《智能信息处理》课程考试

基于本体的地球科学知识图谱的构建

钱 程

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 9 日

基于本体的地球科学知识图谱的构建

钱程

（大连海事大学 信息科学技术学院，大连 116026）

摘　要：地球科学大数据为地球科学带来了研究范式的改革。但由于存在描述规范不统一、共享机制不明、语义异构等问题，在数据集成、共享与复用等方面存在较大困难，使得大数据的众多优势在地球科学相关研究中难以充分发挥。知识图谱的出现为地球科学研究带来了新的思路和方法。知识图谱能够准确、清晰地表达概念及其相互之间的复杂语义关系，为机器所理解，是实现语义翻译、数据融合和复用的关键技术。

本文对知识图谱和地球科学大数据进行了深入分析，探讨了知识图谱在解决地球科学大数据面临的挑战中发挥的重要作用，提出了分层架构地球科学知识图谱的思路，基于本体将地球科学知识图谱描述框架分为学科层、核心描述层、扩展描述层建立了知识描述模型，并以岩浆岩石学本体为例着重描述了核心描述层和扩展描述层的构建，基于分层架构知识图谱模型，建立了基于概念、属性以及相互之间关系约束的知识推理规则和推理机制，实现了基于知识图谱的异构数据的语义翻译，为数据发现和数据集成架设了桥梁，为大数据驱动下的数据挖掘提供了基础。

在此基础上，梳理了地球科学知识图谱建构工具的设计思路和需求，对地球科学知识图谱的逻辑结构进行了设计，分析了地球科学知识图谱建构工具具备的功能，利用实现的建构工具完成了岩浆岩石学知识图谱架构的初步构建，以期推动和完善地球科学知识图谱的建设和应用。

关键词：语义网;本体;本体学习;地球科学知识图谱

Construction of geo-science knowledge map based on ontology

Qian Cheng

(College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China)

Abstract: The big data of Earth Science Brings The reform of Research Paradigm for Earth Science. However, there are some difficulties in data integration, sharing and reuse due to the lack of uniform description, unclear sharing mechanism and semantic heterogeneity, it is difficult to make full use of the advantages of big data in the related research of Earth Science. The Appearance Of Knowledge Atlas brings new ideas and methods to the Earth Science Research. Knowledge Graph can express concepts and their complicated semantic relations accurately and clearly, and is understood by machines. It is the key technology to realize semantic translation, data fusion and Reuse. In this paper, the knowledge map and the big data of earth science are deeply analyzed, and the important role of knowledge map in solving the challenge of the big data of Earth Science is discussed, based on ontology, the description framework of geo-science knowledge map is divided into disciplinary layer, core description layer and extended description layer, and knowledge description model is established, taking the ontology of Magmatic Petrology as an example, the construction of the core description layer and the extended description layer is described, the rules and mechanism of knowledge reasoning based on concepts, attributes and relationship constraints are established, the semantic translation of heterogeneous data based on knowledge graph is realized, and the bridge is built for data discovery and data integration, it provides a foundation for data mining driven by big data. On this basis, the design ideas and requirements of the tool for constructing knowledge atlas of Earth Science are sorted out, the logical structure of Knowledge Atlas of Earth Science is designed, and the functions of the tool for constructing knowledge atlas of Earth Science are analyzed, the preliminary construction of the knowledge map structure of magmatic petrology is completed by using

the constructed tools, in order to promote and perfect the construction and application of the knowledge map of earth science.
Key words: Semantic Web; ontology; ontology learning; Atlas of Earth Science Knowledge

1 引言

知识图谱(Knowledge Graph)是一种知识描述框架，用来进行知识的分享和利用，是人工智能和大数据等相关研究的重要基础。“大数据”的概念最早由 Alvin Toffler 在其 1980 出版的《The Third Wave》一书中提出（Toffler，1981）。网络信息技术的迅速发展以及信息收集设备的不断优化，带来了半结构化、非结构化数据的爆发式增长，数据的产生已经脱离了时空的限制，“大数据时代来临”。目前对“大数据”普遍认可的定义是指在可容忍的时间范围内无法用传统的信息技术和软硬件工具对其进行感知、获取、传输、存储、管理、处理与应用的数据集合，具有 Volume(体量大)、Velocity(速度快)、Variety(模态多)、Veracity(真伪难辨)和 Value(价值高密度低)的 SV 特性。

知识图谱的出现为应对大数据的挑战提供了新的思路和方法。知识图谱并不是一个新兴的领域，其最初的概念和雏形可追溯到二十世纪六十年代，当时语义网络（Semantic Networks）作为知识表示的一种方法被提出(Simmons，1966; Quillian，1968，在语义网络中网格的节点表示概念或实体，网格的边缘线表示概念或实体之间的关系，语义网络主要应用于自然语言理解系统中。

综上，为促进地球科学数据和知识的整合与共享，解决数据和知识建模存在的问题，推动地球科学研究范式的变革，有必要引入知识图谱的概念和思路，建立地球科学知识图谱，消除地球科学数据的语义异构瓶颈，充分挖掘地球科学数据的价值，推动大数据驱动下的知识发现和知识服务，真正实现数据共享、复用和融合，深化地球科学基础研究和应用研究的发展。目前地球科学知识图谱领域仍存在空缺，存在着较大的研究空间，值得地学研究人员开展全面且深入的研究。

2 知识图谱概述

由于地球科学大数据缺乏统一的数据描述标准、数据共享机制不明、语义异构问题显著因此阻碍了大规模数据的集成、共享和互换，限制了大数据驱动下的地球科学研究新发展。此外，还有大量的数据以表格和图片等形式发表于地球科学期刊论文中题数据集，但是由于这类数据需要从以自然语言表达的论文中进行提取也形成了庞大的专无法直接使用，因此在数据共享、集成和复用方面存在较大困难。知识图谱则是通过信息抽取技术将多源异构数据中的知识抽取出来，并利用机器可理解的语言对知识概念和关系进行语义描述，建立知识模型，基于该模型进行知识融合和知识推理，为异构数据语义翻译提供基础，为大数据驱动下的全球数据复用、融合和共享提供服务。

2.1 知识图谱的概念

Google 公司于 2012 年提出知识图谱的概念（Singhal，2012），旨在增强 Google 的搜索引擎功能，增强用户搜索的质量和体验。从本质上来说，知识图谱是一种描述实体及实体之间关系的结构化的语义网络，主要由一个个三元组构成，表现形式有<实体 1，关系，实体 2>和<实体，属性，属性值>实体指的是具有可区别性且独立存在的事物，如某一个国家、某一个人、某一种动物等;概念主要指集合、类别、对象类型、事物的种类等，例如植物、岩石、变质作用等，知识图谱旨在将这些存在于物理世界中的知识抽取出来进行形式化的描述，形成可以被人和机器理解的大规模知识库，为全球数据的集成、共享和重用提供服务。

2.2 知识图谱的概念

知识图谱以准确清晰的知识表达在计算机世界中构建了表示物理世界中信息和知识的有效载体。

（1）准确、清晰的知识表达

知识图谱包含对知识及其相互关系的全面、清晰、明确的描述，而且采用国际标准化的编码对知

识及其相互关系进行形式化的表达，具有科学性、系统性和规范性，提供与其他描述规范进行互操作的基础。知识图谱的架构具有开放性的特点，便于修改和扩充，能够在不同层面上满足对知识的需求，是进行科普、教学和科研的知识库，是科学家进行学术交流的通用语言和基石，更是计算机可理解的数字化、结构化的知识库。

（2）丰富的语义表达能力

知识图谱充分表达了知识点之间的对等关系（例如同义词等）、包含关系（或称为从属关系)、继承关系、实例关系和属性归属关系等丰富的语义信息，可非常清晰和方便的表示成层次化或网状化的知识库。

（3）语义推理能力

除了充分表达知识点之间丰富的原生语义关系以外，知识图谱还具有强大的推理能力，能够从原生知识关联中通过推理产生新的知识，即可将隐性知识显性化，从而为数据挖掘和知识发现提供语义推理服务。

2.3 知识图谱的逻辑结构

从逻辑上可以将知识图谱划分为模式层和数据层两个层次，知识图谱的数据层主要由一个个实体组成，从多源异构数据中抽取的知识则是以实体为单位进行存储。模式层构建于数据层之上，是知识图谱的建设核心。模式层主要通过本体对知识进行统一规范化管理，利用本体对概念即相互之间的关系、属性及其数据类型及取值进行约束和说明，对知识图谱的层次结构进行标准规划化建设，减少冗余。本体作为知识图谱的概念模型，主要强调对概念及概念间的关系进行概括性、抽象性的描述，因此本体不包含过多的实例;知识图谱则是着重于描述实体关系，在数据层对本体内包含的知识进行完善和扩展。

2.4 知识图谱的构建方式

目前知识图谱的构建方式主要有自顶向下和自底向上两种。自顶向下构建方式是指先确定知识图谱的整体框架，然后从粗到细逐渐建立健全知识图谱各分支实体或者概念规范化描述及其相互关系，最终形成完整的知识图谱。自底向上构建方式是指将一个个概念或实体整理出来，分析相互之间的关系，建立具有一定结构的知识模型，最终形成完整的知识图谱。无论是自顶向下还是自底向上的建构方法都可以采用人工方式和计算机自动提取方式相结合来完成。

2.5 知识图谱的建设周期

知识图谱的构建经历了知识获取(Knowledge acquisition、知识建模(Knowledge modeling)、知识融合(Knowledge f}aion)、知识存储(Knowledge storage)、和知识推理(Knowledge reasoning)这五个步骤(见图 2-1) 。

图 2-1 知识图谱的建设周期

2.5.1 知识获取

随着互联网的飞速发展，大量的知识存在于结构化的数据库、半结构化的网页以及非结构化的文本中，知识获取的目的就是将存在于多源异构数据中的概念/实体及相互之间的关系、属性等知识要素提取出来，为知识图谱的构建提供基础。

2.5.2 知识建模

知识建模的核心是构建本体对知识进行规范化、形式化的描述，将知识利用机器可理解的语言进行表达，方便知识图谱的构建与完善。

（1）本体定义与分类

本体(Ontology)起源于哲学领域，是形而上学理论研究的一个分支，主要是对客观世界的事物进行解释和概括，其本质是客观现实的抽象和映射。

本体按照领域依赖程度可分为四大类：顶层(top-level)本体，对物理世界中所有的概念进行描述，不侧重于某一领域；领域(domain)本体，主要是对某一特定领域中的概念模型进行说明；任务(task)本体，对某一特定的任务中的概念模型进行描述；应用(application)本体，主要针对某一任务的概

念模型进行说明。

（2）本体描述语言

本体描述语言可将客观世界中的概念和概念之间的关系描述成计算机可理解的内容，相当于客观世界与计算机交流的媒介，目前应用最广且为W3C（World Wide Web Consortium 万维网联盟）所推荐的有 XML、 RDF(S)、OWL、SKOS。

2.5.3 知识融合

知识融合旨在将信息抽取获得的实体、关系以及属性进行清理和整合，消除概念之间存在的歧义，删除冗余错误概念，提高知识的层次性和逻辑性，从而保证知识图谱的质量。知识融合主要包含实体对齐和知识合并两部分内容。实体对齐主要是通过实体链接技术，将本体中的实体与非结构化文本中的实体建立链接，确定多个实体是否表示现实世界中的同一对象，消除实体间存在的歧义。

2.5.4 知识存储

知识存储的任务就是将知识图谱存储起来，并要保证后期进行知识查询、添加、删除、修改等操作的效率。知识图谱存储的基本单位是知识三元组，目前知识图谱主要以表结构和图结构进行存储。表结构存储主要是将知识三元组保存在二维表中，操作方便简单，但是缺点也是显而易见的，若存储的知识图谱规模过大，当对知识进行查询、添加、删除、修改等操作时，需耗费大量精力，影响知识图谱的应用。图结构的知识存储在表结构基础上进行了性能优化，因此目前知识图谱大多以图的形式对实体以及实体之间的关系进行存储，实体以圆形节点进行展示，用一条边来表示实体之间的关系，使得知识图谱以网状图结构进行展示，该方式不仅实现知识图谱的可视化，扩大知识图谱的存储规模，还提高了增删改查操作的效率，增强了知识图谱的实用性，但是图结构存储还存在数据更新速度慢、对于复杂查询的支持力度不足等缺点。

3 地球科学知识图谱

3.1 构建思路

地球科学包含岩石学、矿物学、矿床学、古生物学、地层学、地质年代学、构造地质学等众多学科门类。每个学科的知识结构不同，知识点及其相互之间的关系有共性也有特性，要整合到统一的知识图谱描述框架下具有相当的难度，必须充分考虑以下两个基本条件:（1）既能体现不同学科知识结构的共性特征，同时又能够体现各个学科的特点；（2）具有开放性和可扩展性，允许通过不同方式或途径建立的地球科学分支的知识图谱整合到一起，不断补充和完善地球科学知识图谱。

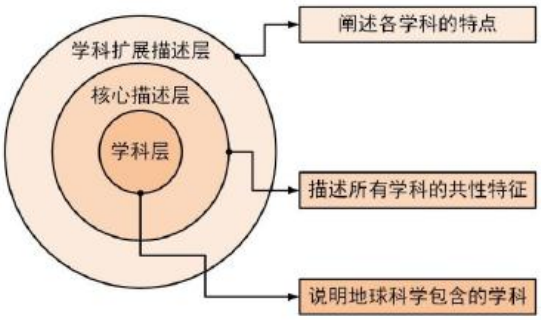

图 3-1 地球科学知识图谱描述框架图

为实现这一目标，我们采用了自顶向下、分层设计的思路来构建知识图谱(见图 3-1)。

整个地球科学知识图谱自顶向下共分三个层级:第一层级为学科层，第二层级为核心描述层，第三层级为学科特征描述层。学科层对地球科学知识图谱涵盖的学科进行整体描述，主要体现地球科学知识图谱包含哪些具体学科。核心描述层旨在描述每个学科的知识框架，重点在于体现不同知识点之间的层级关系，提炼所有学科知识点的共性特征组成核心描述集，涵盖每个学科的所有知识点。通过对这些知识点之间相互关系的描述，建立起每个学科的知识描述框架。每个学科的描述层可在统一的规范标准下由不同途径建立后整合到整体框架中。学科扩展描述层则可根据不同学科的特点自定义描述集，对知识点的特征进行详细描述，体现学科特色。

4 总结

本文分析了地球科学知识图谱的建构特点和需求，提出了自顶向下、分层设计的建设思路，利用规范化的本体描述语言对地球科学知识图谱描述框架进行了描述。在此基础上根据本体的关系约束探讨了基于地球科学知识图谱的数据发现与集成机制。基于地球科学知识图谱建构和应用需求对地球科学知识图谱建构工具进行了分析，并对其功能进行了设计。主要研究内容和取得的主要成果表

述如下:

（1）对目前地球科学大数据在描述标准、共享机制、语义翻译等方面存在的问题进行了分析，阐明了知识图谱在大数据驱动的地球科学研究中发挥的重要作用。在此基础上提出了分层描述的知识框架模型，设计了统一知识描述框架，对不同学科知识结构的共性特征以及各分支的特有性质进行了分级描述，既满足地球科学各学科普遍描述需求，又能体现各学科特定描述要求，具有较强的开放性、灵活性和可扩展性。

（2）建立了基于描述框架中关系约束的知识推理规则和推理机制。通过知识推理，将知识图谱中的概念与多源数据中的实体建立映射关系，消除了多源异构数据之间的语义异构，实现了基于知识图谱的语义翻译，为数据发现和数据集成架设了桥梁，为大数据驱动下的数据挖掘提供了基础。

（3）为满足地球科学知识图谱建设需求，对地球科学知识图谱建构工具进行了需求分析和功能设计，并利用实现的建构工具完成了岩浆岩石学知识图谱架构的初步构建，验证了设计思路的有效性和可行性。

参 考 文 献

[1] 林龙成.语义网中 OWL 本体概述及其构建方法研究[J].电脑知识与技术,2020,16(12):203-204.

[2] Jian-Jun Qi. Attribute reduction in formal contexts based on a new discernibility matrix. Journal of Applied Mathematics and Computing. 2009, 30(1-2): 305-314.

[3] 李京杰.基于语义本体的个性化学习推荐研究[J].软件导刊(教育技术),2016,15(09):77-78.

[4] 余霞,刘强,叶丹.基于规则的关系数据库到本体的转换方法[J].计算机应用研究,2008(3):767 -770,785.

[5] Fortuna B , Grobelnik M, Mladenic D. OntoGen :Semi - automati c Ontology [A] / / Smith M J, Salvendy G (Eds.). Human Interface[C],Part II , HCII 2007 , LNCS 4558 , 2007:309 -318 .

[6] 刘小乐,马捷.语义网环境下基于本体的知识集成研究进展[J].现代情报,2015,35(01):159-163+169.

[7]]Hearst M A . Automated Discovery of WordNet Relations[A] Christiane F . WordNet：An Electronic Lexical Database[M]，MIT Press，1992：132—152.

[8] 杜小勇,李曼,王珊.本体学习研究综述[J].软件学报,2006,17(9),1837-1847 .

[9] Deng ZH，Tang SW，Zhang M，Yang DQ，Chen J．Overview of ontology． Acta Scientiarum Naturalium Universitatis Pekinensis，2002，38(5)：730-738(in Chinese with English abstract