·科学论坛·

科学数据智能:人工智能在科学发现中的机遇与挑战

孟小峰*

中国人民大学 信息学院,北京 100872

[摘 要] 随着全球各科学领域大科学装置的出现,科学发现进入了大数据时代。科学发现无法完全依赖于专家经验从海量数据中发现稀有科学事件,大量历史数据无法有效利用,同时愈发突出实时性和高精度,科学事件的模式具有稀有性,通用的算法并不适用于科学领域,由此科学数据智能发现问题应运而生。科学数据智能发现旨在使用数据智能的方法加速科学事件的发现。然而,科学数据智能发现缺少整体框架设计,具体表现为缺乏科学数据的一体化分析体系和异构科学数据高效知识融合机制,并且海量历史数据长期存储及挖掘低效。本文从数据管理的角度提出科学数据智能发现与管理框架和相关挑战,以期推动科学发现的进步。

[关键词] 科学数据;数据智能;数据管理;智能发现;知识融合;长期存储

DOI:10.16262/j.cnki.1000-8217.20210611.006

科学数据是指人类在科学活动中,经由科学装置的不断发展而产生,通过实验、观测、探测、调查、挖掘等途径获取的用于研究活动的原始数据及衍生数据,这些积累的数据能够反映客观事物的本质、特征、变化规律。随着科学观测装置、观测技术的发展,科学数据已进入信息丰富的大数据时代[1]。天文学、遥感科学、高能物理学等领域都面临着科学数据激增,需要探索更加高效、智能的方法从大规模科学数据中发现有价值的科学事件。

科学事件的探索和发现往往具有时效性,以时域天文学为例,大视场短时标巡天以其阵列式观测覆盖组合大视场和高时间分辨率的数据采集,具备了高效发现短时标科学事件(持续时间较短的科学事件)的能力,但也对数据管理带来前所未有的挑战。大视场短时标巡天每天都以 TB 量级的速度快速采集数据,并形成大规模数据流,短时标科学事件就蕴含其中,但是短时标科学事件极其稀有且稍纵即逝,因此对分析的实时性要求很高,此外高噪声和伪事件又导致其真伪判断愈加困难[2-4]。不仅在于天文学领域,其他科学领域数据收集类似,都愈发强调实时性和高精度。

事实上,上述例子的挑战主要表现为“快、准、全”三方面。首先,大科学装置产生的大多为科学数据流,大规模流式处理和分析是必须的,其本质为“大”数据中发现“小”概率的科学事件,要求系统具备实时智能分析的能力[5]。其次,系统需要提供对科学事件快速验证的能力,因此不同的数据源的高精度融合和多尺度实体画像构建能够助力科学家做出准确判断,即整体发现不仅要“快”,更要“准”。最后,由于科学事件的稀有性,系统需要实现智能地自我更新,以不断提高整个系统的发现能力,因此,必须借助历史数据的高效分析以实现科学事件发现的“全”面[6]。

孟小峰 博士,中国人民大学教授,博士生导师,CCF 会士。主要研究方向为数据库理论与系统、大数据管理系统、大数据隐私保护、大数据融合与智能、大数据实时分析、社会计算等。

基于此,针对科学事件的发现目标,要解决大规模科学数据的智能发现问题,本质上是实现大规模科学数据的智能管理,本文从数据管理的角度来解决智能发现问题。

具体而言,大规模科学数据智能发现与管理主要面临着如下三方面的挑战:

收稿日期:2020-06-04;修回日期:2021-01-11

* 通信作者,Email:xfmeng@ruc.edu.cn

本文受到国家自然科学基金项目(61941121,91846204)的资助。

（1）实时智能的科学事件分析

实时智能的科学事件分析事实上主要面临数据处理和智能发现两方面问题。科学数据中的观测目标极多，即数据基数大，就要求报警率极低（可达十万分之一），才能保证科学家对报警的重视程度，因此不仅需要具备实时处理大规模科学数据的能力，同时需要具备高精度的智能发现能力。

（2）快速高效的科学事件验证

高效的科学事件验证主要解决的问题是对于科学事件报警信号的实时验证，以快速识别其价值。例如，在时域天文学中，天文学家的验证工作繁琐，虽然有集成的数据库平台可以使用，但这些数据库都只停留在数据的集成阶段，未能高效地从集成的数据库中抽取数据间的关系和知识并加以融合，也不能充分利用历史科学文献中积累的科学事件知识，导致验证工作困难[7]。

（3）大规模科学数据的长期存储

当前科学数据的收集效率越来越高，然而长期历史数据由于管理能力限制呈现出价值逐年递减的态势，如同矿业领域的煤矸石一样，不能被高效利用，影响了长期数据服务于提高系统发现能力的效率，因此，对长期历史数据的存储和分析是必须解决的问题。如何有效组织并以低成本解决大量历史数据的查询分析问题，使得能够从底层数据角度服务于智能分析和验证任务是科学数据面临的普遍问题。

1 科学数据智能发现与管理框架

前文所述的挑战如果得以解决，将为科学发现打开一扇崭新的窗口，极大地助力科学家对科学事件的发现工作。基于此，本文提出大规模科学数据智能发现与管理框架，如图 1 所示，包含智能分析层、知识融合层和数据存储层三个部分：

（1）科学事件的实时智能化分析：针对科学事件的实效性和特殊科学装置数据采集特点设计新的流数据处理框架适应科学数据要求的实时性能约束和处理模式，此外计算任务从数据和模型两个角度助力科学事件的高效智能分析。

（2）多尺度科学数据的全景化融合：科学观测不是单方面的观测，存在多个观测角度、观测装置、观测地点等，针对科学数据特有的多尺度、多源观测特性，采用知识融合及知识图谱技术实现不同科学数据源的交叉融合，构建海量科学事件观测目标知识图谱，加速科学事件验证。

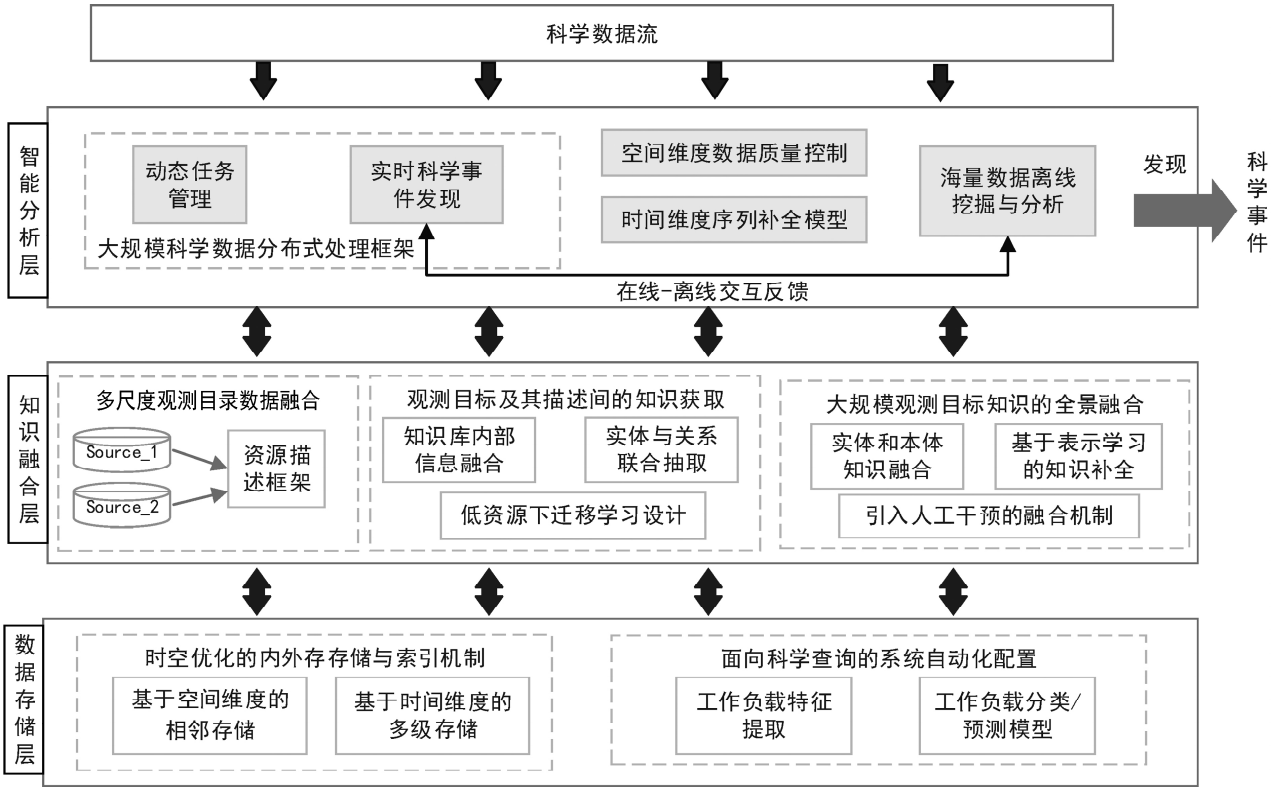

图 1 大规模科学数据智能发现与管理框架

（3）大规模科学数据的协同化存储：科学数据的长期存储主要解决高效查询问题。因此，首先从存储优化角度提高整体数据访问性能，科学数据通常具有时间和空间特性，可以通过时空优化的内外存协同存储与索引机制保证数据存储的合理性；其次从系统合理配置角度提高特定查询的效率，通过科学数据工作负载运行时的特征收集和分析，动态设置系统的最优化配置方案。

事实上，本文提出科学数据智能管理框架以科学事件智能分析、高效科学事件验证和大规模科学数据长期存储三大基础性关键技术作为支撑，三个关键技术作为一个有机整体共同助力科学事件的快速发现。

2　科学事件的实时智能化分析

本节主要聚焦于从处理框架和分析方法两个角度介绍科学数据流的智能化分析。

2.1　科学数据的分布式处理框架

科学数据流的形式是多样的，最终都可以归结为观测值的时间序列，但采集方式会有不同。对于元组采集方式而言，每个采集终端负责对一个目标或极小区域采集样本值，如海洋中的观测浮标收集温度湿度等，每次数据采集都是一个极小的数据元组。对于批量采集方式而言，观测单元对海量目标同时进行数据采集，如时域天文学中观测阵列对天体光度采集，每次都会形成海量目标的观测值的数据块，且这类数据块又不适宜拆分成元组处理，因为会损失块内邻域元组之间的关联特性。

针对以上特性，科学数据的分布式处理框架需要能够结合不同的领域知识动态适应不同采集方式。对于元组采集模式而言，处理框架需要使用非阻塞式元组处理模式或阻塞式微批处理模式[8, 9]，即 Apache Storm 和 Apache Spark streaming 采用的方式处理。对于批量采集方式而言，处理框架需要使用非阻塞式实时块数据处理模式，该处理不同于上述两种处理模式。由于块数据不能拆分元组处理又要保证块数据处理的实时性，因此处理框架必须兼顾块邻域关联特点的基础上动态对块数据分区进行分布式处理，且分区数据的处理要进一步有实时性保证。这就要求处理框架底层支持基于块数据分布式处理的实时约束技术。此外，还需要通过资源隔离的方式隔离不同的处理模式并保证它们有机地协同工作。

2.2　交互反馈的科学发现机制

科学发现中常用的方法是时间序列异常检测[10−13]，主要方法可分为：基于分类、基于聚类、基于统计学、基于信息论以及基于人工智能的异常检测等技术[24]等。而当下科学数据通常是以时序流形式呈现[14]，且异常发现模式不能够完全穷尽，导致传统的时间序列异常检测算法不能够胜任。

科学数据具有连续采集特性，因此科学发现可分为离线挖掘与在线分析两部分，从模型角度提高科学数据分析精度。离线层数据量大，使得离线数据训练的模型精度高，更能够涵盖数据的全局特征，但离线训练模型耗费时间长；实时层数据量少，实时层的模型训练要求快，但训练的模型精度低，只能够涵盖数据的最新特征（局部）。因此需要研究在线与离线交互分析反馈机制，用离线精度高的模型，支持实时的异常检测，从系统和算法两方面实现实时序列异常发现算法体系的演化，并实现离线分类模型自适应更新，图 2 为本文提出的实时—离线闭环反馈策略。

基于反馈机制的科学发现使得系统的离线层和实时层形成闭环，从而持续提高科学发现的精度，形成科学数据处理的工作流[15]。

2.3　数据质量控制与序列补全

科学数据作为一系列观测值容易受到外界环境干扰，会导致数据的畸变或缺失，因此考虑从数据角度提高分析精度。

对于典型的元组采集方式而言，目前有很多抗噪声的方法用于数据的质量控制，如小波变换等。但是对于批量采集方式而言，这类方式是不适用的。因为每次干扰都是局部空间相关的，如时域天文学

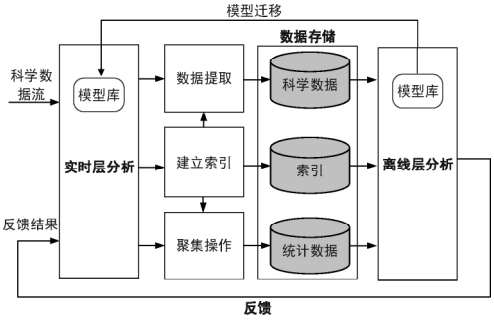

图 2　实时—离线闭环反馈机制

中云雾对天体光度的遮挡都是范围性的。因此，分析这种局部空间的相关性干扰是数据质量控制的核心。这要求质量控制技术需要区分不同数据采集方式，以动态适配。

无论是哪种数据采集方式，最终都是以时间序列形式呈现的，由于观测计划改变、设备故障等，常常导致时间序列残缺不全，缺失的数据比例之大导致已有的方法无法解决，严重影响了后续对观测数据的分析以及科学事件发现。针对时间序列的补全，主要有基于统计量的统计学、基于相关时间序列以及基于深度学习的方法[16, 17]。这些方法通常只能在离线层使用，要求相关序列非完全缺失，而且无法处理连续大量的缺失数据。

在科学发现的真实场景下，序列的缺失情况千差万别，不仅需要科学数据的实时补全方法，同时需要保证在缺失数据无法补全时的发现精度，结合反馈机制来不断完善补全算法，具有重要意义。

3 多尺度科学数据的全景化融合

在科学发现场景下，科学事件的验证往往需要借助多个数据源的数据对观测到的科学事件候选体进行统一化的多维度描述形成观测目标的多尺度画像，以辅助科学家更为清楚地验证候选体的真伪，同时能够对数据进行溯源[18]。为了对观测目标的知识进行可粒度缩放、可跨界关联、可全局视图的融合与管理[19, 20]，本节提出基于知识表示学习的全景式科学数据知识融合机制（图 3），帮助科学家实现智能验证，突破目前验证的高延迟瓶颈。

3.1 多尺度观测目标之间的数据融合

科学数据可以来源于不同的观测设备、观测地点、观测方式、观测顺序等，其数据形式可以是数据集、数据库、文本或文档等，数据格式可以是图像、文字等，因此其数据的多源异构特性非常明显。传统方法主要结合语义信息和多辅助信息来计算相似度，需要设计不同的学习模型来适应不同数据的特征，十分不便[21]。

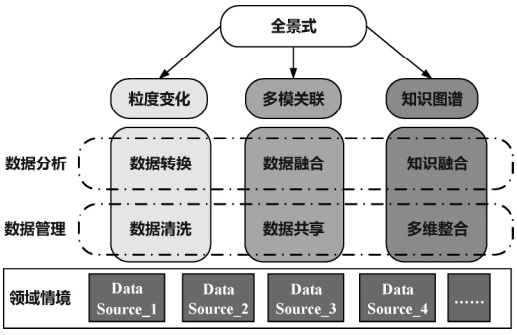

图 3 大数据驱动的“全景式”科学数据融合机制

针对科学数据的多源异构特性，需要实现多模态数据之间的表示转换，以便将不同的特征映射到统一的知识表示空间中[22]。该实体融合方法将来自于不同观测设备的、以不同格式存储的观测目标信息转化为统一的资源描述框架，并据此进行知识的对齐和消歧，具体使用基于知识图谱技术的有监督数据转化方法，通过参数共享、正则项添加等方式完成观测实体的融合。

3.2 观测目标及其描述间的知识获取

科学领域有着大量的科学数据库和本体库，与此同时还有海量的科学研究文献数据，关于科学发现和科学事件分析等科学论文可以从相关网站或数据库中自由获取，这使得抽取大量以文本形式存在的科学知识变为可能，而且这也为知识的更新和质量控制提供了保证。

因此在科学发现中，可以通过基于知识表示学习的科学知识获取方法，将科学文献中存在的科学知识进行挖掘和抽取，具体研究基于初始知识库和本体库的双向嵌入式学习，对实体和本体都进行嵌入式学习，以此增强从科学文献中提取实体和关系的效率，同时研究在低资源情境下基于迁移学习方法来把开放领域中的研究模型引入到科学文献中的知识发现过程中来。

3.3 大规模观测目标知识的全景融合

针对大规模观测目标的知识全景融合，旨在刻画大数据驱动的“全景式”科学数据知识图谱。这里提出将对齐的多源科学数据和获取的科学知识从概念层和实例层对齐后再次融合到一个全局视图的全景化知识图谱中[23]。

首先，需要在已有的科学数据上进行知识融合，需要对已有数据中的概念和实例进行对齐[24]；其次，基于上述两个研究基础，对从开源数据中获取的观测目标科学知识与已知的观测数据库进行再一次知识的对齐验证，同样需要从概念和实例两个层次来完成，由于需要较强的观测领域背景知识，也为了方便服务于科学工作者，利用众包技术或者交互设计技术将人工部分融入到集成过程中

来[25]，使得融合后的知识质量得到有效控制；最后，基于融合后的最终知识图谱设计链接预测方法，比如利用图嵌入式学习或表示学习方法进行标注缺失数据的标签预测，以便补全观测目标知识中的缺失或遗漏部分。

4　大规模科学数据的协同化存储

在科学领域观测产生的数据主要服务于实时智能的科学发现，但是随着数据源源不断到来，系统依然需要将数据进行长期存储，以提供智能分析层、数据融合层和上层科学家查询。由于科学场景的查询具有典型的时空局部性，因此，本节主要研究高效的科学数据存储框架和查询性能优化。

4.1　时空优化的多级存储架构

实时性和快速性是智能管理场景下科学数据长期存储的核心要求。而传统的长期科学数据的管理，主要研究目标是批式大数据管理系统，不能够满足智能管理的实时性和快速性。新的采样数据不断到来，系统不仅需要实时地处理和查询这些数据，而且需要持久化地保存历史数据，以便支持数据的全时态查询与分析。

针对科学数据的时间和空间特性，可以通过使用内存或高速存储设备实现内外存协同存储，并结合科学数据的时空相关性进行优化。通过时空优化的多级内外存协同存储与索引机制可以将不同时间段的数据合理存放以兼顾实时性和空间消耗，从而实现科学大数据快速持久化，图 4 即为内外存协同多级存储架构。

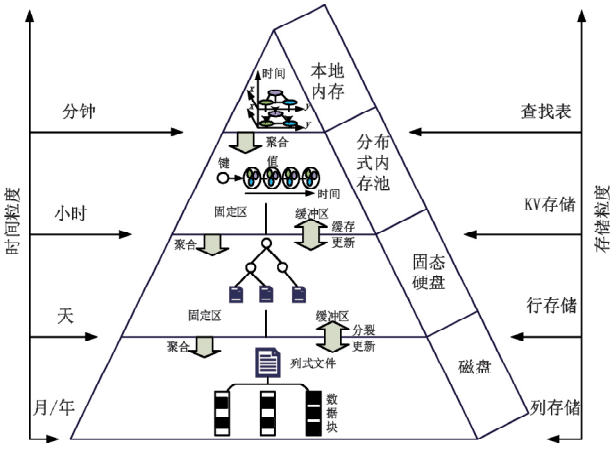

图 4　内外存协同多级存储架构

通过事先存储供聚集分析使用的粗粒度概要数据，并使用精度感知存储机制，在适当放宽查询的精度要求下访问近似或者部分概要数据来给出可以接受的结果，以提高查询分析性能。

4.2　运行时系统自动化配置

面向智能管理的科学查询需要实时性约束（特别是针对短期历史数据），由于观测周期的限制，每次发起的查询最好能在一次观测周期内完成，以确保查询结果能够用于下次数据处理，即查询延迟要小于给定的时间限制[26]。

由于科学查询是复杂多变的，对满足实时性的系统资源的要求也是不同的，若系统资源配置不合理，会严重影响系统查询的整体延迟。可以构建增量性能模型的方式预测查询延迟，运行时特征可以包括查询规模、查询算子、系统配置、资源使用等。最终通过预测的延迟选择合适任务配置与调优方法，保证在尽可能满足实时性的条件下资源消耗最小，最后快速实现新配置方案的部署。

5　总结与展望

科学数据进入信息丰富的大数据时代，其具有多样性和复杂性特点，目前的大数据分析方法主要依赖于常规的标准数据类型，缺乏科学数据一体化分析体系。此外，科学数据的统一表达、建模、操作计算方法明显欠缺，难以实现多维度、多尺度的科学数据知识融合与分析，使得科学家在科学事件验证时面临效率低、耗时久的瓶颈。科学大数据的长期存储和高效查询也是目前科学发现工作面临的重要问题。

要实现科学数据智能发现与管理由挑战到机遇的华丽转身，就需要提出新的发现与管理框架。本文从数据管理的角度提出科学数据发现与管理框架，将科学数据智能管理分解为智能分析、知识融合、数据存储三个层面，为大规模科学数据智能发现打开了新窗口，为科学领域的观测和科学事件的发现提供了新思路。

可预见的未来，大科学装置蓬勃发展，面向不同的科学目标产生的科学数据形态各异，需要的分析技术也是不尽相同的，如果都从零开始构造科学大数据分析系统，不仅研发动辄几年，而且耗费大量人力物力且不具备复用性。因此，对部件的复用显得至关重要。事实上，建筑领域中北宋李诫的《营造法

式》就提出了“凡构屋之制，皆以材为祖”的理念，元件“材”为基础的思想道出了中国古建筑的灵魂，即标准件、模数化和装配式，实现了营造效率、成本和建筑美观的内在平衡，这是古代匠人的永恒智慧。对科学大数据来说，是否存在一种“营造法式”，通过建设科学计算元件库，以实现大型复杂的科学分析系统能够像古建筑般高效构建且重复利用，“多快好省”地支持科学发现，这就是值得思考的重要方向之一。

参 考 文 献

[1] 黎建辉，沈志宏，孟小峰. 科学大数据管理：概念、技术与系统. 计算机研究与发展，2017，54(2)：235——247.

[2] Ivezic Z，Kahn SM，Tyson JA，et al. LSST：from science drivers to reference design and anticipated data products. The Astrophysical Journal，2019，873(2)：44.

[3] Yang C，Meng XF，Du ZH. Cloud based Real-Time and low latency scientific event analysis. Big Data，2018，498—507.

[4] Yang C，Meng X，Du Z，et al. Data Management in time-domain astronomy：requirements and challenges. BigSDM，2018，32—43.

[5] 孟小峰，慈祥. 大数据管理：概念、技术与挑战. 计算机研究与发展，2013，50 (1)：146—169.

[6] 杨晨，翁祖建，孟小峰，等. 天文大数据挑战与实时处理技术. 计算机研究与发展，2017，54 (2)：248—257.

[7] 孟小峰，杜治娟. 大数据融合研究：问题与挑战. 计算机研究与发展，2016，53 (2)：231—246.

[8] Wan M，Wu C，Wang J，et al. Column store for GWAC：a high-cadence，high-density，large-scale astronomical light curve pipeline and distributed shared-nothing database. Publications of the Astronomical Society of the Pacific，2016，128(969)：15.

[9] Medvedev D，Lemson G，Rippin M. SciServer compute：bringing analysis close to the data. Proceedings of the 2016 ACM International Conference on Scientific and Statistical Database Management，2016，27：1—4.

[10] Chandola V，Banerjee A，Kumar V. Anomaly detection：a survey. ACM Computing Surveys，2009，41(3)：1—58.

[11] Malhotra P，Vig L，Shroff G，et al. Long short term memory networks for anomaly detection in time series. // European Symposium on Artificial Neural Networks，2015.

[12] Movahedinia R，Chaharmir MR，Sebak AR，et al. Realization of large dielectric resonator antenna ESPAR. Ieee Transactions on Antennas and Propagation，2017，65 (7)：3744—3749.

[13] Ding D，Zhang M，Pan X，et al. Modeling extreme events in time series prediction. // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining，2019，1114—1122.

[14] Feng TZ，Du ZH，Sun YK，et al. Real-time anomaly detection of short Time-Scale GWAC survey light curves. // IEEE 6th International Congress on Big Data，2017，224—231.

[15] Deelman E，Gannon D，Shields M，et al. Workflows and e-Science：an overview of workflow system features and capabilities，2009，25(5)：528—540.

[16] Zhang YF，Thorburn PJ，Xiang W，et al. SSIM-A deep learning approach for recovering missing time series sensor data. IEEE Internet of Things Journal，2019，6(4)：6618—6628.

[17] Arous I，Khayati M，Cudre-Mauroux P，et al. RecovDB：accurate and efficient missing blocks recovery for large time series. // 2019 IEEE 35th International Conference on Data Engineering，2019，1976—1979.

[18] Simmhan YL，Plale B，Gannon D. A survey of data provenance in e-science，2005，34(3)：31—36.

[19] Dong X，Gabrilovich E，Heitz G，et al. Knowledge vault：a web-scale approach to probabilistic knowledge fusion. // Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining，2014，601—610.

[20] Dong XL，Srivastava D，Acm S. Knowledge curation and knowledge fusion：challenges，models，and applications// Proceedings of the 2015 Acm Sigmod International Conference on Management of Data，2015，2063—2066.

[21] 王雪鹏，刘康，何世柱，等. 基于网络语义标签的多源知识库实体对齐算法. 计算机学报，2017，40(3)：701—711.

[22] Kong C，Gao M，Xu C，et al. EnAli：entity alignment across multiple heterogeneous data sources. Frontiers of Computer Science，2019，13(1)：157—169.

[23] 王硕，杜志娟，孟小峰. 大规模知识图谱补全技术的研究进展. 中国科学：信息科学，2020，50(4)：551—575.

[24] Ren X, Wu ZQ, He WQ, et al. CoType: joint extraction of typed entities and relations with knowledge bases// Proceedings of the 26th International Conference on World Wide Web, 2017, 1015—1024.

[25] Doan A, Ardalan A, Ballard JR, et al. Human-in-the-Loop challenges for entity matching: a midterm report. ACM HILDA, 2017, 12:11—16.

[26] Wang CK, Meng XF, Guo Q, et al. Automating characterization deployment in distributed data stream management systems. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2669—2681.

Scientific Data Intelligence: AI for Scientific Discovery

Meng Xiaofeng*

School of Information, Renmin University of China, Beijing 100872

Abstract The large-scale scientific infrastructure has been accelerating all fields of science into Big Data Era. Although many interesting scientific events are contained in such a huge amount of data, it brings many a lot of trouble to scientists. Scientists can no longer rely on their experience to discover rare scientific events from massive data as they did before. The data intelligence technology is one of important topics to discover scientific events automatically. However, the key challenge is the lack of an intelligent discovery framework, involving the intelligent analysis methodology for scientific events, the intelligent verification mechanism for scientific events and the long-term storage architecture of scientific data. Based on this, we propose an intelligent management framework and its details from the view of data management to promote the intelligent scientific discovery.

Keywords scientific data; data intelligence; data management; intelligent discovery; knowledge fusion; long-term storage

（责任编辑 张 强）

* Corresponding Author, Email: xfmeng@ruc.edu.cn