地球表层系统数据目录关联网络构建研究及展望

邱芹军[1,2]，郝孟堪[1,2]，谢忠[1,2,*]，陶留锋，李伟杰[1,2]，王洋[1,2]，刘建东[1,2]

1. 中国地质大学（武汉）计算机学院，武汉 430074
2. 地理信息系统国家地方联合工程实验室，武汉 430074
3. 中国地质大学（武汉）地质探测与评估教育部重点实验室，武汉 430074

摘要：针对地球表层系统（以下简称地表系统）开放数据分散、多源、异构、多模态的问题，关联网络作为一种强大的知识表示方式，它通过将元数据作为节点，将元数据之间的关系作为边，以节点间关联性的强弱作为边的值，构建了元数据之间可以被计算机理解的结构化、语义化的网络。关联网络可以为地表数据的开放和共享提供完整的解决方案。本文介绍了地表关联网络的发展现状、构建方法和构建内容设计。首先，通过评估和筛选研究对象，筛选并确定了国内外典型关联网络作为研究对象。然后，通过综合分析和评估现有文献和典型关联网络，从顶层构建方法和具体内容实现两方面分别进行比较分析。在构建方法方面，分析了关联网络的数据来源、构建方式、自动化程度和更新方式。在内容设计分析方面，介绍了关联指标的选择，同时，详细讨论了空间特征、时间特征和内容特征提取、表示和计算的方法。此外提出了对未来地表关联网络的启示：从构建方法上，应提高数据收集质量、发展地表数据目录共享程度分析方法；从构建内容设计上，应完善特征提取及表示方法、完善地表特征关联及计算方法；从应用上，应共同推动应用案例和积攒实践经验，提升地表关联网络的应用成效。

关键词：地表关联网络，数据目录，关联网络构建，关联指标，特征关联

Abstract:

Keywords:

1 引言

地球表层系统（以下简称地表系统）是地球各圈层交互作用和人类活动最为活跃的区域，其数据覆盖面广、类型丰富、变化快速、数据海量且开发潜力巨大。地球表层系统是由大气圈、水圈（含冰冻圈）、生物圈、土壤圈和人类圈所构成的地表自然社会综合体[1]，是人类圈与地圈相互作用的复合物质系统，与周围的地球圈层其他部分存在物质能量交换关系[2]，是一个开放的复杂次级巨系统。地表系统是地球各圈层交互作用和人类活动最为活跃的区域，其数据覆盖面广、类型丰富、变化快速、数据海量且开发潜力巨大。围绕地表系统领域的科

学数据管理与服务是当前的全球前沿和热点。

地球表层系统开放数据包括地理空间数据、环境数据、地质数据、气象数据、遥感数据、地球观测数据等[3]，涵盖了地球表层系统的地理、环境、气候、地质等多个方面的数据集[2]。这些数据集通过开放获取、共享和使用，可以极大地推动地球科学研究，地球表层开放数据呈现以下几个特征：一、数据来源丰富，随着遥感技术、地面观测网络和地球科学研究的发展，地球表层开放科学数据来源越来越丰富；二、开放数据平台建设，为了更好地共享和利用地球表层开放科学数据，各国纷纷建立了开放数据平台。例如，美国地质调查局（USGS）的地球资源观测与科学数据中心（EROS）[4]、欧洲空间局（ESA）[5]的地球观测网站等。这些平台为研究人员提供了方便的数据获取途径；三、数据标准化和互操作性，地球表层开放科学数据的标准化和互操作性是数据共享和利用的关键。为实现这个目标，相关机构制定了一系列数据标准和规范[6]，如 OGC（开放地理空间联盟）的地理信息标准[7]，以及 ISO[8]（国际标准化组织）的地球观测数据和服务标准[9]。四、数据处理和分析方法创新，随着地球表层开放科学数据量的不断增加，数据处理和分析方法也在不断创新。例如，人工智能和机器学习技术的应用，可以帮助研究人员更高效地处理和分析大量数据，从而获得更深入的科学认识。以上四点表明，地球表层开放科学数据正迅速发展。

然而，由于科学数据分散在不同的存储库、平台和系统中，不同的系统采用不同的数据格式和类型，提供的元数据信息也略有不同，阻碍了科学数据的进一步关联、集成和集成，也阻碍了科学数据价值的充分发挥和科研效益的最大化[10]。如何提取这些分散的存储库中的元数据信息并其形成统一的数据目录，深入剖析地球表层系统科学数据目录专题内容、时间、空间等本质属性，以及数据格式、类型结构、坐标基准等形态特征，选取用于关联的数据特征，通过被选特征之间的综合语义关系，研究建立多维、定量的地球表层系统科学数据综合关联模型；结合最新的表示学习方法、权重计算理论、专家打分，研究顾及地球表层系统科学数据目录主题内容、空间拓扑、空间精度、时间拓扑、时间粒度、数据类型、数据格式等[11]单一相似度度量方法，构建数据语义相关度计算的线性模型；在此基础上，通过地球表层系统科学数据目录语义关系和相应的语义相关度值，开展地球表层系统科学数据目录的关联共享，为国家科学数据中心平台智能化、精细化的数据推荐和语义检索等提供支撑服务，成为亟待解决的问题。

关联网络作为一种强大的知识表示方式，可以为上述问题提供完整的解决方案。关联网络以 RDF（Resource Description Framework）[12]为基本单元，元数据为节点，以元数据之间的关系为边，以节点间关联性的强弱为边的值，构建了数据与数据之间能被计算机理解的结

构化、语义化的链接[13,14]。关联网络中的节点和边可以组成复杂的网络结构，这些结构可以自动识别和推断概念之间的关系，从而实现语义搜索和自然语言处理等任务。关联网络还具有自适应性，它可以根据经验学习和自我调整，从而提高知识表示的准确性和可靠性。现有的关联网络主要从构建方法和研究应用两个方面开始。在构建方法方面，大多是从本体构建[15]、数据（目录）挖掘[16]和关联网络构建（包括关联指标获取[17]，特征表示[18]，特征关联，特征计算[19]等步骤）进行研究。在应用方面，国内和国外已经有很多组织或科学团队进行了研究，形成了许多关联网络，这些关联网络可以分为通用型和领域型[20]，领域型关联网络是指针对特定领域中的特定对象或实体构建的关联网络。例如地理空间元数据关联网络[18]和极地科学数据关联网络[17]等研究。但这些研究多是科研人员针对特定场景形成，更新和维护时效较短，无法对地表系统关联网络的发展形成较大的推动。

相较而言，通用型关联网络，例如 Linked GeoData、GeoSciML (Geoscience Markup Language)、OSM Semantic Network、GeoWordNet 等。它们有以下优点：通常由大型商业团队进行开发维护，所以具有长期的维护和稳定的更新频率，同时覆盖范围更广，能够提供多方面的地球科学研究和应用服务；数据量大，能够提供更加全面和精细的地球科学信息；语言通用强，节点和边的定义采用通用的国际标准，使得通用地表关联网络在不同地表领域之间具有较强的语言通用性；应用广泛且构建方法多样。因此通用关联网络对地表系统关联网络有较大推进作用。但是对于地球表层系统这样大量大规模的数据集，如何获取它们的元数据，使用哪些特征来精准表示元数据，如何将元数据的不同特征进行关联和计算，目前还缺乏相关研究。

2 地球表层系统科学数据目录关联网络

2.1 地球表层系统科学数据目录

目前海量地球表层系统科学数据分布在不同数据网站中[21]，并随着科学研究发展而不断扩展。因此，需要构建地球表层系统科学数据目录，数据目录是一个记录和组织数据资源的清单或索引[22]，通常用于帮助用户查找和访问所需的数据资源。数据目录可以包含多种类型的信息，如数据资源的名称、描述、关键字、数据提供者、数据格式、数据访问链接等[23]。其中数据目录主要存储的就是元数据，元数据提供了对数据的描述信息[24]，而数据目录则用于记录和组织这些元数据。

地表系统数据目录相应的就是一个存储和管理地表系统各个圈层数据信息的文件夹或目录，它主要由数据集的描述信息也就是元数据构成，元数据包括地表系统数据集的核心要素如地理位置、关键词、时间信息等。

2.2 地球表层系统科学数据目录的体系结构

地球表层系统科学数据目录的体系结构，首先需要知道数据目录的核心要素都包括什么，根据专家经验总结得到数据目录需要包含数据集名、数据格式、空间范围、时间要素、比例尺、空间坐标参考系、关键词、数据质量、数据量、数据分类、数据快照、数据生产者、数据可访问性、数据地址、访问参数、访问方式、源元数据、论文引用、数据网站、数据网页快照、发现日期等多方面的信息。依据上述 21 条数据目录核心要素，可以得到地表系统科学数据目录的体系结构（如图 1 所示）包括：元数据汇聚获取、元数据清洗处理、元数据标准化和分类编目，通过构建开放数据目录挖掘模型实现多模态科学数据目录精准及高效挖掘。主要包括：①元数据采集汇聚，根据设定的数据引接、爬取规则，实现地表系统开放源数据常态汇聚接入；②元数据清洗处理，删除重复信息、纠正存在的错误，并提供一致性数据检测；③元数据标准化处理；④元数据分类编目，在地球表层系统开放科学数据本体库支持下，研发基于大模型的异构元数据标准化处理方法，实现原始元数据的自动规范化和标准化处理。最终将其分类编目形成数据目录。

数据目录形成后，经过关联指标获取、数据目录语义特征提取及表示、数据目录语义特征关联和数据目录语义特征计算等一系列关联网络构建步骤后最终形成大规模地球表层数据目录关联网络。

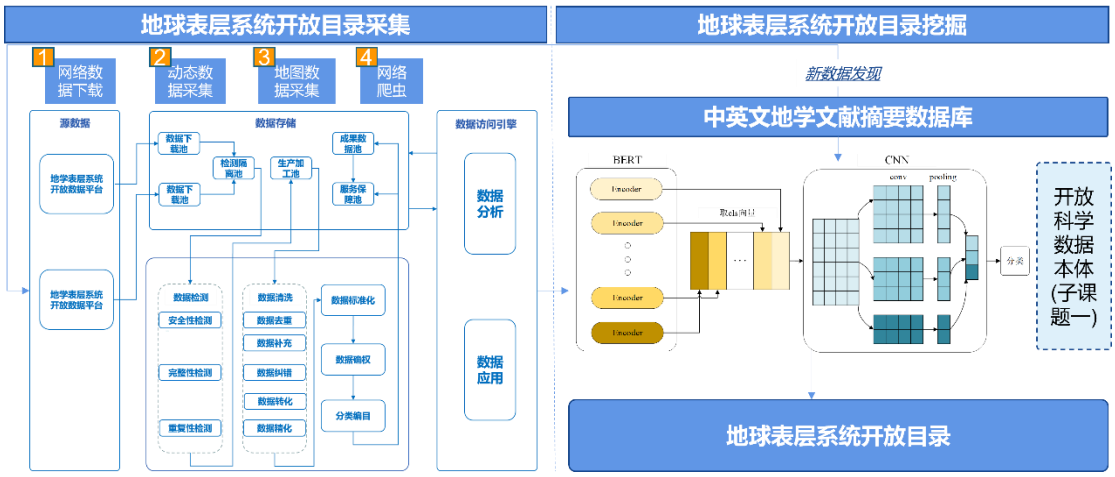

图 1 地球表层系统科学数据目录体系结构

3 材料与方法

本文采用软件工程领域通用的文献综述方法[25]对地球表层系统数据目录关联网络进行比较分析，共包括三个步骤。（1）研究对象评估和筛选，为了更注重实践和工程应用，本文同时选取国内外典型关联网络作为研究对象。（2）研究对象确定，由于关联网络参差不齐，有许多规模较小或已经停止维护更新，需要设计相关标准对关联网络进行进一步筛选与确定。（3）现状比较与分析，通过综合分析文献和评估现有关联网络，从顶层构建方法和具体内容实现两方面对其进行比较分析，并提出未来研究方向和建议。

3.1 研究对象初步筛选

通过与专家沟通和大量查阅相关文献，对现有关联网络进行筛选，最终得到了十六个关联网络，包含领域关联网络：GeoSciML (Geoscience Markup Language)(http://geosciml.org/)、OSM Semantic Network(https://wiki.openstreetmap.org/)、Geospatial Data Cloud(https://www.gscloud.cn/)、LinkedGeoData (http://linkedgeodata.org/)、SWEET(https://github.com/ESIPFed/sweet)、GeoWordNet(https://old.datahub.io/dataset/geowordnet)、Data.gov.uk(Data.gov.uk)和 OEG(https://www.oeg.net/en)；通用关联网络有：DEU (https://data.europa.eu/en)、DBpedia(https://www.dbpedia.org/)、YAGO(http://openkg.cn/dataset/yago)、Wikidata(https://www.wikidata.org)、ConceptNet(https://conceptnet.io/)、VIVO(http://vivo.library.cornell.edu)、XLore(https://www.xlore.cn)和 Knowledge Vault(https://knowledgevault.online)。这里选取了通用关联网络和领域关联网络，主要有以下原因；由于通用关联网络发展较早、更为成熟，因此通过与通用关联网络的比较，可以在多个方面对领域关联网络的构建及应用有所启发。通用关联网络在构建过程中采用了大量的自动化方法，如特征提取、特征表示和特征计算等，这些方法可以为领域型关联网络的自动构建关键技术研发提供借鉴。此外，通用关联网络中时空信息的表示也可以为地表时空信息的表示提供启示。另外，通用关联网络已经被广泛应用，这可以在一定程度上启发未来地表关联网络的应用推广。为了研究对象的进一步确定，需要从相关关联网络网站或维基百科收集上述关联网络的具体信息，包括：发布网址、定位目标、创建时间、定位目标、创建者、创建国家、网络规模、是否开源等，搜集结果如表 1 所示。

表 1 初步筛选的关联网络基本情况

关联网络	发布网址	定位目标	创建时间	创建者及所在国家	数据规模	是否开源	最新版本

DEU	https://data.europa.eu/en	欧洲联盟的官方开放数据平台	2011	European Union	1 608 830 欧洲公共部门数据集	是	实时
Linked GeoData	http://linkedgeodata.org/	以 OpenStreetMap 为数据源，创建大型知识库。	2009	莱比锡大学，德国	超过 30 亿个节点和 3 亿条边，约 200 亿个三元组。	是	2016
OSM Semantic Network	https://www.openstreetmap.org/	OSM 中的实体包括地理位置、地名等，这些实体通过空间位置、属性特征等相互关联。用户可以通过该语义网络实现地理信息检索	2004	Marc Wick 创立 瑞士	它包含超过 2500 万个地名和超过 1100 万个独特的特征	是	
OEG	http://geo.linkeddata.es/	它提供了用于管理、发布、检索和使用地理信息资源的开放式平台	2015	Longley, P.A 等人		是	
Data.gov.uk	http://data.gov.uk/linked-data/	提供了从各个政府部门和机构收集的各种数据集，这些数据集可以免费下载和使用。	2010	英国政府	超过 40 万个数据集	是	1.0
Geospatial Data Cloud	http://www.cgsdata.org.cn/	汇聚全国各地的各类地质信息	2012	中国	收录数据约 6000 余种，数据总量超过 110PB	半	1.0
GeoWordNet	https://old.datahub.io/dataset/geowordnet	它基于 WordNet 架构，通过对地表术语进行分类和关联来实现知识表示和信息检索。	2007	美国	包含超过 8 万个概念，这些概念覆盖了地球科学的各个领域	是	2.0
DBpedia	https://www.dbpedia.org/	以 Wikipedia 为信息源，从中提取结构化的数据并构建关联网络。	2007	莱比锡大学等，德国	1219 个本体，2.2 亿个实体，14.5 亿个三元组。	是	Largest Diamond
YAGO	https://yago-knowledge.org/	创建包含人、城市、国家和组织等通用性数据的关联网络。	2006	Max Planck 信息学研究所，德国	超过 5000 万实体，20 亿事实。	是	4
Knowledge Vault	https://developers.google.com/knowled	从互联网数据中抽取知识并构建关联网络	2014	谷歌公司，美国	16 亿个三元组，4500 个概	否	—

	ge-graph				念，4469 种关系。		
Wikidata	https://www.wikidata.org/wiki/Wikidata:Main_Page	它包含各种领域的知识，例如人物、地点、事件、作品、科学知识等，并使用 RDF 格式存储	2012	Wikimedia 基金会，美国	2.6 亿个实体以及数十亿条关系数据	是	1.37
ConceptNet	https://conceptnet.io/	基于人工智能的开放源代码项目，旨在构建一个跨语言的关联网络	2004	MIT Media Lab 自然语言处理小组	涵盖约 40 万个概念和 3500 万个关系	是	5.7.1
VIVO	http://vivo.library.cornell.edu	使用实体关系本体模型来组织和呈现有关人员、研究和教育活动的信息	2003	康奈尔大学	25 个国家和地区的 300 多个实例，覆盖了众多学科领域	是	
XLore	https://www.xlore.cn/index	中英文双语的百科关联网络		清华大学计算机系知识工程研究室	1.49 亿个实体，51 万个关系	是	
GeoSciML	http://geosciml.org/	创建一套服务于地表数据共享传输的数据模型	2003	国际地球科学信息委员会，—	包含 1772 个概念	是	4.1
SWEET	https://github.com/ESIPFed/sweet	创建一套描述地球科学领域的本体库。	2009	国家航空航天局，美国	4533 个概念以及 359 个属性。	是	3.5.0
GeoCODES	https://geocodes.earthcube.org/#/landing	项目旨在改善地球科学数据集的互操作性和发现性，从而促进科学合作和数据共享	2011	美国国家科学基金会	共享平台	是	

3.2 研究对象确定

初步筛选的关联网络中存在数据量不足、活跃度低等问题，需要根据一定原则进行剔除。本文考虑使用以下原则：

（1）通用性：研究对象应为通用型地表关联网络，而面向小型案例的任务型关联网络将被剔除。

（2）数据独特性：当多个关联网络中的数据存在重复时，只保留其中一个。

（3）可获取性：关联网络中的数据应该容易获取，能够在研究中方便地使用和操作，

不易获取的网络将被剔除。

（4）活跃度：关联网络应该保持一定的活跃度，能够反映当前研究领域的最新发展，同时也能够吸引更多的研究者关注和参与，活跃度低的应被剔除。

（5）数据量充足：关联网络中的数据量应该足够丰富，能够满足研究需要，同时也要避免数据量过大导致研究效率降低，数据量不足的网络将被剔除。

经过上面的原则处理后，保留 5 个地表关联网络 GeoSciML、OSM Semantic Network、OEG、SWEET 和 Geospatial Data Cloud 以及 4 个通用关联网络 DBpedia、YAGO、DEU 和 Wikidata，共同构成本文地表关联网络比较分析的目标对象。

4 关联网络发展现状比较分析

本节针对已经筛选出的关联网络，分别从顶层构建方法和具体内容设计实现两方面对其进行比较分析。

4.1 构建方法分析

构建方法是关联网络研究的核心。通过比较分析被选择的研究对象，从数据来源、构建方式，自动化程度，更新方式对现有关联网络进行分析，并对其进行总结。

4.1.1 数据来源

关联网络中元数据的来源主要有两类[27]，第一类是互联网上公开的开放科学数据网站中[28]，这类网站通常包含大量的数据集信息，此类数据信息的获取较为简单；第二类是研究论文中为了支持研究结果和结论而公布的数据集，这类数据信息通常分布分散且不易收集。

通常，筛选出的关联网络中的数据来源都属于第一类，第二类数据集信息的获取可以作为后续研究的重点关注。

4.1.2 构建方式

关联网络的构建方式可以分为自顶向下、自底向上和二者混合三种[29]。自顶向下（Top-down）构建方式：这种构建方式是由先验知识或领域专家的经验来指导网络的构建。通常是先定义一些高层次的概念或目标，然后通过分解和细化这些概念或目标来构建一个层次结构。最终，通过将这些概念或目标之间的关联关系加入到网络中，来构建一个完整的关联网

络。自底向上（Bottom-up）构建方式：这种构建方式是从具体的数据或实例中提取出关联关系来构建网络[30]。通常是通过对数据进行聚类、分类或分析等操作，将相似的数据或实例进行组合，然后通过探索这些组合之间的关联关系来构建网络。二者混合的构建方式：这种构建方式是自顶向下和自底向上的结合，即同时考虑先验知识和具体数据的特征。通常是将先验知识作为一种先验概率，然后通过对数据进行学习和推理，更新先验概率，来构建一个更加准确和可靠的关联网络。

GeoSciML、GeoNames、SWEET、YAGO 和 DBpedia 均采用自顶向下的构建方式[31]，OSM Semantic Network、Wikidata 和 DEU 均采用混合构建方式，通常情况下，自顶向下的构建方式适用于具有明确目标和高度结构化的数据，适用于规模较小的网络；自底向上的构建方式适用于无结构和非结构化的数据，这种方式更新快、支持大数据量的关联网络构建，但数据噪音大、准确性不高；而二者混合的构建方式则适用于复杂的数据和目标模糊的情况，但是其灵活度高且构建难度大。

4.1.3 自动化程度

从自动化程度来看，自动化程度指关联网络构建过程中的人工参与程度，可分为人工、半自动和自动[32]。在人工构建的关联网络中，所有的节点和边都是由人工手动添加的。这种方式需要大量的人力和时间，适用于小规模的网络或者需要高度控制的情况；在半自动构建的关联网络中，首先通过算法或者其他自动化的方式生成一部分的节点和边，然后由人工对其进行进一步的筛选和修正。这种方式相对于人工构建，可以减少人工的参与程度和时间，同时也可以保证网络的质量和准确性；在自动构建的关联网络中，所有的节点和边都是通过算法或者其他自动化的方式生成的，完全不需要人工的参与。这种方式适用于大规模网络的构建，可以大大减少人力和时间成本，但是也可能会存在一定的误差和不准确性。自动化程度与构建方式相关，采用自底向上构建方式的关联网络通常均为全自动化；而采用自顶向下构建方式时则有不同程度的人工参与，从已有的关联网络来看也基本遵循这一点，GeoSciML，OSM Semantic Network 和 SWEET 都是人工构建的[33,34]，有大量志愿者参与其中。DBpedia 和 Wikidata 是半自动的，YAGO 是自动的[35]。

4.1.4 更新方式

关联网络的更新方式通常可以分为增量式更新、全量式更新[29]、增量-全量混合更新和

增量式快照更新。增量式更新是指在原有关联网络的基础上，新增加一些节点和边，或者对已有节点和边进行修改。这种方式适用于需要频繁更新的情况，可以减少更新的成本和时间。例如，在地表关联网络中新增加元数据信息或者修改元数据信息等；全量式更新是指在更新关联网络时，需要重新构建整个网络。这种方式适用于原有网络结构发生较大变化，或者需要对整个网络进行重新优化的情况。但是，全量式更新需要耗费较大的计算资源和时间；增量-全量混合更新是指在更新关联网络时，先进行增量式更新，然后在一定时间间隔后或服务器空闲时，再进行全量式更新。这种方式可以在保证更新效率的同时，保证网络的准确性和稳定性；增量式快照更新是指在更新关联网络时，先对原有网络进行快照备份，然后再在快照备份的基础上进行增量式更新。这种方式可以保证数据的安全性和可恢复性，同时也可以减少更新的成本和时间。对于大型关联网络如 DEU、YAGO[36]和 Wikidata[37]，使用增量式更新，而对于小中型关联网络如 GeoSciML [38](Geoscience Markup Language)一般使用全量更新。

4.2 构建内容设计分析

本章将介绍关联网络构建的内容分析的基本原理和方法，包括关联指标构建、特征提取及表示、特征关联、关联网络构建，关联网络评价等方面。首先是关联指标构建：介绍如何根据研究对象和研究目的选择适当的关联指标，如空间特征、时间特征和内容特征等，并介绍这些指标的选取方法。然后是特征提取及表示：介绍如何从数据中提取关键特征，如空间、主题、时间等，以及如何将特征表示为向量或矩阵的形式，以便于后续的关联分析分析。接下来，我们将详细介绍特征关联与计算：介绍如何将特征关联及计算特征之间的关联度或相似度，以及如何选择合适的关联度或相似度度量方法，如余弦相似度[39]、皮尔逊相关系数[40]等。接着，我们将介绍介绍如何根据特征之间的关联度或相似度构建关联网络和相关算法，以及网络的参数设置方法。最后，介绍如何对构建好的关联网络进行评价，如网络密度[41]、聚类系数[42]、平均路径长度[43]等指标，以及如何利用这些指标来评估关联网络的质量和可靠性。

4.2.1 关联指标构建

关联地表开放数据，需要建立一个完整的关联指标体系，以表示地表数据中模糊、不精确和不完全的知识与关系。这个指标体系不仅为特征提取和表示提供支持，还为关联关系的

计算奠定了基础[17]。

关联指标的选取至关重要，可以为语义抽取，特别是逐级关联计算奠定基础[44]，应尽量按照“全面性、代表性、数据可获取性”的原则[17]。元数据通常包括数据集名、数据格式、空间范围、时间要素、比例尺、空间坐标参考系等多方面的特征。这些特征可以总结为: 本质特征、形态特征和来源特征[18]。罗侃等[17]构建的关联指标体系包括数据本质特征，形态特征，来源特征三种。其中数据本质特征包括数据内容、空间特征、时间特征；数据形态特征包括数据结构、数据基准、数据精度、数据语言和存储介质；数据来源特征包括数据的来源及处理过程。使用专家打分的方式为这些指标赋予计算权重。此方法几乎选取了元数据的所有特征，精确度较高，但计算起来复杂度较高。赵红伟等[18]构建的关联指标体系选取空间特征，时间特征，专题特征这三个本质特征和数据来源特征，此方法只考虑了最重要的数据特征，但计算起来较为方便。

当涉及到语义关联时，考虑到数据目录中的每个特征都会增加关联网络的复杂性，而且可能会降低主要语义的应用效果。因此，在建立语义关联网络时，必须做出选择并仅选择最重要的元数据信息，以使网络具有更高的灵活性、可调控性和更明确的应用目标，同时需要根据地表数据的特征进行针对性的调整。

4.2.2 特征提取及表示

首先是在特征提取方面，由于数据目录的特征不能够完全获取，例如，网页缺少描述或数据特征存在于一段摘要中，因此地球表层数据目录的特征主要有两个来源，一是存在于数据目录的核心要素中，有时间要素项表示时间特征，有空间范围项和空间坐标参考系项表示空间特征，有关键词项和数据分类项表示内容特征；二是存在于数据集名和摘要中，数据集名和摘要中可能存在时间、空间和内容特征。因此需要对数据集名和摘要进行特征提取。对于不同类型的关键词，提取算法也不尽相同，下面根据不同特征所使用的算法进行介绍。

1. 对于时间和空间这类关键词，它们通常有固定的格式，因此提取方法也相似，主要有：基于规则的方法、基于词性标注的方法、基于统计的方法[45]三种，基于统计由于简单、快速是比较常用的，如 TextRank [46]和 TF-IDF (term frequency–inverse document frequency)[47]。

2. 对于内容关键词地表领域有专门基于领域专业词汇的方法：地表领域有很多专业词汇，如地理名称、地质学术语、地球物理参数等，可以通过事先构建地表专业词汇表，并利用词汇表对文本进行匹配来提取地表关键词。这种方法简单直观，适用于特定的地表领域。

3. 对于时间、空间和内容关键词都适用的方法主要有下面两种。

（1） 基于机器学习的方法[48]：通过训练模型来自动识别文本中的时间和空间关键字，如基于决策树[49]、支持向量机（Support Vector Machine，SVM）[50]、条件随机场（ConditionalRandom Field，CRF）[51]及最大熵模型等[52]。这种方法可以考虑词性、上下文和语义信息，具有一定的准确性和灵活性，但需要手工设计特征和进行模型训练。

（2）基于深度学习的方法：它通过使用深度神经网络模型，如循环神经网络（RNN）、卷积神经网络（CNN）、LSTM（长短期记忆网络）、Transformer[53]以及这些模型的组合等，来学习文本中的空间和时间关键字，近来基于 BERT[54]等大模型融合领域特征进行调优的方法也可以取得不错的效果，而且可以避免从头开始训练模型的时间和计算成本。这种方法可以捕捉词义和语义信息，对复杂的空间表达方式和多语言场景具有较好的适应性，但需要大量的训练数据和计算资源。

然后是特征表示方面，对于内容特征关键词，由于地表开放数据目录中提取的关键词往往是名词的组合词汇如森林面积、草地类型，将数据转换成向量表示，采用 One-Hot Encoding[55]、Bag of Words、Word Embedding[56]（词嵌入，如 Word2Vec、GloVe、BERT 等）模型实现词向量的生成。

时空位置的表示主要有数字（例如经纬度等）和文本（例如地名等）两种方式。

对于空间特征关键词，经过标准化处理的空间特征一般为一个或多个地理名词，用来表示所属数据目录的空间范围。对于空间特征的表示，需首先考虑点位置的编码。目前，点位置的编码方法主要有离散化方法、直接位置编码、正弦函数编码、正弦函数多尺度编码等方法。对于线或面等复合实体的空间位置，则需要对其包含的点位置的编码进行聚合。目前，主要的聚合方法有基于核函数的方法、全局聚合方法、局部聚合方法和层次聚合方法等。

对于时间特征关键词，主要包含时间点与时间点、时间点与时间段、时间段与时间段之间的拓扑关系，这三种关系可以分类进行比较判断。有学者将时间点全部转化为时间段，如时间点“2022 年”，转换后的时间段就为“2022 年 1 月 1 日到 2022 年 12 月 31 日”，将时间拓扑关系计算全部转换为时间段时间拓扑关系计算，简化了计算。与空间位置编码相似，时间位置编码同样首先需要对点时间进行编码，然后再进行聚合得到段时间或复合时间的编码。目前，已有 CTDNE、IGE、DynamicTriad、DDNE、DANE、DynGem、NP-GLM、DGNN、DyRep 等模型。

4.2.3 特征关联与计算

科学数据语义特征关联模型是通过数据特征表示信息之间的综合语义关系构建，建立不同数据特征之间的关联关系表示，进而形成描述数据间关系的词汇集，基于词汇集计算它们的关联度，并将其归化到[0,1]区间内，将关联关系和关联度作为节点之间的边，其中值越靠近 1 证明关联度越大，越靠近 0 证明关联度越小。

1. 内容关联关系主要是内容类别关系和内容语义关系。

对于内容类别关系，目前，国内外存在多种专题分类体系，它们大都以分类树的形式描述，一般用 Brother-of、Son-of、 Father-of 等词语来描述它们之间的关系[18]，可以由专家打分获取权重，进而得到内容类别相似度。

对于内容语义关系，语义相似度算法是一种用于计算两个关键词之间语义相似程度的算法。通常情况下，该算法基于词向量模型，将每个单词表示为一个向量，并使用这些向量来计算两个单词之间的相似度。主要的算法有余弦相似度[57]、欧几里得距离[58]、词嵌入模型（Word2Vec 和 GloVe）[59]、皮尔逊系数（Pearson correlation coefficient）、Jaccard 相相似系数 (Jaccard Similarity Coefficient)、基于《知网》的相似度算法[60,61]等。这些算法各有优劣，如余弦相似度适用范围广，词嵌入模型需要大规模语料，基于知网的相似度算法偏重于中文等。通过选择上述算法之一可以得到内容语义相似度，将其与内容类别相似度加权最终得到内容相似度。

2. 空间关联关系，地理空间是一个相对空间，是一个目标组合排列集，在早期的研究中，有学者把空间关系主要分为拓扑关系、度量关系和顺序关系[62,63]。随着地理信息系统的广泛研究与应用，众多学者对空间关系进行了大量深入细致的研究，但大都是在拓扑、度量和顺序三种关系的基础上进行扩展。拓扑关系是指拓扑变换下的不变量，如空间实体的重叠和相接关系；顺序关系指的是空间实体在空间中的某种排序，如前后、左右、上下、东西南北等；度量关系是指用某种度量空间中的度量尺度来描述空间实体之间的关系，如距离关系和尺度大小等。在这些关系中，拓扑关系最为重要[64]。

空间拓扑关系目前主要有两大类，Randell 等人提出的区域连接演算 RCC(region connection calculus)理论[65~67]和 Egenhofer 等人提出的求交模型[68,69]。

（1）RCC 分为 RCC-8 和 RCC-5。RCC-8 包括：不连接（DC）、外部连接（EC)、部分交叠（P0)、正切真部分（TPP)、非正切真部分（NTPP)、相等（EQ)、反正切真部分（TPPI)和反非正切真部分（NTPn)。由于 RCC5 对边界不敏感，即将 DC 和 EC 合并为分离（DR)，

将 TPP 和 NTPP 合并为真部分（PP), TPPI 和 NTPPI 合并为反真部分(PPI)[70]。RCC-8 所做的拓扑区分非常符合人类对空间拓扑关系的认知，Knauff 等人[71]，Renz 等人[72]也从认知实验角度对这种拓扑区分进行了评价。很多学者使用了 RCC-8 的这种拓扑关系区分方式[73]。

（2）4-交模型基于四个基本的空间关系：包含（containment）、相交（intersection）、相离（disjointness）和相等（equality），用于描述空间对象之间的相互关系。但是 4-交模型对空间对象形状和位置的要求较高，对空间关系的描述较为简单等。后面 Egenhofer 以点集拓扑学为理论依据提出了 9-交模型[74]，它包括 Equals（相等）、Disjoint（不相交）、Contains（包含）、Within（被包含）、Intersects（相交）、Touches（相邻）、Crosses（穿过）、Overlaps（重叠）、Covers（覆盖）共九种。基于空间拓扑分类体系，判断空间拓扑关系方法也不尽相同；有学者使用自己创建的地理空间基础信息库进行判断，由于信息库是人工生成的，因此判断结果较为准确，但也会导致普适性较差；有学者通过调用开放 API 的地理编码,例如谷歌地图 API 或使用 ArcGIS 等 GIS 软件，将地理位置区域转化为用经纬度为顶点的多边形（四至点等），进而判断拓扑关系并分配权重，拓扑关系的权重可以由专家打分得到或使用训练数据集生成。

在拓扑关系确定的基础上，判断空间度量关系，它包含空间重叠比例和空间距离两个指标。空间重叠比例是几何实体重叠部分的面积或长度与实体面积或长度的比值。空间距离是两个空间实体的最短距离。空间度量值的计算方式由空间拓扑的类别决定。最终空间关联度由空间拓扑值和空间度量值加权得到。

3. 时间关联关系，时间关联关系主要包含时间拓扑关系和时间度量关系[75]。在时间拓扑关系研究方面有：Point Algebra 模型：这是一种基于点代数的时间拓扑关系模型，通过使用点代数的操作符和谓词来描述时间上的拓扑关系，如在、在之前、在之后等。TimeML 模型：这是一种用于表示和处理时间信息的语义标记语言，包含了丰富的时间拓扑关系描述，例如：包含、重叠、开始、结束等。基于时间区间代数（ Interval Algebra）理论有学者于 1983 年提出了一种包含 13 种时间拓扑关系的模型，总结了常见的关系如相等、包含、在之后、在之前等，提出了 equal、contain 等 13 种基本的时间拓扑关系，成为时间关系研究的基础，同样时间拓扑关系权重值可以由专家打分得到，也可以由数据训练得到。

时间度量关系是指时间上的量度和计量关系，用于描述和比较不同时间点、时间段或时间跨度之间的时间差异。时间度量关系可以基于不同的标准和度量单位进行计算和表示，常见的包括时间间隔和时间长度。最终时间关联度由时间拓扑关联度和时间度量关联度加权得到。

4.2.4 关联网络构建与评价

在特征提取以及关联后，关联网络的构建主要是存储方式的不同，对于小规模的网络通常使用邻接矩阵[76]，邻接矩阵中，每个节点对应矩阵中的一行和一列，每个边对应矩阵中的一个元素。如果两个节点之间存在一条边，则矩阵中对应的元素值为关联度。对于大规模的网络一般使用图数据库[77]进行存储，图数据库是一种专门用于存储和查询图形数据的数据库。它可以高效地存储和查询网络中的节点和边，支持高级查询和分析操作。图数据库适用于大规模网络数据的存储和查询。

关联网络构建完成后，还有学者进行了阈值设置，子图提取与可视化，分析与评价等步骤，进而提高关联网络的准确性和健壮性。

（1）阈值设置，在特征计算过程中，所有的元数据两两之间都形成了关联度，会生成大量有低关联度的边（噪声），使得网络具有极高的密度。阈值提取就是设置一个阈值来筛选出关联度较高的特征对，比如只保留关联度高于某个阈值的特征对。可以提高关联网络的准确性和可靠性。

（2）子图提取与可视化：有的网络构建完成后是非常庞大和复杂的，会使用户难以找到所需要的部分，可以筛选特征对，提取出一个子图，其中每个节点表示一个特征，每条边表示两个特征之间的关联关系。对于提取出的子图，进行可视化分析，以便更好地理解特征之间的关联关系。例如，可以使用图形布局算法将节点和边布局在二维平面上，以便观察和分析。

（3）分析与评价，评价内容主要有，网络结构分析：关联网络的结构特征可以通过度分布、聚集系数、连通性等指标进行分析。例如，可以计算网络的平均度、直径、平均路径长度等指标来了解网络的规模和复杂性。社区检测[78]：社区是指网络中密集相连的节点的集合，通常表示节点之间存在相关性或功能相似性。社区检测可以帮助发现网络中的重要模块和组织结构，并深入理解网络的功能。中心性分析：中心性是指网络中节点的重要程度[79,80]，可以通过度中心性、接近中心性、介数中心性等指标进行评估。中心性分析可以帮助了解网络中的关键节点和信息传播路径。

5 讨论与启示

关联网络可以帮助我们更加高效地管理和利用地表数据，通过关联网络，可以将不同来源的数据资源进行关联和整合，形成一个统一的数据资源库，方便不同用户和应用程序的使

用和共享。在对现有地球表层关联网络构建方法和构建内容设计比较分析的基础上，本文发现，当前关联网络泛化能力不足，构建的自动化程度也较低，现有关联网络在关联指标构建还不够完善；特征关联技术方法还不够成熟；计算关联度时，对于各个特征的权重分配偏向主观。本文从数据来源、构建方法和构建内容设计分析三方面对地表关联网络的构建提出以下几点建议。

（1）提高数据收集质量

第一，地表系统开放数据存在分散、多源、异构、多模态的特点，通常以专题共享网站、数据服务、元数据、期刊或数据论文等形式存在，因此，研究发展适应不同模态的地表系统开放数据目录挖掘方法，形成地表系统开放数据目录，是充分共享和利用这些数据的关键科学问题。第二，元数据质量也是关联网络建设的关键。地表系统数据涉及到多个学科领域，元数据的质量和完整性对于后续分析和应用具有重要的影响。为了提高数据的质量，需要建立有效的元数据采集、过滤和审核机制。

（2）发展地表数据目录共享程度分析方法

在构建关联网络过程中，对于收集到的元数据，使用数据目录将它们存储起来，这些数据多源、异构，质量也良莠不齐，数据质量是影响开放科学数据及其共享效果的关键，应受到科研人员的高度重视。科学数据的高质量开放共享必须保证数据的准确性、完整性、一致性、及时性、可靠性、关联性。应考虑发展定量语义解析的科学数据共享评价指标与模型对数据目录的共享程度进行分析，对科学数据的可发现、可访问、数据描述信息等共享质量进行定量化评价与分析。

（3）完善特征提取及表示方法

特征抽取方面，由于各个数据网站采用的数据标准不尽相同以及词汇和语言的差异，它们无法提供完整的元数据描述信息，这些信息可能存在于数据集名字、数据集摘要中。需要使用特征提取算法对元数据描述信息进行抽取。由于地表数据特征通常包含许多专业术语或名词，使得地表领域的特征抽取效率较低或者只能依靠人工。因此，在未来关联网络的构建中，需要考虑引入领域知识，研究更适合地表领域的特征抽取算法，或建立全球统一的地表元数据标准和规范。

特征表示方面，可以考虑通过嵌入式表示将关联网络中的特征表示在低维实数向量空间中，是一种常见的特征表示方法。以使得相似的特征在向量空间中距离更近，不相似的特征距离更远。目前，地表领域对于数据特征的嵌入式表示研究还比较缺乏，需要加强利用人工智能方法来实现地表数据的嵌入式表示，为地表数据的计算和推理打下基础。

（4）完善地表特征关联及计算方法

现有的地表关联网络在内容拓扑方面通常仅具有“父子、兄弟类”等简单的语义关系，在内容语义方面仅存在“相同”等简单的语义关系，以及“包含”、“相接”关系。这种简单的关系表达不足以描述地表数据中实体间复杂的计算、演化等非线性关系，也无法支持地表数据的复杂推理。因此，在未来地表关联网络的构建过程中，应充分考虑地表数据的复杂关系和规则，以增强其在复杂关系上的推理能力。同时，已有的地表关联网络主要集中在地质学和地理学，应拓展其研究范围，包括水文学、大气科学在内的其他地学学科的关联网络构建。

（4）促进地表领域关联网络应用实施

与通用关联网络在个性化推荐、智能问答等领域已经取得的成效相比，地表领域关联网络的研究和应用还处于初级阶段，缺乏具体的应用案例和实践经验。因此，在未来的关联网络研究中，应考虑具体实践落地，尝试选择典型案例进行实际应用研究，以提升地表关联网络的应用成效。

致谢：感谢中国地质大学（北京）王成善院士、中国科学院地理科学与资源研究所周成虎院士、中国地质大学（北京）成秋明院士等的指导，感谢深时数字地球国际大科学计划（DDE）大知识工作组，感谢匿名审稿人和编辑对论文提出的修改意见和建议。

参考文献

[1] 刘昌明, 刘璇, 杨亚锋,等. 水文地理研究发展若干问题商榷[J]. 地理学报, 2022, **77**(01): 3-15. [Liu Changming, Liu Xuan, Yang Yafeng et al. Discussion on Some Problems in the Development of Hydrographic Research. Journal of the Geographical Society of China, 2022, **77**(01): 3-15.]

[2] 张猛刚, 雷祥义. 地球表层系统浅论[J]. 西北地质, 2005,(02): 99-101. [Zhang Menggang, Lei Xiangyi. On the Earth Surface System. northwestern geology, 2005,(02): 99-101.]

[3] 吴绍洪, 高江波, 戴尔阜,等. 中国陆地表层自然地域系统动态研究:思路与方案[J]. 地球科学进展, 2017, **32**(06): 569-576. [Wu Shaohong, Gao Jiangbo, Dai Erfu et al. Study on the dynamics of terrestrial surface natural geographical system in China : ideas and schemes. Advances in Earth Science, 2017, **32**(06): 569-576.]

[4] Duda K A, Abrams M J. Aster and usgs eros disaster response: emergency imaging after hurricane katrina[J]. Photogrammetric Engineering and Remote Sensing, 2005,**71**: 1346-1350.

[5]Desnos Y, Borgeaud M, Doherty M et al. The european space agency's earth observation program[J]. IEEE Geoscience and Remote Sensing Magazine, 2014,**2**(2): 37-46.

[6] 王卷乐, 孙九林. 地球系统科学数据共享标准规范体系研究与应用[J]. 地理科学进展, 2009, **28**(6): 839-847. [Wang Juanle, Sun Jiulin. Study on Scientific Data Sharing Standards and Specifications System for Earth System Science and its Application. Progress in Geography, 2009,

28(6): 839-847.]

[7] 马胜男, 魏宏, 刘碧松. 地理信息标准研制的国内外进展及思考[J]. 武汉大学学报(信息科学版), 2008,(09): 886-891. [Ma Shengnan, Wei Hong, Liu Bisong. The progress and thinking of geographic information standard development at home and abroad. Journal of Wuhan University (Information Science Edition), 2008,(09): 886-891.]

[8] Nah E, Cho S, Kim S et al. International organization for standardization (iso) 15189[J]. Annals of laboratory medicine, 2017,**37**(5): 365-370.

[9] 诸云强, 孙九林, 廖顺宝,等. 地球系统科学数据共享研究与实践[J]. 地球信息科学学报, 2010, **12**(1): 1-8. [Zhu Yunqiang, Sun Jiulin, Liao Shunbao et al. Research and Practice of Earth System Science Data Sharing. Journal of Earth Information Science, 2010, **12**(1): 1-8.]

[10] 邢文明, 郭安琪, 秦顺,等. 科学数据管理与共享的FAIR原则——背景、内容与实施[J]. 信息资源管理学报, 2021, **11**(02): 60-68. [Xing Wenming, Guo Anqi, Qin Shun et al. FAIR Principles for Scientific Data Management and Sharing : Background, Content and Implementation. Journal of Information Resource Management, 2021, **11**(02): 60-68.]

[11] 宋佳, 高少华, 杨杰,等. 科技资源元数据的关联与推荐方法[J]. 中国科技资源导刊, 2017, **49**(05): 37-44. [Song Jia, Gao Shaohua, Yang Jie et al. Association and recommendation method of science and technology resource metadata. china science & technology resources review, 2017, **49**(05): 37-44.]

[12] Lassila O. Resource description framework (rdf) model and syntax specification[J]. http://www. w3. org/TR/REC-rdf-syntax/, 1999.

[13] 赵红伟. 地理空间元数据语义关联网络的构建及其应用研究[D]. 北京: 中国科学院大学, 2016.[Zhao Hongwei. 2016. Research on the construction and application of geospatial metadata semantic association network. 博士, BeiJing: university of chinese academy of sciences, 2016.]

[14] 张骏骁. 语义约束的滑坡灾情评估优势信息智能选取方法[D]. 成都: 西南交通大学, 2019.[Zhang Junxiao. 2019. Intelligent selection method of dominant information for landslide disaster assessment based on semantic constraints. Chengdu: Southwest Jiaotong University, 2019.]

[15]吉雪强, 张跃松. 长江经济带种植业碳排放效率空间关联网络结构及动因[J]. 自然资源学报, 2023, **38**(03): 675-693. [Ji Xueqiang, Zhang Yuesong. Spatial correlation network structure and motivation of carbon emission efficiency of planting industry in Yangtze River Economic Belt. Journal of Natural Resources, 2023, **38**(03): 675-693.]

[16] Waterworth D, Sethuvenkatraman S, Sheng Q Z. Advancing smart building readiness: automated metadata extraction using neural language processing methods[J]. Advances in Applied Energy, 2021,**3**: 100041.

[17] 罗侃, 诸云强, 程文芳,等. 极地科学数据关联方法及应用研究[J]. 极地研究, 2016, **28**(3): 361-369. [Luo Kan, Zhu Yunqiang, Cheng Wenfang et al. Research on data association method and application of polar science. Chinese Journal of Polar Research, 2016, **28**(3): 361-369.]

[18] 赵红伟, 诸云强, 侯志伟,等. 地理空间元数据关联网络的构建[J]. 地理科学, 2016, **36**(08): 1180-1189. [Zhao Hongwei, Zhu Yunqiang, Hou Zhiwei et al. Construction of geospatial metadata association network. geographical science, 2016, **36**(08): 1180-1189.]

[19] 赵红伟, 诸云强, 杨宏伟,等. 地理空间数据本质特征语义相关度计算模型[J]. 地理研究, 2016, **35**(01): 58-70. [Zhao Hongwei, Zhu Yunqiang, Yang Hongwei et al. Geospatial data essential feature semantic correlation calculation model. geography study, 2016, **35**(01): 58-70.]

[20] 诸云强, 孙凯, 李威蓉,等. 地球科学知识图谱比较分析与启示：构建方法与内容视角[J]. 高校地质学报, 2023, **29**(3): 382-394. [Zhu Yunqiang, Sun Kai, Li Weirong et al. Comparative

Analysis and Enlightenment of Geoscience Knowledge Graph : Construction Method and Content Perspective. Journal of University Geology, 2023, **29**(3): 382-394.]

[21] 梁顺林, 陈晓娜, 陈琰,等. 陆表卫星遥感GLASS产品集的研发新进展[J]. 遥感学报, 2023, **27**(04): 831-856. [Liang Shunlin, Chen Xiaona, Chen Yan et al. New progress in research and development of land surface satellite remote sensing GLASS product set. journal of remote sensing, 2023, **27**(04): 831-856.]

[22] 于梦月. 基于本体的开放政府数据的元数据方案及其应用研究[D]. 大连: 大连海事大学, 2018.[Yu Mengyue. 2018. Research on metadata scheme and its application of open government data based on ontology. dalian: Dalian Maritime University, 2018.]

[23] Krishnamurthy R, Awazu Y. Liberating data for public value: the case of data.gov[J]. International Journal of Information Management, 2016,**36**(4): 668-672.

[24] 邱春艳, 陈可睿. 科学元数据标准的现状、特点与改进建议[J]. 数字图书馆论坛, 2022,(12): 10-18. [Qiu Chunyan, Chen Kerui. The status quo, characteristics and improvement suggestions of scientific metadata standards. digital library forum, 2022,(12): 10-18.]

[25] Kitchenham B A, Charters S. Guidelines for performing systematic literature reviews in software engineering[R]., 2007.

[27] 李涓子, 侯磊. 知识图谱研究综述[J]. 山西大学学报(自然科学版), 2017, **40**(03): 454-459. [Li Juanzi, Hou Lei. A review of knowledge graph research. Journal of Shanxi University (Natural Science Edition), 2017, **40**(03): 454-459.]

[28] 李涛, 王次臣, 李华康. 知识图谱的发展与构建[J]. 南京理工大学学报, 2017, **41**(01): 22-34. [Li Tao, Wang Cichen, Li Huakang. The development and construction of knowledge graph. journal of nanjing university of science and technology, 2017, **41**(01): 22-34.]

[29] 黄恒琪, 于娟, 廖晓,等. 知识图谱研究综述[J]. 计算机系统应用, 2019, **28**(06): 1-12. [Huang Hengqi, Yu Juan, Liao Xiao et al. A review of knowledge graph research. computer system application, 2019, **28**(06): 1-12.]

[30] 刘峤, 李杨, 段宏,等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, **53**(03): 582-600. [Liu Qiao, Li Yang, Duan Hong et al. Summary of knowledge graph construction technology. computer research and development, 2016, **53**(03): 582-600.]

[31] Emile Geay J, Khider D, Daniel et al. (2018). The linked earth ontology : a modular , extensible representation of open paleoclimate data.

[32] 徐增林, 盛泳潘, 贺丽荣,等. 知识图谱技术综述[J]. 电子科技大学学报, 2016, **45**(04): 589-606. [Xu Zenglin, Sheng Yongpan, He Lirong et al. Summary of Knowledge Graph Technology. Journal of the University of Electronic Science and Technology of China, 2016, **45**(04): 589-606.]

[33] Girres J, Touya G. Quality assessment of the french openstreetmap dataset[J]. Transactions in GIS, 2010,**14**(4): 435-459.

[34] Frontini F, Del Gratta R, Monachini M. (2016). Geodomainwordnet: linking the geonames ontology to wordnet. Springer International Publishing, Cham. 229-242.

[35] Mahdisoltani F, Biega J A, Suchanek F M. (2015). Yago3: a knowledge base from multilingual wikipedias. eds. *Conference on Innovative Data Systems Research*.

[36] Demidova E, Oelze I, Nejdl W. (2013). Aligning freebase with the yago ontology. eds. *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. Association for Computing Machinery, San Francisco, California, USA. 579-588.

[37] Ilievski F, Pujara J, Shenoy K. (2022). Does wikidata support analogical reasoning? Springer International Publishing, Cham. 178-191.

[38] Sen M, Duffy T. Geosciml: development of a generic geoscience markup language[J]. Computers & Geosciences, 2005,**31**(9): 1095-1103.

[39] 张振亚, 王进, 程红梅,等. 基于余弦相似度的文本空间索引方法研究[J]. 计算机科学, 2005, **32**(9): 4. [Zhang Zhenya, Wang Jin, Cheng Hongmei et al. Research on text spatial index method based on cosine similarity. computer science, 2005, **32**(9): 4.]

[40] Mukaka M M. Statistics corner: a guide to appropriate use of correlation coefficient in medical research[J]. Malawi medical journal : the journal of Medical Association of Malawi, 2012,**24**(3): 69-71.

[41] Intanagonwiwat C, Estrin D, Govindan R et al. Impact of network density on data aggregation in wireless sensor networks[J]. Proceedings 22nd International Conference on Distributed Computing Systems, 2002: 457-458.

[42] Hamilton W L, Ying R, Leskovec J. (2017). Inductive representation learning on large graphs. eds. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., Long Beach, California, USA. 1025-1035.

[43] 周云龙. 复杂网络平均路径长度的研究[D]. 合肥: 合肥工业大学, 2013.[Zhou Yunlong. 2013. Research on the average path length of complex networks. Hefei: Hefei University of Technology, 2013.]

[44] 刘志辉, 魏娟霞, 张均胜. 基于知识图谱的科技创新指标自适应计算方法研究[J]. 情报学报, 2019, **38**(08): 826-837. [Liu Zhihui, Wei Juanxia, Zhang Junsheng. Research on adaptive calculation method of scientific and technological innovation index based on knowledge graph. information learned journal, 2019, **38**(08): 826-837.]

[45] 胡少虎, 张颖怡, 章成志. 关键词提取研究综述[J]. 数据分析与知识发现, 2021, **5**(03): 45-59. [Hu Shaohu, Zhang Yingyi, Zhang Chengzhi. Summary of keyword extraction research. Data Analysis and Knowledge Discovery, 2021, **5**(03): 45-59.]

[46] Mihalcea R, Tarau P. (2004). Textrank: bringing order into text. eds. *Proceedings of the 2004 conference on empirical methods in natural language processing*. 404-411.

[47] Turney P D. Learning algorithms for keyphrase extraction[J]. Information Retrieval, 2000,**2**(4): 303-336.

[48] Firoozeh N, Nazarenko A, Alizon F et al. Keyword extraction: issues and methods[J]. Natural Language Engineering, 2020,**26**(3): 259-291.

[49] Belgiu M, Drăguţ L D. Random forest in remote sensing: a review of applications and future directions[J]. Isprs Journal of Photogrammetry and Remote Sensing, 2016,**114**: 24-31.

[50] Chang C, Lin C. Libsvm: a library for support vector machines[J]. ACM Trans. Intell. Syst. Technol., 2011,**2**(3): 27.

[51] Bale T L, Vale W W. Crf and crf receptors: role in stress responsivity and other behaviors.[J]. Annual review of pharmacology and toxicology, 2004,**44**: 525-557.

[52] 张杰. 文献结构化的细粒度检索技术研究[D]. 南京: 东南大学, 2019.[Zhang Jie. 2019. Research on structured fine-grained retrieval technology of literature. Nanjing: Southeast University, 2019.]

[53] Vaswani A, Shazeer N, Parmar N et al. Attention is all you need[J]. arXiv, 2017.

[54] Devlin J, Chang M, Lee K et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[55] Bengio Y, Ducharme R, Vincent P et al. A neural probabilistic language model[J]. J. Mach. Learn. Res., 2003,**3**(null): 1137-1155.

[56] Mikolov T, Chen K, Corrado G S et al. (2013). Efficient estimation of word representations in vector space. eds. *International Conference on Learning Representations*.

[57] Nguyen H V, Bai L. (2010). Cosine similarity metric learning for face verification. eds. *Proceedings of the 10th Asian conference on computer vision - volume part II*. Springer-Verlag, Queenstown, New Zealand. 709-720.

[58]Danielsson P. Euclidean distance mapping[J]. Computer graphics and image processing, 1980,**14**(3): 227-248.

[59] Ruder S, Vulić I, Søgaard A. A survey of cross-lingual word embedding models[J]. J. Artif. Int. Res., 2019,**65**(1): 569-630.

[60] 葛斌, 李芳芳, 郭丝路,等. 基于知网的词汇语义相似度计算方法研究[J]. 计算机应用研究, 2010, **27**(09): 3329-3333. [Yan Gebin, Li Fangfang, Guo Silu et al. Research on HowNet-based lexical semantic similarity calculation method. application research of computers, 2010, **27**(09): 3329-3333.]

[61] 王小林, 王东, 杨思春,等. 基于《知网》的词语语义相似度算法[J]. 计算机工程, 2014, **40**(12): 177-181. [Wang Xiaolin, Wang Dong, Yang Sichun et al. Word semantic similarity algorithm based on ' HowNet '. computer engineering, 2014, **40**(12): 177-181.]

[62] Sloman S A, Love B C, Ahn W. Feature centrality and conceptual coherence[J]. Cognitive Science, 1998,**22**(2): 189-228.

[63] 陈军, 赵仁亮. GIS空间关系的基本问题与研究进展[J]. 测绘学报, 1999,(2). [Chen Jun, Zhao Renliang. The basic problems and research progress of GIS spatial relationship. Journal of Surveying and Mapping, 1999,(2).]

[64] 张雪伍, 苏奋振, 石忆邵,等. 空间关联规则挖掘研究进展[J]. 地理科学进展, 2007,(06): 119-128. [Zhang Xuewu, Su Fenzhen, Shi Yishao et al. Research progress on spatial association rule mining. Progress in Geography, 2007,(06): 119-128.]

[65] Randell D A, Cui Z, Cohn A G. (1992). A spatial logic based on regions and connection. eds. *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*. Morgan Kaufmann Publishers Inc., Cambridge, MA. 165-176.

[66] Randell D, Cohn A. Modelling topological and metrical properties in physical processes.[J]., 1993.

[67] Randell D A, Cohn A G. (1989). Modelling topological and metrical properties in physical processes. eds. *Proceedings of the first international conference on Principles of knowledge representation and reasoning*. Morgan Kaufmann Publishers Inc.. 357-368.

[68] Egenhofer M J, Franzosa R D. Point-set topological spatial relations[J]. International Journal of Geographical Information Systems, 1991,**5**(2): 161-174.

[69] Egenhofer M, Herring J. Categorizing binary topological relations between regions, lines and points in geographic databases, the 9-intersection: formalism and its use for naturallanguage spatial predicates[J]. Santa Barbara CA National Center for Geographic Information and Analysis Technical Report, 1990,**94**: 1-28.

[70] 虞强源, 刘大有, 谢琦. 空间区域拓扑关系分析方法综述[J]. 软件学报, 2003,(04): 777-782. [Yu Qiangyuan, Liu Dayou, Xie Qi. A survey of topological relation analysis methods in space domain. journal of software, 2003,(04): 777-782.]

[71] Knauff M, Rauh R, Renz J. (1997). A cognitive assessment of topological spatial relations: results from an empirical investigation. In: Hirtle S C, Frank A U Springer Berlin Heidelberg, Berlin, Heidelberg. 193-206.

[72] Renz J, Rauh R, Knauff M. (2000). *Towards cognitive adequacy of topological spatial relations*.

Springer Berlin Heidelberg, Berlin, Heidelberg. 184-197.
[73] 王芳. 空间方向关系推理及多种空间关系结合推理的研究[D]. 吉林: 吉林大学, 2012.[Wang Fang. 2012. Research on spatial direction relation reasoning and the combination of multiple spatial relations reasoning. Jilin: Jilin University, 2012.]
[74] Brahim L, Okba K, Robert L. Mathematical framework for topological relationships between ribbons and regions[J]. Journal of Visual Languages & Computing, 2015,**26**: 66-81.
[75] 高云亮. 地理信息资源关联关系的可视化方法研究与实践[D]. 郑州: 解放军信息工程大学, 2017.[Gao Yunliang. 2017. Research and practice on visualization method of geographic information resource association relationship. Zhengzhou: PLA University of Information Engineering, 2017.]
[76] J. R G. (2018). Graph algorithms in the language of linear algebra: how did we get here, and where do we go next? eds. *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. 268.
[77] Angles R, Gutierrez C. Survey of graph database models[J]. ACM Comput. Surv., 2008,**40**(1): 1.
[78]吕天阳, 谢文艳, 郑纬民,等. 加权复杂网络社团的评价指标及其发现算法分析[J]. 物理学报, 2012, **61**(21): 145-154. [Lu Tianyang, Xie Wenyan, Zheng Weimin et al. Evaluation index of weighted complex network community and its discovery algorithm analysis. Acta Physical Sinica, 2012, **61**(21): 145-154.]
[79] 于会, 刘尊, 李勇军. 基于多属性决策的复杂网络节点重要性综合评价方法[J]. 物理学报, 2013, **62**(2): 1-9. [Hui Yu., Liu Zun, Li Yongjun. Comprehensive evaluation method of complex network node importance based on multi-attribute decision making. Acta Physical Sinica, 2013, **62**(2): 1-9.]
[80] 张琨, 沈海波, 张宏,等. 基于灰色关联分析的复杂网络节点重要性综合评价方法[J]. 南京理工大学学报, 2012, **36**(4): 579-586. [Zhang Kun, Shen Haibo, Zhang Hong et al. Comprehensive evaluation method of complex network node importance based on grey correlation analysis. journal of nanjing university of science and technology, 2012, **36**(4): 579-586.]