王振波. 2020. 基于位置大数据的青藏高原人类活动时空模式. 地理学报, 75(7): 1406-1417) [DOI: 10.11821/dlxb202007006]

Yan P C. 2019. Analysis on the application of large data fusion in public security video surveillance. Digital Technology and Application, 37(9): 30-31 (颜培超. 2019. 浅析公共安全视频监控大数据融合应用. 数字技术与应用, 37(9): 30-31) [DOI: 10.19695/j.cnki.cn12-1369.2019.09.17]

Yao F G, Zhong X X and Zhou J C. 2018. Granular computing: a new method of intelligent modeling for big data fusion. Journal of Nanjing University of Science and Technology, 42(4): 503-510 (姚富光, 钟先信, 周靖超. 2018. 粒计算: 一种大数据融合智能建模新方法. 南京理工大学学报, 42(4): 503-510) [DOI: 10.14177/j.cnki.32-1397n.2018.42.04.017]

Zhang F, Wu L, Zhu D and Liu Y. 2019. Social sensing from street-level imagery: a case study in learning spatio-temporal urban mobility patterns. ISPRS Journal of Photogrammetry and Remote Sensing, 153: 48-58 [DOI: 10.1016/j.isprsjprs.2019.04.017]

Zhang H, Yu P, Li G and Yu J. 2019. “One Map” planning based spa-

tial plan information sharing. Planners, 35(21): 11-15 (张恒, 于鹏, 李刚, 于靖. 2019. 空间规划信息资源共享下的"一张图"建设探讨. 规划师, 35(21): 11-15) [DOI: 10.3969/j.issn.1006-0022.2019.21.002]

Zhang J X. 2011. The trend of development of multi-source remote sensing data fusion. Geomatics World, 9(2): 18-20 (张继贤. 2011. 多源遥感数据融合的发展趋势. 地理信息世界, 9(2): 18-20) [DOI: 10.3969/j.issn.1672-1586.2011.02.004]

Zhang L F, Peng M Y, Sun X J, Cen Y and Tong Q X. 2019. Progress and bibliometric analysis of remote sensing data fusion methods (1992—2018). Journal of Remote Sensing, 23(4): 603-619 (张立福, 彭明媛, 孙雪剑, 岑奕, 童庆禧. 2019. 遥感数据融合研究进展与文献定量分析(1992—2018). 遥感学报, 23(4): 603-619) [DOI: 10.11834/jrs.20199073]

Zhang L P and Shen H F. 2016. Progress and future of remote sensing data fusion. Journal of Remote Sensing, 20(5): 1050-1061 (张良培, 沈焕锋. 2016. 遥感数据融合的进展与前瞻. 遥感学报, 20(5): 1050-1061) [DOI: 10.11834/jrs.20166243]

Zhang L Y, Pei T, Wang X, Wu M B, Song C, Guo S H and Chen Y J. 2020. Quantifying the urban visual perception of Chinese traditional-style building with street view images. Applied Sciences, 10(17): 5963 [DOI: 10.3390/app10175963]

Zhu Y Q, Tan Y J, Zhang J T, Mao B, Shen J and Ji C F. 2015. A framework of Hadoop based geology big data fusion and mining technologies. Acta Geodaetica et Cartographica Sinica, 44(S1): 152-159 (朱月琴, 谭永杰, 张建通, 毛波, 沈婕, 汲超飞. 2015. 基于Hadoop的地质大数据融合与挖掘技术框架. 测绘学报, 44(S1): 152-159) [DOI: 10.11947/j.AGCS.2015.F059

Big geodata aggregation: Connotation, classification, and framework

PEI Tao[1,2,3], HUANG Qiang[1,2], WANG Xi[1,2], CHEN Xiao[1,2], LIU Yaxi[1,2], SONG Ci[1,2], CHEN Jie[1,2], ZHOU Chenghu[1,2]

1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China;
2. University of Chinese Academy of Sciences, Beijing 100491, China;
3. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

Abstract: Big geodata can be regarded as the "footprints" generated by geographic objects, which are divided into two types: big earth observation data and big social behavior data. Big earth observation data refer to remote sensing data and all kinds of monitoring data from ground observation stations; big social behavior data are defined as the data generated by humans, such as mobile phone data, traffic card data, social media data, and trajectory data of floating cars. Big earth observation data contain the information of land surface, and big social behavior data record the social behavior of humans. The essence of big geodata mining is to reveal the man-land relationship via the retrieval of the "footprint" of geographic objects. Magnificent progresses in recent years demonstrate that most of them have been made through the aggregation of multiple types of big geodata. The significance of big geodata aggregation can be summarized into three aspects: to represent a new research paradigm, to produce a new research angle, and to improve the research quality by combining different types of information. In contrast to image fusion, which is regarded as the combination of imageries with the same data structure, big geodata aggregation is defined as the merging of different types of big geodata into a multidimensional dataset through transformation in terms of structure, content, and representation. In general, big geodata aggregation includes big earth observation data and big social behavior data. Given the differences in data source, data structure, connotation, and granularity, the aggregation procedure is various and complicated. Considering the characteristics and the classification of big geodata, we classify big geodata aggregation into four types: spatiotemporal aggregation, object-oriented aggregation, topic-oriented aggregation, and model-oriented aggregation. In spatiotemporal aggregation, data are merged via their common locations, such as the aggregation between POI and remote sensing data. In object-oriented aggregation, data are connected through the same object, such as the aggregation of indoor and outdoor trajectory data via the same phone carrier. In topic-oriented aggregation, data with the same topic are integrated, such as the aggregation of different types of information sourced from the same earthquake event. In model-oriented aggregation, data are assimilated via the same model, such as the aggregation of street view data and POI data when using a deep learning model for the recognition of building style. As an important part of big geodata mining, the procedure of big geodata aggregation can be divided into four steps: the determination of the aggregation core, information back tracking, retrieving, and converging. The key issues in big geodata aggregation include unifying spatiotemporal benchmarks, uniting data representation and storage, matching multisource big geodata from the same geo-object, centralizing the spatiotemporal scope of multisource big geodata, settling the changing velocities of different objects, solving data-sharing and privacy protection problems, and aggregating big and small geodata. With the development of methods and techniques of big geodata aggregation, some research areas, including the sensing of places, the identification of land functions, the detection of crowd activities, the determination of the relationship between urban functions and multisource flow, and the introduction of a land surface complex giant system, may be substantially expanded.

Key words: big geodata philosophy, urban functional area, crowds activities, geographic information science, data fusion
Supported by National Natural Science Foundation of China (No. 41525004, 42071436)