城市计算：用大数据和AI驱动智能城市

郑　宇[1,2]
[1] 微软亚洲研究院
[2] 上海交通大学

关键词：城市计算　大数据　时空数据分析　城市感知

近年来，随着感知技术和计算环境的成熟，各种大数据悄然而生，如交通流、气象数据、道路网、兴趣点、移动轨迹和社交媒体等。同时，人工智能（尤其是机器学习）算法的成熟也为数据分析提供了利器。如果使用得当，我们就可以利用这些数据和智能算法来解决城市所面临的问题，如空气污染、交通拥堵、能耗增加、规划落后等。城市计算通过对多源异构数据的整合、分析和挖掘来提取知识和智能，并结合行业知识来创造“人 - 环境 - 城市”三赢的局面。

城市计算的概念和框架

城市计算是一个交叉学科，是计算机科学以城市为背景，与城市规划、交通、能源、环境、社会学和经济学等学科融合的新兴领域。具体而言，城市计算是一个通过不断获取、整合和分析城市中多源异构的大数据来解决城市所面临的挑战的过程。城市计算将无处不在的感知技术、高效的数据管理和强大的机器学习算法，以及新颖的可视化技术相结合，致力于提高人们的生活品质，保护环境和促进城市运转效率，帮助我们理解各种城市现象的本质，甚至预测城市的未来[1]。

图1给出了城市计算的基本框架，包括城市感知及数据采集、数据管理、城市数据分析、服务提供等四个环节。

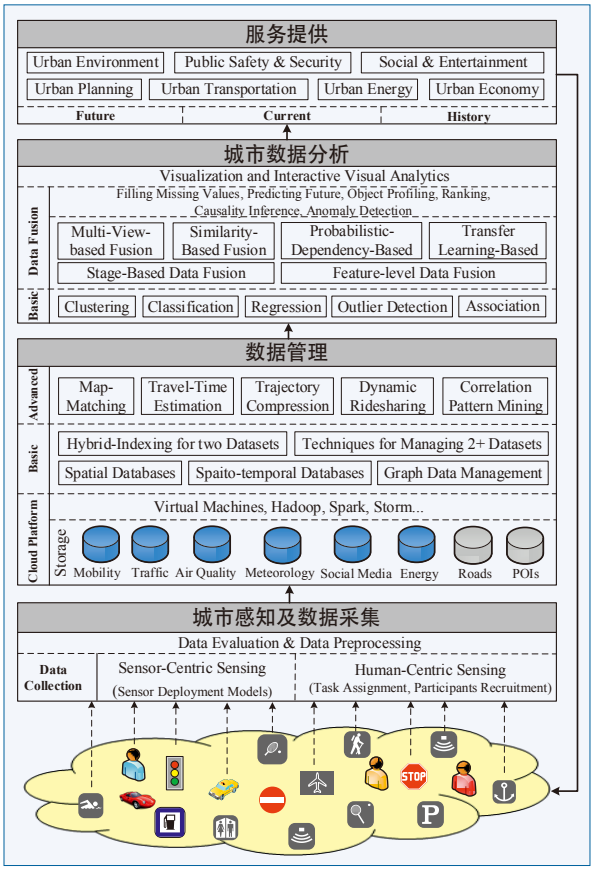


图1　城市计算的基本框架

在城市感知层面，我们可以通过车载 GPS 或用户的智能手机产生的轨迹数据来不断感知司机在道路上的驾驶状态，也可以收集用户发布在社交媒体上的信息。在数据管理层面，我们通过时空索引结构把司机产生的大规模轨迹和社交媒体数据高效地组织和管理起来，以供后续实时分析和挖掘。在数据分析层面，当城市出现异常时，我们可以根据这些轨迹数据较为准确地确定异常发生的空间范围和时间区间。因为，当异常发生后，各条道路上的车流量以及人们选择的行车路线都会发生改变。我们可以有针对性地利用与这些地方及时间段相关联的（而不是全部）社交媒体来分析异常出现的起因。在服务提供层面，这些信息会被及时地传递到交通管理部门和周边通行的人群，以快速处理异常并避免更多的人陷入混乱 [2]。

城市感知

感知模式

城市感知层面主要通过以传感器为中心和以人为中心的两大类感知方式来不断收集和获取城市里的各种数据。

在以传感器为中心的感知场景里，我们可以把传感器部署到固定的地点（称为固定感知）或者移动的物体上（称为移动感知），让传感器自动地收集并发送数据。如图 2(a) 和 (b) 所示。一旦传感器部署完毕，人就无须参与数据的采集过程。

在以人为中心的感知场景里，每个人被看作一个传感器，不断感知周边的情况，利用人群的数量和移动来捕获城市的韵律。如图 2(c) 和 (d) 所示。这个场景可以进一步细分为被动群体感知和主动群体感知。在被动群体感知的场景里（比如使用手机通信，或者刷卡进出地铁站），人们并不知道自己在贡献数据，也不清楚被收集的数据的用途（比如用户出入地铁站的刷卡数据可以用来感知地铁站里的人流，从而优化列车调度）。在主动群体感知的场景里，人们知道感知任务的目的、时间和用途，可以选择参加或者不参加。为了吸引用户贡献数据，主动群体感知通常采用用户激励机制（如通过金钱或积分来激励用户更多地参与和贡献数据）[3]。

城市感知面临的挑战

数据的偏斜采样问题 我们在城市里获取的数据只是数据全集的一个样本，很多属性在此样本上的分布与其在全集上不一致。例如，我们获得了一个城市里出租车的轨迹数据，但出租车只是所有行驶在路面上的车辆的一部分。出租车在路面上的分布与所有车流在路面上的分布可能很不一样，如图 3 所示。某些路面上有很多出租车，却没有太多私家车，反之亦然。因此，不能简单地根据路面上出租车的数量来估计道路的整体车流量 [4]。如何从偏斜的采样数据中准确推断出完整数据蕴含的知识是一大挑战。

数据稀疏性 由于资源的限制，我们只能部署有限的传感器，如何根据有限的传感器产生的稀疏

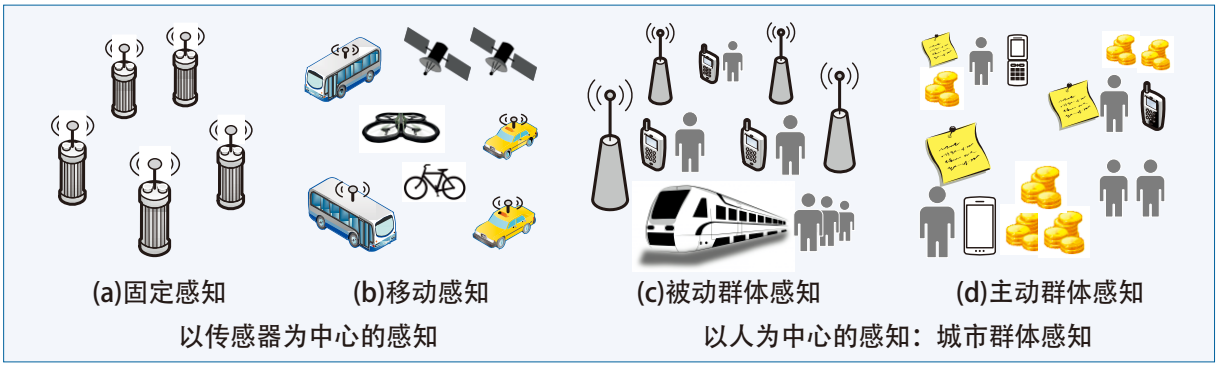


图2 城市感知的四种模式

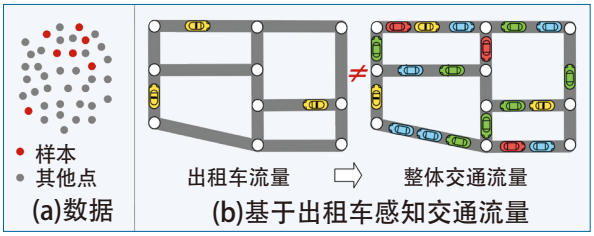


图3 城市感知中的数据偏斜采样问题

数据来感知整个城市状态的全貌是一个难点。如图 4 所示，北京城市面积很大，但由于空气质量监测站点价格昂贵，且占地面积较大，目前只部署了 35 个监测点。如何根据这 35 个监测点的数据推断出没有部署监测点的地方的空气质量就变得很关键。由于城市的空气并不是均匀分布的，受到交通、污染源分布和扩散条件等诸多因素的影响，使得我们不能通过简单的线性插值方法来计算未部署监测点的地方的空气质量 [5]。

图4 城市感知中的数据稀疏性问题

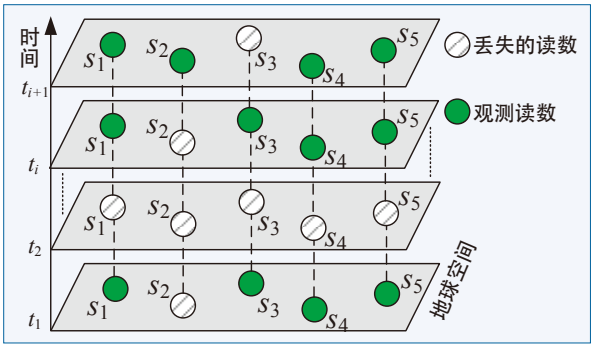


图5 城市感知中的数据丢失问题

数据丢失 由于传感器和通信链路会出现故障，使得本应该获取到的数据会出现丢失。图 5 中的白色圆圈表示丢失的数据，每一个纵向序列表示一个传感器在不同时间点的读数。数据的缺失会给城市状态的感知以及后面的分析和挖掘过程带来很大的挑战。由于我们无法预先得知数据将在什么时候丢失，丢失的形式也很多样，而且，由于存在地域差异（工业区和公园）和环境的突变（比如刮大风），有些时候并不是时间或空间上越临近的点的读数越相似 [6]，如何高效、准确地填充这些缺失的数据是一个难题。

资源的合理配置 我们拥有的资源（如土地、经费等）是有限的，如何合理使用这些资源，并保证较好的感知效果是一个难题，原因有两个方面：

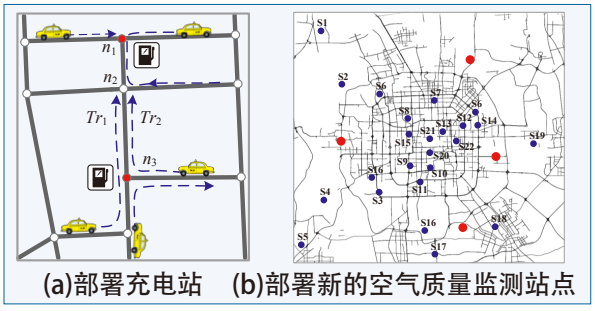


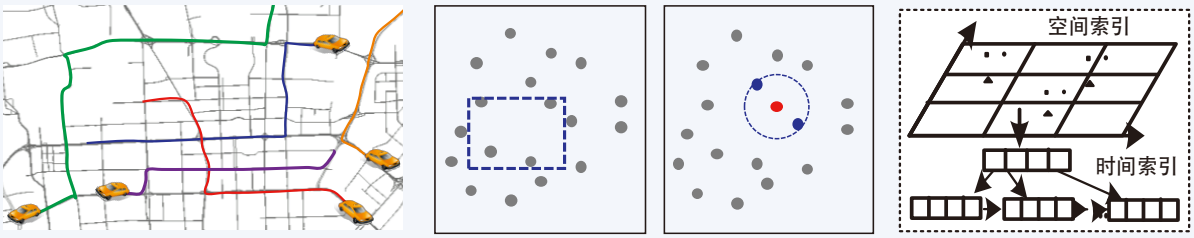
图6 城市感知中的资源配置难题

第一，从大量的候选集中挑选有限的 k 个地方来部署传感器，从而使得某种指标性能最优，这往往是一个 NP-Hard 问题。如图 6(a) 所示，假设我们的经费只能部署 5 个充电桩，把这些充电桩放在哪 5 个路口才能使得它们覆盖的总体车流最大化，这是一个 NP-Hard 问题 [7]。

第二，衡量标准未知。如图 6(b) 所示，假设我们可以在北京新增 4 个空气质量监测站点，这些站点应该放在哪里才能使得整个空气质量监测系统的监控效果最好？在一个地方尚未部署监测站点之前，我们其实并不清楚那个地方的空气质量如何，又如何清楚该在什么地方部署呢？弄清楚这个标准本身就是一个难题 [8]。

城市数据管理

城市中的数据规模大、种类繁多、变化快，而且都带有很强的时空属性，通常被称为时空大数据。城市数据管理通常都需要云计算平台的支撑，但由



(a)独特的数据结构　(b)特有的查询方式　(c)混合式索引

图7　城市数据管理的难题

于以下四个方面的挑战，现有的云平台并不能很好地支持大规模时空数据的管理。

独特的数据结构：时空数据的结构与文本和图像有很大的区别。比如，一张照片一旦被拍摄下来，它的大小就不会再变化。但一辆车的轨迹会随着它在城市中的移动而不断变长，我们无法预知这条轨迹的长度和大小，如图 7(a) 所示。轨迹数据通常以一种流数据的形式存在，一条轨迹中两个点的顺序不能交换，否则会带来完全不一样的语义 [9]。现有云计算平台的存储方式，对支持时空数据的高效查询和分析并不友好 [10]。

特有的查询方式：我们通常会利用关键词来查询文档，通过关键词与文档中包含词的精确或者近似匹配把相关文档找出来。但面对时空数据，我们通常会做时空范围查询以及最近邻查询。例如，查找过去两分钟内在我周边 100 米范围内空驶过的出租车，这就是一个空间范围（我周边 100 米）加上时间范围（过去两分钟）查询（如图 7(b) 所示）。当我在开车时，查找离我最近的加油站，这是一个连续的最近邻查询。有时为了估计道路通行时间，我们甚至需要查找出经过一条给定线路的车辆轨迹 [11]。处理这些查询需要用到独特的时空数据管理和索引技术，而这些技术在现有的云平台上是缺失的。

混合式索引：在很多城市计算的应用场景中，我们需要用到多种数据，并融合多源数据中的知识。如果孤立地管理这些数据，在线学习时就不能及时提取由多个数据共同带来的知识，更来不及做知识融合。除了对每个数据单独索引之外，还需要对多源异构数据建立混合式索引，提高数据之间的关联效率（如图 7(c) 所示）。比如我们要学习空气质量、交通流量、楼房密度和天气之间的关联关系，查找出诸如“当天气有雾、交通拥堵，且周边楼房密度大于 500 栋 / 每平方公里时，则此处的空气质量就是污染”的跨域关联规则，我们需要看每两种数据共同出现的规律。

时空索引和分布式系统的整合：传统云计算平台里配备的分布式计算环境（如 Hadoop, Spark 和 Storm）虽然能利用多个计算节点来提升算法的运行效率，但其本身并没有集成时空索引结构，在处理大规模的动态时空数据时，为保证数据处理的实时性，不得不动用更多的计算资源，而且性能面临瓶颈。将时空索引算法和分布式计算环境有机结合，可以利用更少的计算资源来获得更高效的计算性能，从而可以尝试更为复杂和精准的算法。如何将两者有机结合，需要同时具备时空数据库和分布式计算环境的知识。

城市数据分析

城市数据分析是指利用人工智能技术挖掘和融合城市大数据中的知识，直观地展现结果，并最终解决行业问题。在这个层面，我们面临四方面的挑战。

针对时空数据的机器学习算法：早年间提出的机器学习算法大部分都是用来处理图像、视频和文本数据。但由于时空数据与图像和文本有着本质的区别，这些机器学习算法并不能在时空数据上发挥最大的效力。时空数据的特性包括空间距离、空间层次、时间临近性、周期性和趋势性（如图 8 所

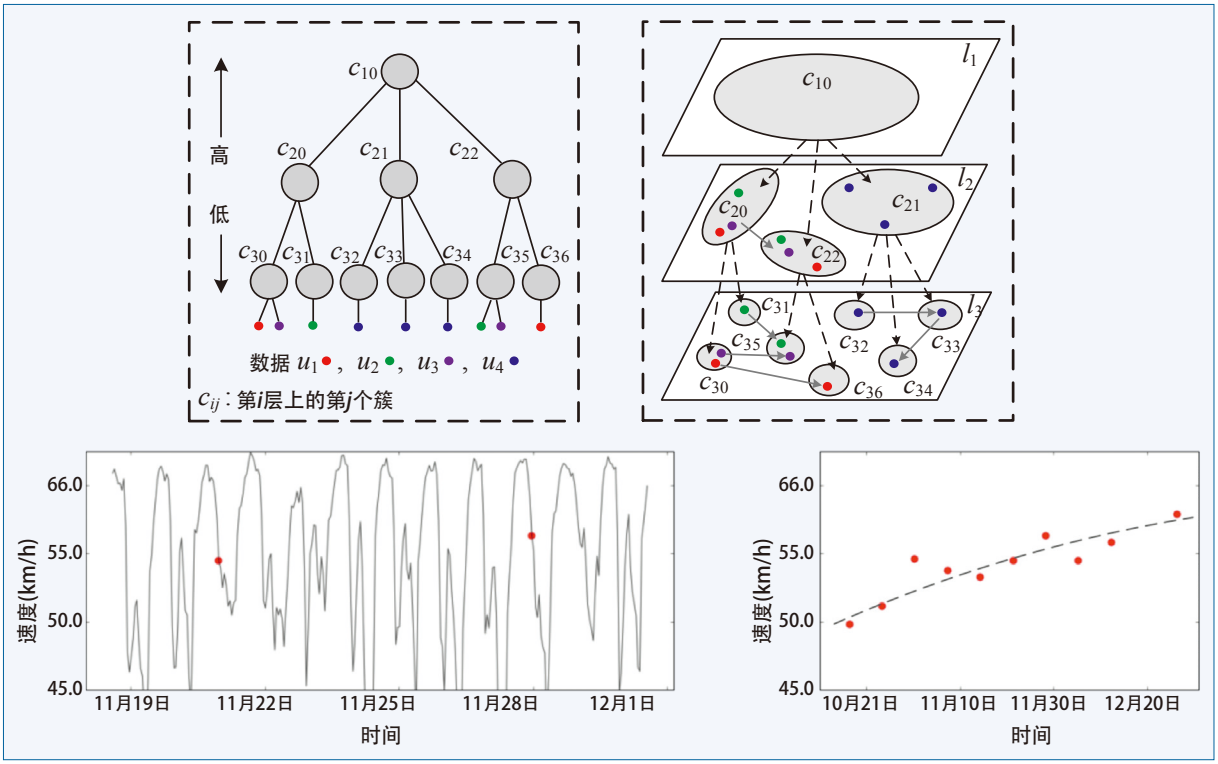


图8　时空数据的特性

示）。这些特性在文本和图像数据中并不存在或者不明显。如何针对时空数据的这些特性来设计更为合理的机器学习算法是一大挑战。比如在深度学习领域针对图像数据设计的卷积神经网络 (CNN) 和针对文本数据提出的长短时记忆 (LSTM) 都不能很好地预测城市区域的人流量 [12]。针对时空数据需要特别设计新的深度学习模型 [13]。

多源数据融合：传统的数据挖掘往往只从单一数据中获取知识。比如，“啤酒和尿布”这个例子就是从单一的超市交易记录里获取的关联规则。但在城市计算的场景里，我们往往需要融合来自多个数据源的知识才能解决一个复杂的问题。比如，通过融合来自交通、气象和地理信息的知识来预测空气质量。如何融合那些看似毫不相干的领域的数据知识是一大挑战，也是大数据领域研究的关键问题 [14]。

融合数据管理和机器学习：长期以来，机器学习和数据库一直是两个独立的领域，两者有各自的研究课题和社区，交集很少。但在一个实际的城市计算应用中，我们需要将两者的知识有机地融合在一起，才能设计出又快又好的算法，解决复杂且实时性要求高的问题。这需要对两个领域都有深刻的理解。

交互动态可视分析：早年间的数据挖掘往往是利用算法从数据中获取结果的一个单向的、静态的过程。在城市计算的场景里，我们不仅需要具备数据科学知识，还需要具备行业背景知识，只有将两者结合才能解决实际问题。但是数据科学家和行业从业者有着各自的知识体系，双方鲜有交集。这给人工智能技术在行业中落地带来了很大的挑战。除了让数据科学家尽量学习行业知识，使其研究贴近应用需求外，交互可视分析也提供了一个很好的知识共享机制，它通过把行业专家纳入数据挖掘的环节中来实现人与机器的智能融合，以及数据科学与行业知识的对接。比如，让算法考虑简单的标准，产生初步的计算结果，然后让行业专家根据自己积累的行业知识对结果进行筛选和过滤，去掉不合理

的结果，保留合理的结果，然后再让算法去计算。通过这样的迭代，最终促使人机智能的有效融合。

服务提供

按照时效性，城市计算的服务可以分为厘清现状、预测未来、洞察历史三种类型。以空气污染为例，根据有限的空气质量监控站点，结合气象、交通等其他数据源来计算整个城市任意角落的空气质量，此为厘清现状[5]。对未来两天空气质量的估计，为预测未来[15]。根据多年历史数据分析出污染物的来源和成因，此为洞察历史[16]。

按照服务的行业划分，城市计算涵盖城市交通和规划、城市环境、城市能源、城市商业、公共安全、教育、医疗、社交和娱乐等。

在服务提供层面，我们面临三个方面的挑战：

融合行业知识和数据科学：数据科学家和行业专家通常掌握各自的技能和知识体系，两者鲜有交集。但一个城市大数据项目落地需要同时具备这两方面的知识。如何发掘有效的沟通和协作机制来融合两方面的知识是一大挑战。

系统对接：在传统的城市行业里，有很多业务系统。系统中有很多复杂而重要的业务逻辑。一个新兴的城市计算系统不太可能（也没有必要）完全替换现有的行业系统。如何实现两个系统的无缝对接存在挑战。如果要把新的城市计算技术完全嵌入到现有系统的内部，需要数据科学家深入、彻底地学习行业知识和业务逻辑，负担较大。而且，由于很多机器学习算法需要定期重新训练模型，这种完全的嵌入模式容易将（数量极其有限的）数据科学家牢牢捆绑在某个项目上不得脱身，限制了数据科学家为更多的行业服务，最终导致服务不可扩展。另一方面，由于很多行业数据存在机密性和隐私，数据不能向外提供，因此，完全孤立在外的城市计算系统也难以获取数据输入。

培养数据科学家：数据科学家需要自己寻找问题、分析问题、选择数据、设计模型、搭建基于云的系统，最后解决行业问题。

首先，数据科学家需要理解行业问题。要知道是什么因素导致了某个行业问题的出现（如空气污染与厂矿废气排放、汽车尾气、扩散条件和气象因素均有关系），原有行业的研究有哪些成果和思想值得我们借鉴，为什么原有的方法有缺陷等（如图9所示）。这样，我们才能选择合适的数据，并设计合理的基于数据科学的方法来解决行业问题。也只有这样，我们才能够掌握行业的语言，与行业进行有效的沟通，并让行业接受基于数据和人工智能的方法，最终让城市计算的技术落地。

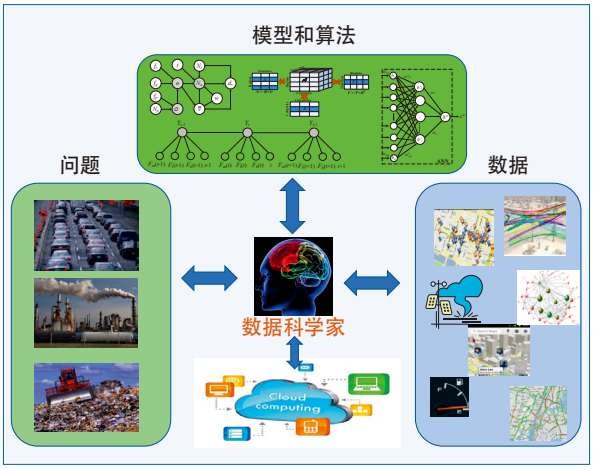


图9　数据科学家需要理解行业问题

其次，数据科学家要深知各种数据背后蕴含的知识，而不是仅仅停留在数据的格式和表征上。比如，路面上出租车的GPS轨迹既反映了路面的速度和流量信息，也反映了人们的出行规律。大量这样的出行数据进一步反映了一个区域的功能，而区域的功能又影响到这个区域的经济、能耗和环境等。有了这样的思路，我们就可以利用领域A的数据解决领域B的问题，既实现了跨域数据的融合，也缓解了某个领域数据不足（或缺失）带来的压力。

第三，数据科学家需要懂得数据科学里各个环节的知识，包括数据管理、数据挖掘、机器学习和可视化等。要解决一个实际的大数据问题，需要多项技能的综合运用（不仅仅只是人工智能技术）。

第四，数据科学家需要懂得如何使用（甚至是如何修改和增强）云计算平台。云平台的架构

会影响到数据管理和挖掘算法的设计。很多算法在单机上可以正常运行，但在云平台上就需要进行重大改进。

培养一位优秀的数据科学家周期长、成本高、成材率低，很难量产，这很可能会成为未来人工智能与传统产业结合的一大瓶颈。

城市大数据平台

要解决以上各种挑战，我们需要具备四个方面的知识：大数据、人工智能、云计算和行业知识。要让城市计算的技术落地，我们还需要搭建一个城市大数据平台。该平台可以基于传统的云计算平台来搭建，但不仅仅是现有的云计算，因为现有的云计算平台缺乏对时空数据的有效管理机制（如时空索引、混合式索引以及这些索引与分布式系统的结合），也缺乏针对时空数据特有的人工智能算法。

图 10 描述了城市大数据平台的架构。

首先，根据数据结构及其关联的时空属性，我们将城市大数据分为 6 大类，如图 10 最下层所示。根据数据结构，我们将城市大数据分为点数据和网数据。其次，根据数据关联的时空属性，我们将城市大数据分为三类：第一类数据在空间和时间维度上都是静态的；第二类数据在空间维度是静态的，但在时间维度不停变化；第三类数据在时空维度都不停地变化。

点数据：一个加油站或商场是一个兴趣点。这个兴趣点一旦修建好，它的空间位置就不会随着时间而变化，它的各种属性（如面积大小、楼层高度等）也不会随时间而变化，因此，它是一个时空都静态的点数据。有些传感器的位置虽然不变，但它每个小时产生的读数（比如一个地方的温度）不断发生变化，因此，它是一个空间静态但时间动态的点数据。对于滴滴专车和摩拜单车，在不同的时间和地点我们会收到不同的用户请求，因此，它是一个时空都动态的点数据。

网数据：简单的路网是一个时空静态的网络结构数据。道路一旦修建完毕，它们的位置和属性

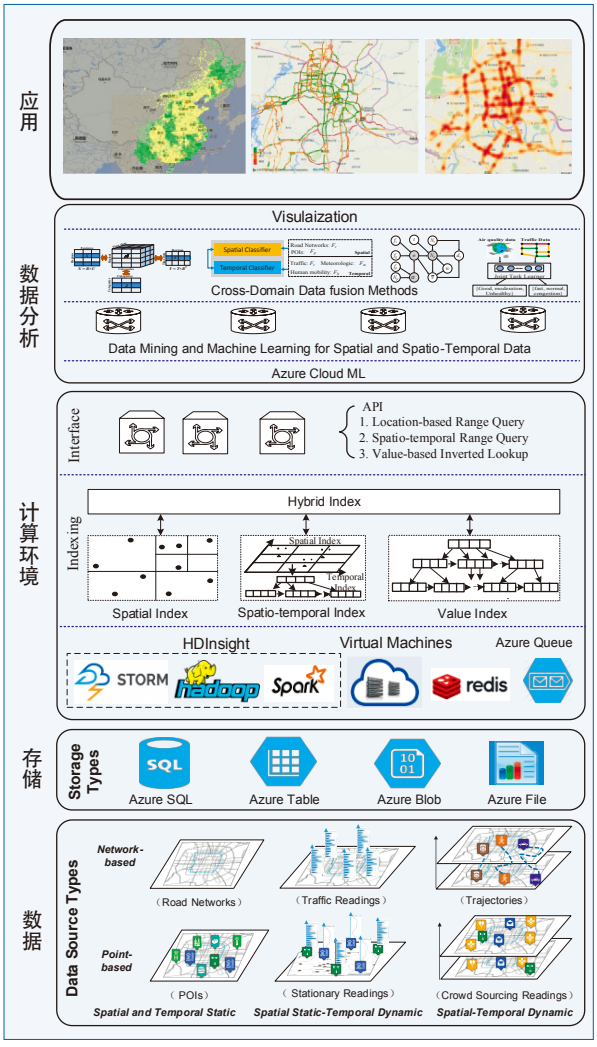


图10　城市大数据平台

就不会随时间变化。我们把时变的交通流量叠加到路网上，就形成了空间静态但时间动态的网数据。轨迹数据是时空都变化的网数据，例如一辆汽车在不同的时间处于不同的位置，车的油耗、速度和朝向等属性也在不停地变化。一个移动物体产生的轨迹点形成一个链式的网络结构。轨迹数据是城市大数据中结构最为复杂，且蕴含信息最为丰富的一种数据 [9]。

其次，我们可以借助已有云计算平台上的存储资源，如 Azure SQL, Table 和 Blob 等，来存储不同的数据模型。

第三，我们针对不同的数据模型，设计不同的

空间或时空索引算法。然后，把这些索引结构集成到分布式计算环境中，比如Hadoop, Spark和Storm。这样结合的优点是可以利用更少的计算资源来更快地处理更大的数据和更为复杂的任务，使得平台能够支撑上层的大规模高实时性的数据挖掘任务。

第四，我们需要建立城市数据分析层面。这个层面既包含了常用的机器学习工具（比如分类、聚类和回归等模型），还包含了针对时空数据专门定制的高级机器学习算法，以及多源数据融合算法[14]。

基于这样的城市大数据平台，我们可以组合各个层面的模块，快速地搭建各种垂直应用，在保证平台可扩展性的前提下，提供高效、稳定的服务。 ■

郑 宇

CCF杰出会员、CCCF编委、ADL工作组组长。微软亚洲研究院主任研究员，上海交通大学讲座教授、博导，香港科技大学客座教授。主要研究方向为基于位置的服务、时空数据挖掘、地理信息系统和普适计算等。

msyuzheng@outlook.com

参考文献

[1] Zheng Y, Capra L, Wolfson O, et al. Urban Computing: Concepts, Methodologies, and Applications[J]. *ACM Transactions on Intelligent Systems & Technology*, 2014, 5(3):1-55.

[2] Pan B, Zheng Y, Wilkie D, et al. Crowd sensing of traffic anomalies based on human mobility and social media[C]// *ACM Sigspatial International Conference on Advances in Geographic Information Systems*. ACM, 2013:344-353.

[3] Ji S, Zheng Y, Li T. Urban sensing based on human mobility[C]// *ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016:1040-1051.

[4] Shang J, Zheng Y, et al. Inferring gas consumption and pollution emission of vehicles throughout a city[C] // *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014:1027-1036.

[5] Zheng Y, Liu F, Hsieh H P. U-Air: when urban air quality inference meets big data[C]// *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2013:1436-1444.

[6] Yi X, Zheng Y, Zhang J, et al. ST-MVL: Filling Missing Values in Geo-sensory Time Series Data[C]// *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (IJCAI 2016).

[7] Li Y, Bao J, Li Y, et al. Mining the most influential k-location set from massive trajectories[C]// *ACM Sigspatial International Conference on Advances in Geographic Information Systems*. ACM, 2016:51.

[8] Hsieh H P, Lin S D, Zheng Y. Inferring Air Quality for Station Location Recommendation Based on Urban Big Data[C]// *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015:437-446.

[9] Zheng, Y. Trajectory data mining: an overview[J]. *ACM Transactions on Intelligent Systems and Technology* (TIST).2015; 6(3):29.

[10]Jie Bao, Ruiyuan Li, Xiuwen Yi, Yu Zheng. Managing Massive Trajectories on the Cloud. in Bao J, Li R, Yi X, et al. Managing massive trajectories on the cloud [C]// *Proceedings of the 24th ACM International Conference on Advances in Geographical Information Systems*.

[11]Li R, Ruan S, Bao J, et al. Querying Massive Trajectories by Path on the Cloud (Short Paper) [C]//*Proceedings of the 25th ACM International Conference on Advances in Geographical Information Systems*.

[12] Hoang M X, Zheng Y, Singh A K. FCCF: forecasting citywide crowd flows based on big data[C]//*ACM Sigspatial International Conference*. ACM, 2016:1-10.

[13]Zhang J, Zheng Y, Qi D. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction[C]// *Proceedings of the 31st AAAI Conference* (AAAI 2017).

[14]Zheng Y. Methodologies for cross-domain data fusion: An overview[J]. *IEEE Transactions on Big Data*, 2015;1(1):16-34.

[15]Zheng Y, Yi X, Li M, et al. August. Forecasting fine-grained air quality based on big data[C]//*Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM,2015: 2267-2276.

[16] Zhu J Y, Zhang C, Zhang H, et al. pg-Causality: Identifying Spatiotemporal Causal Pathways for Air Pollutants with Urban Big Data[J]. *IEEE Transactions on Big Data*, 2016, (99):1-1.