知识发现中异构信息标准化处理研究
——以资源环境领域文献为例

曲建升[1] 刘红煦[1,2]
[1]中国科学院兰州文献情报中心 兰州 730000 [2]中国科学院大学管理学院 北京 100049

摘要: [目的/意义] 以数据集成过程中异构信息的集成为研究目标,在保证文献综合集成系统对信息提取的准确性要求的基础上,以资源环境学科为例,提出一种异构信息的标准化处理方式。[方法/过程] 采用团队自建的资源环境学科知识本体为依据,通过对资源环境学科异构信息在地理空间、时间单位及属性提取中的标准化分析,提出异构信息标准化处理的思路,指导搭建实现信息集成、支持综合集成的人机交互的文献综合集成平台。[结果/结论] 最终主要针对不同数据格式、不同来源的文献进行知识格式化提取及处理,完成文献综合集成的数据准备阶段的工作。异构信息标准化处理仅仅是知识发现过程的起点,后续将重点关注标准化的信息统计分析及可视化展示,完整实现文献综合集成的知识发现过程。

关键词: 知识发现 异构信息 标准化 领域文献
分类号: G250
DOI: 10.13266/j.issn.0252-3116.2016.06.013

1 引言

图情研究中,知识发现或文献数据分析通常以中国知网、WOS 等为数据来源,相对较为单一。邱均平等[1]曾指出在对学科知识结构进行分析时,若能在数据来源方面有所创新,将会在很大程度上促进计量分析的推广应用。因此,在分析文献的同时,如果能采集全文数据,对多源异构的文献全文进行深入的内容分析,将有助于得出更有价值的结论。但由于多源异构数据存在着较大的异构性,用户很难通过统一的基于标准的方式对数据进行处理,这对文献数据集成提出了挑战。因此,笔者提出了一种文献研究结论描述方式——将不同来源的文献全文数据中有价值的信息(这里主要指研究结论)进行格式化提取,实现异构数据的综合集成,并以资源环境领域文献为例,验证这一标准化模式的可行性。该过程与传统的资源发现不同之处在于:通过对大量研究成果进行整合、集成,以得到具有一定可信度的研究结论,是突破资源发现的知识发现过程。通过基于文献研究结论的知识发现,可以为科学研究提供新的线索,为科技创新研究提供合理假设。

本文针对知识发现中异构信息处理这一研究主题,主要研判了基于文献的知识发现的研究背景及现有异构信息标准化的方式,提出处理异构信息需要解决数据的多元性及多源性问题,即从类型及格式两方面对异构信息进行标准化,具体处理方式视文献所属研究领域而定。笔者以资源环境领域文献为例,探索具体领域异构信息的标准化处理模式,以解决知识发现层面的异构性问题,继而研究如何构建异构领域文献处理的框架,并实现具有标准化处理接口的服务,该服务以规范化接口处理多种异构文献数据,向用户提供诸如单位转换、格式转换等数据定制操作的接口,并同时提供相应的元数据信息,并将相关处理过程的信息记录到数据库中。

2 国内外相关研究评述

目前,基于文献的知识发现逐渐成为信息资源整合建设中的关键问题,而知识发现过程中异构信息处理则成为解决这一问题的拦路虎。情报学研究人员早在 20 世纪 80 年代就对知识发现进行了相关探索,D. R. Swanson 最早提出“未被发觉的公开知识”这一概念[2],知识发现[3]是一个以数据库、人工智能、数理统计、可视化四大技术为基础,多学科交叉融合形成的新的交叉学科,目的是把大量原始的数据转换成有价值

作者简介: 曲建升(ORCID: 0000-0002-2806-3447),研究员,博士,硕士生导师; 刘红煦(ORCID: 0000-0002-4809-1437),硕士研究生,通讯作者,E-mail: hbelhx@163.com。
收稿日期: 2016-01-21 修回日期: 2016-03-05 本文起止页码: 84-90 本文责任编辑: 刘远颖

的知识，挖掘出隐含、精炼和高质量的信息。最常见的知识发现是“基于文献的知识发现”，其哲学思想来自波普尔关于科学发现的“猜想－反驳方法论”以及“知识是客观的，本质上是猜测性的”的知识论[4]，基于文献的知识发现可以直接促进新知识的产生。公开发表的文献中有时也蕴含着未被发现的及尚未预见到的知识，国内外图情学者对以揭示“未被发觉的公开知识”为目的的“基于文献的知识发现”进行了不断的探索。荣毅虹等[5]总结出基于文献的发现是以揭示蕴含于公开发表的文献，但尚未被人们认识或发觉的知识片段间的逻辑联系，从而提出知识假设，以便专业研究人员进一步证实，促使新知识产生为目的的情报研究。

2.1 异构信息处理

自20世纪90年代开始，文献数字资源在数量和规模上迅猛增长，形式异常丰富，由于文献来源各异，必然导致这些资源在形式上多元异构，文献数字资源兼具形式“多元化”和来源“多源化”的双重特点：一方面在形式上囊括文本、影像、音频等多种类型，另一方面涵盖中外文文献、中外文不同数据库的多种来源。如何处理多元异构的数据形式及多源的信息资源，国内学者开展了大量研究，探索对这些资源的整合、集成和融合问题[6]。牛奉高[7]指出需采用数据融合处理多元异构的数据形式，同时通过整合集成的研究处理多源的数据资源，从而将资源聚合的概念拓展到知识发现的层面。窦天芳等[8]基于Web环境下开放与协作的理念，提出了图书馆应用多源数据开展集成服务的思路，并对多源数据的选择、抽取、加工及集成应用进行了介绍。

针对多种来源的数据资源，为了实现分布和异构系统间的互操作和信息共享，学者们提出了通过数据整合消除异构数据源在不同数据库系统的分布性和异构性。肖希明等[9]认为数字文化资源整合是将图书馆、博物馆、档案馆、美术馆、文化馆等公共文化服务机构相对分散的、异构的馆藏数字资源进行类聚、融合、重组，从而实现资源“一站式”查找与获取。通过一站式检索，将原本交叉重复的各个数据库内容整合集成，统一检索。C. Calistru等[10]提出了一种描述、存储、传递数字化馆藏资源的理论模型。F. Hernández等[11]分析了解决数字馆藏资源分布式存储带来的信息标准及存储格式标准化等问题。目前较为主流的资源整合模式是基于知识本体的数字资源整合模式，基于语言建模等语义互操作技术及基于元数据整合的关联数据也在为推进旨在实现分布式异构数据共享的数字资源整合服务提供技术支持。

为处理多元异构的数据形式，国内外学者提出了信息整合、知识整合等概念，以解决信息实体间复杂关系的整合问题。W. B. Rayward[12]最早提出了数字资源整合，相比传统载体形式，数字资源整合需揭示信息实体间的概念和语义关系，信息整合的结果应是实现知识的集合以满足用户的知识需求，知识整合应运而生。知识整合是以知识组织方法为指导，以数据整合、信息整合为基础，以知识组织体系为支撑，组织资源知识结构中概念及概念关系的一种整合方式[13]。哥伦比亚大学开发的多文档自动文摘系统Newsblaster[14]，将网上每天的重要新闻进行聚类，对文档进行冗余消除、信息融合、文本生成等处理后生成一篇简明扼要的摘要文档。主题图、本体是新型知识组织体系，本体由于具有知识组织体系的功能，能够实现对知识结构的描述与揭示，从而成为知识组织的主流技术。由于不同的文献领域、实践项目具有不同的目标和特征，因此并没有一个广泛适用的整合模型在国际上通用，这也成为研究者需要着重突破的问题之一。

2.2 知识发现中异构信息标准化

基于文献的知识发现通过文献综合集成实现，即针对结构不同、用途不同、特征及性质不同的数据，通过一定的人工、技术手段最终实现在物理上的集中或逻辑上的集中，数据集成的核心任务就是对各种信息进行标准化或规范化[15]。从数据类型、数值的表示方法、数据的取值范围、数据语义等方面，所要集成的各个数据源具有很强的异构性。对于不同的集成对象，其集成的资源异构性各不相同，也即随着集成方法的发展，要解决的问题也在发生变化。文献集成系统的最终目标是实现知识集成，因为系统之间交换的数据和信息都是知识的载体，知识集成最具潜力也最难于实现。

异构信息知识发现主要涉及到信息提取、数据库知识发现、数据仓储等技术。由于需要处理的数据十分复杂，没有特定的模型描述，信息提取技术根据处理页面的类型可分为自由文本、半格式化文本及格式化文本信息提取；根据抽取原理不同分为基于规则、基于统计和多策略混合提取；根据用户与系统的交互方法分为手工构造、监督、半监督和无监督[16]。目前已经存在一些识别抽取模式，如直接利用HTML网页内容表示的内在规律，通过自主归纳，获得网页相关内容，该过程无需用户参与网页内容模式的获取过程，但准确率普遍不高。根据文献信息的页面格式大致相似的特点以及知识发现对信息提取准确性的要求，目前可采用的识别抽取模式为基于特定规则描述语言以及人

工标记训练样本，通过多种归纳学习方法，自动获得HTML网页内容，初期可采用人工手动获取方法，即阅读文献，手动格式化提取信息。

综上，处理异构信息需要解决数据的多元性及多源性问题，传统的资源整合集成重在解决信息资源分布的异构性或是由于分布异构而导致的数据格式异构等问题，尚不存在一个广泛适用的信息资源整合模型，需要一个较为成熟的领域知识集成和发现模式。而文献综合集成的最终目的则是对资源进行深度知识发现，通过语义层面实现文献内容的完全融合，使新知识获取成为可能，而这离不开语义层面的文本挖掘和知识发现。笔者以数据和信息集成过程中异构信息的标准化为研究目标，通过提出异构信息标准化处理的思路，指导未来如何搭建实现信息集成、支持综合集成的人机交互的文献综合集成平台。

2.3 文献信息资源异构性分析

本文的异构信息指来源不同、表示形式各异的文献信息，是广义上的“异构”，而非传统数据库分析角度的“异构”，笔者将主要针对不同数据格式、不同来源的文献，即数据的多元性及多源性问题进行知识的异构性分析，通过进一步的标准化提取及处理，完成文献综合集成的数据准备阶段的工作。

文献信息资源异构性主要体现在文献资源类型异构和格式异构上，具体来说数字文献资源属于狭义信息资源的范畴，即属于人类通过科学的手段获取、加工等创造的信息和知识。图书情报学界一般根据出版物形式和内容来划分信息资源的类型，具体以出版物形式为主、知识的内容和载体形态为辅来执行。按照这种划分方式，信息资源主要包括图书、连续出版物、特种文献、非书资料（统称为传统信息资源）和网络信息资源等几种基本类型，数字文献资源的建设正在由传统的“图书秩序”走向现代化的“数字秩序”[17]，随着信息技术和人类社会的发展，网络信息资源以外的4种信息资源都会以数字的形式存在或者数字形式和纸本形式并存，而且传统形式的资源会不断被数字化和网络化。数字化资源逐渐成为信息资源的主要形式，也是本研究的主要对象形式。文献资源格式异构是指把大量的网络文本及印刷型文本二进制序列的信息数字化。数字文献资源主要以文本信息资源为主，如PDF格式文档、DOC格式文档以及CNKI采用的CAJ格式和KDH格式等。由于受使用权的限制，处理不同的文档格式使得基于全文的分析耗时耗力[7]。

3 面向知识发现的异构信息标准化

3.1 从资源发现到知识发现

对异构信息进行标准化的过程体现了从资源发现到知识发现的转变，对异构信息的处理，在面向资源发现和知识发现中有显著的不同，具体见表1。基于文献的资源发现已经较为成熟，2015年7月多所高校开始试用的维普智立方·发现系统[18]是资源发现产业化的代表之一，而基于文献的知识发现尚没有较为通用的模式或案例。

表1 异构信息标准化在面向资源发现与面向知识发现的文献综合集成中的区别

项目	面向资源发现	面向知识发现的文献综合集成
本质	文献信息的“物理”整合	文献信息的“化学”整合。对原始知识进行拆分整合，产生新知识
目的	将广泛分布的信息进行整合，便于检索或提供给科研人员关于学科发展态势等的预测	对某一研究主题的文献进行内容拆分，提取不同文献的研究结论进行统计学意义上的集成，得出具有一定可信度的结论，以指导后续研究
方法	对文献中涉及的各类知识对象做唯一标识、粒度分析、关联呈现，实现从情报分析视角对隐含知识关联做深入挖掘	将原始数据在携带语义的基础上进行“结构化描述”，解决数据的表述问题，使机器理解数据的结构信息，以支持语义检索、机器识别和资源的细粒化等。目前主要通过人工处理语义信息，用户参与，灵活处理异构数据
对象	文献计量角度的基本信息	通常是大型数据库或数据仓库，广义角度也可以是文件系统，或其他数据集合，如网络信息资源、知识库等。本文主要指网络文献资源，并深入到研究内容、研究结论等
信息检索方式	基于“关键词/检索词匹配”找出馆藏的资源，具体知识需要读者自己习得，依赖于读者个人的科研素养	根据读者的问题，找出解决方案，检索结果通过知识有序组织和描述（提供结果或知识线索），帮助读者解决科研和学习问题，提高科研效率

知识发现的对象通常是大型数据库或数据仓库，广义角度也可以是文件系统，或其他数据集合，如网络信息资源、知识库等。本文所指的知识发现是从方法论意义上的广义知识发现，即关注于整个知识发现活动的全部过程和基本规律的综合体系，重点关注知识发现初期，多源异构数据的处理过程。现阶段，笔者主要通过人工处理一系列语义信息，由具体用户参与，使得原始数据可在标准化过程中体现更多的灵活性。

3.2 面向知识发现的异构信息标准化处理方式

通过对比面向资源发现与面向知识发现的文献集成中异构信息标准化的区别，本节旨在提出通用或规范的异构信息标准化处理方式。整体来说，一方面要

格式化提取异构信息,另一方面,对于格式化提取后的数据仍需进行二次规范化,将语义进行统一,以便于统计分析。

3.2.1 异构信息格式化提取 伴随着集成研究内容的不同,知识提取的模板有所不同,加入空间分析、统计分析数据时,新的模板应满足不同学科的研究人员的需求,此时应允许用户自定义工作模板,对信息进行标准化提取。信息提取的主要功能[19]是从文本中抽取出特定的事实信息。被抽取出来的信息通常以结构化的形式存入数据库中,供用户查询或进一步分析使用,信息提取往往需要通过对文本中的句子以及篇章进行分析处理后才能得到。信息提取系统将数据格式化后,可通过信息检索获得相关的信息。信息标准化模式如表2所示:

表2 异构信息标准化提取模式

指标类型	信息提取					
	基本特征指标	基本定性指标	影响因素特征指标		研究主体特征指标	
细分指标	编号	学科分类	时间段	年份	时间段	年份
	题名	研究区域		季节		季节
	作者	研究时间	主因素1	细分因素1	主体指标值	
	关键词	研究方法		细分因素2	主体指标辅助值	细分指标值1
	发表时间	出版物类型		细分因素3		细分指标值2
	期刊名	数据来源	主因素2	细分因素4		细分指标值3
	单位	其他		细分因素5		细分指标值4
	摘要			细分因素6		……
	其他		其他		其他	
说明	自动提取。用于文献计量学统计,系统自动生成	用户提取。其中学科分类、出版物类型、数据来源在元数据表中选取。研究区域、时间、方法规范化提取	关系型提取。关系型的主体与客体由用户自行确定,可添加删除影响因素、影响主体等		数值型提取。注意数据是有范围的,而非准确数值,需要两个字段限定	

对不同学科文献进行Meta分析时,考虑到领域文献各自的特点——文献普遍具有题名、摘要、关键词、发表时间、期刊名、作者等,可对这些基本信息从文献计量学角度进行集成。

基本定性指标主要是在包含一定属性的元数据表格中进行选取,只做定性统计。其中,研究方法是指如观察性研究还是试验性研究等方法学特性,研究时间一般指研究内容涵盖的全部时间范围。

对于非结构化的文献结论,笔者初步将其分为关系型数据与数值型数据,即影响因素特征指标列及研究主体特征指标列。

(1)影响因素特征指标是指所研究对象的影响因素,有些研究会对不同时间段的研究主体进行分析,从而得出不同结论,因而在对关系型或数值型信息进行提取时对不同时间段的研究结论分别提取。

(2)研究主体特征指标是指如研究对象的长宽高等基本特征、一系列标准的选择等;由于Meta分析的效应指标有的可以直接从文献中获取,有的需要经过对文献中的数据进行计算后获得,故设定研究主体指标辅助值,由用户自定义辅助值计算方式。

3.2.2 二次规范化处理 二次规范化处理是指针对提取的信息进行数据值、数据单位的规范化。属性一般分为三部分:属性名、属性范围、单位。对于准确的属性值,其属性范围下限与上限相等即可。二次规范也可以说是二次调整,即将每一项研究拆分成细分结论,将各属性值统一单位,将不同时间段的结论按照一定顺序排列(一般按照时间增序排列)。示例见图1。

4 异构信息标准化实例研究

笔者拟以构建资源环境学科文献的综合集成平台为例,探索具体领域异构信息的标准化处理模式。已有研究解决了部分理论难题,如提取数据的基础信息和空间信息等,笔者主要针对异构数据标准化处理。尽管目前可对文献的基本知识点进行人工提取和集成,但如何标准化提取处理数据仍然是困扰进一步研究的难点所在。由于目前数据采集过程中存在多个异构数据源,并且不同数据源的数据模式不同,因此源数据和目标数据在结构上也会存在不同。在进行数据集成时,首先应将数据以整齐的格式描述出来。

4.1 资源环境学科数据的异构性

资源环境领域空间和时间的二维性是地理现象的两个基本特征,资源环境领域异构数据集成的关键技术是如何用统一的数据模式描述各个数据源中的数据,屏蔽平台、数据结构等异构性,实现数据的无缝集

研究	时间	因素		
		最小值	最大值	单位
研究1	1950–1960	1.6	1.9	mm
研究2	1977–1979	1.4	1.8	mm
	1980–1990	1.3	1.6	mm
	1990–2000	1.8	2	mm
研究3	2000–2005	0.14	0.16	cm
	2006–2010	0.18	0.18	cm
研究 4	2008–2008	1.7	1.7	mm
研究 5	1988–1998	0.16	0.17	cm
研究 6	1972–1975	1.5	1.7	mm
……				

单项研究拆成细分结论，统一属性单位，将研究时间段按一定顺序排列

结论	时间	因素(mm)	
		最小值	最大值
结论1	1950–1960	1.6	1.9
结论2	1972–1975	1.5	1.7
结论3	1977–1979	1.4	1.8
结论4	1980–1990	1.3	1.6
结论5	1988–1998	1.6	1.7
结论6	1990–2000	1.8	2
结论7	2000–2005	1.4	1.6
结论8	2006–2010	1.8	1.8
结论9	2008–2008	1.7	1.7
……			

图 1 二次规范化示例

注: 如某一项研究结论为“某一年的指标值是一个具体的固定值”，而非给出范围，此时需将时间范围年份起始值设为相同，研究因素最大值与最小值设为同一值，如研究 4 的意义为“2008 年的具体指标值是 1.7mm”

成和对接。数据的异构性体现为不同层次，资源环境信息系统集成面临 4 个层次的异构性: ①系统异构。指操作系统和硬件不兼容。②语法异构。指不同的语义表述和数据表示。在资源环境领域，数据按结构类型可分为结构化数据和非结构化数据。结构化数据主要包括二维关系表; 非结构化数据主要涉及文本文件、XML 文件、Excel 表格和图形图像文件等。③结构异构。结构异构是指采用不同的数据模式表示相同的数据含义。具体包括两种表现形式: 数据格式异构和数据表之间复杂的关联关系。数据格式异构，也指数据表现形式的差别，如文本的数据表示和经纬度的数据表达等。④语义异构。语义异构是指不同系统中对数据的意义解释不一致。语义异构通常分为内涵异构和外延异构两个层次，指词语和概念在不同的上下文中有不同的含义，不同领域人员用不同的词语表示同一个含义，或者同一个词语表示不同的含义[20]，采用不同的结构来表示相同或相似的信息[21]，各种数据源中的概念之间存在着多种联系，但由于各个数据源的分布、自治性使得这种隐含的联系不能形式化地体现出来。

4.2 资源环境学科知识本体选择

近年来，异构数据集成研究的焦点已从解决数据语法、系统层面的异构逐渐发展到致力于解决数据语义异构问题。为此，研究人员在数据集成领域引入了“本体”这种能表达较强概念语义的工具，提出了基于本体的异构数据集成方法。本体是共享概念模型的形式化规范说明[22]，它通过对概念和概念之间关系的形式化定义来确定概念的精确含义，表示共同认可的、可共享的知识，使不同团体或组织的人都能对要表达的信息达成一致的理解，从而解决数据语义异构的问题。因此基于本体的异构数据集成是当前数据集成的重要发展方向之一。其中，混合本体方法综合了单一本体方法和多本体方法，每个数据源都有各自的本体描述以及各自的局部本体，是较为灵活的本体，此外也拥有一个全局本体作为共享词汇表，该方法的优点是增添、更新数据源时不需要对映射和共享的词汇表做过多的改动。

知识提取的前提是系统中存在一个完整的学科知识体系。在拟构建的文献综合集成系统中需加入一个学科知识本体，至少包含资源环境学科所有的二级学科，用以界定学科分类。当用户建立一个任务时，选择研究主题所在的学科体系，系统通过自动嵌入全部学科的目录树，供用户选择其任务的具体位置。这样，用户完成的每项研究任务都可在系统中相应的位置储存下来，使得全部用户的每项研究都成为平台积累的一项成果，为后人做相应研究提供指导。由于目前国际上并不存在通用的较完善的学科知识本体，笔者拟采用中国科学院兰州文献情报中心及中国科学院资源环境科学信息中心联合构建的“资源环境学科知识本体”，规范语法语义的异构问题。

4.3 资源环境领域数据标准化

在数据标准化处理过程中，由于不同研究中数据类型往往不一样，需要针对具体情况设立相对指标，进行标准化。资源环境领域数据标准化过程涉及以下几个方面:

4.3.1 地理空间标准化 资源环境学科不同研究区域的具体范围和表示方法不同。有些研究采用地理名称标识研究范围，这时可统一转换成经纬度; 地图中的不同点往往代表具体研究区域，如果研究区域较小，反过来也可采用区域名称标识。因此，是通过经纬度统一标准化处理位置信息，还是用地理名称标识常见的地理区域，此时需要一个较为完善的地名词典覆盖不同研究区域，需要依据具体情况处理。

鉴于地理空间元数据的重要性，国际社会制定了一系列的元数据标准来规范地理空间数据的内容和形式[23]。主要的标准有美国联邦地理数据委员会(FG–

DC)制定的CSDGM[24]、国际标准化组织推荐的ISO/TC211元数据标准、美国NASA开发的目录交换格式标准DIF。我国于20世纪90年代开始地理空间元数据的研究，各部门和研究机构也都制定了一些标准，代表性的有原国家测绘局颁布的《基础地理信息数字产品元数据》、军队颁布的《军用数字地图产品元数据要求》以及国家标准化管理委员会颁布的《地理信息元数据》等。其中FGDC地理空间元数据内容标准的目的是确定一个描述数字地理空间数据的术语及其定义集合，包括需要的数据元素、复合元素及其定义和域值以及描述数字地理空间数据集的元数据信息内容，笔者拟借鉴该标准草案，指导研究人员标准化地提取地理信息。

4.3.2 时间单位标准化 资源环境学科不同研究的时间单位、阈值等不一，需要进行统一。如速度的单位是长度单位和时间单位的合成单位，常用的有米/秒(m/s)、千米/小时(码)(km/h)、英里/小时(mile per hour)等，国际单位制中采用米/秒(m/s)，因此需将所有表示速度的单位统一换算成国际标准单位。系统内部需嵌入常用单位的换算标准并自动转换，对于系统未纳入的单位，允许用户个性化添加更新。笔者希望为集成研究人员建立一个体系化的平台，平台系统化地辅助用户实现具体功能，集成研究的初始步骤就是对异构数据的研究时间进行标准化。无论是时间单位还是空间属性等，数据标准化时的基本原则是依照国际标准，提供不同单位之间的转化率并允许用户扩增，因此该平台是动态可自操作的开发平台。

国际单位制是国际计量大会(CGPM)采纳和推荐的是一种一贯单位制，将单位分为基本单位、导出单位(含辅助单位)。据此，系统嵌入的资源环境学科的单位换算规范(部分)如表3所示：

表3 文献综合集成系统嵌入的资源环境学科单位换算规范(部分)

单位分类	物理量名称	单位名称	单位符号	基本换算情况
基本单位	长度	米	m	
	质量	千克(公斤)	kg	
	时间	秒	s	
	电流	安培	A	
	热力学温度	开尔文	K	
	发光强度	坎德拉	cd	
	物质的量	摩尔	mol	
导出单位	力	牛顿	N	$1N=1Kg\cdot m/s^2$
	压强	帕斯卡	Pa	$1Pa=1N/m^2$
	光照度	勒克斯	lx	$1lx=1lm/m^2$
	温度	摄氏度	℃	$1℃=K-273.15$
	时间	年	a	
	时间	小时	h	
	降水量	毫米	mm	

4.3.3 属性提取标准化

(1)基本定性指标中的研究时间，时间范围为1900年至2016年，由用户选取研究内容所属的年份，基本时间单位为年。

(2)基本定性指标中的研究区域，以经纬度两个字段进行提取，同时支持地名提取，一些小型区域可用地名代替，大范围区域则通过经纬度限定，即字段：经度、纬度、地名。针对一个研究区域，存在四组经纬度的点，以界定矩形研究区域。

(3)影响因素特征指标及研究主体特征指标中，嵌入基本属性，如温度、湿度、高度等，以下拉列表的形式供用户选择，加入允许用户自定义的接口。

5 结语

通过对资源环境学科异构信息源的标准化，将原始知识拆分集成后产生新知识，该知识发现过程的愿景是研究人员输入研究主题或任何一个相关概念，能够得到该主题相关的所有文献(科研成果)，根据读者研究需要进行个性化的信息提取，系统自动对提取的结构化信息进行拆分整合、分类整理，同时对集成的结果进行多维度分析，得到相关研究内容的知识图谱或知识脉络以及地理信息系统的可视化展示，对研究主题相关的知识点提供深入、准确的分析，扩大思维范围，并对关联的知识或文献进行有序组织，从而得到多主题的最新科研情报。

对异构信息进行标准化处理仅仅是知识发现过程的起点，笔者以团队自建的资源环境学科知识本体为依据，通过对资源环境学科异构信息在地理空间、时间单位及属性提取中的标准化分析，提出异构信息源标准化处理的思路，指导搭建实现信息集成、支持综合集成的人机交互的文献综合集成平台。后续笔者将重点关注标准化的信息统计分析及可视化展示，完整实现文献综合集成的知识发现过程。

参考文献：

[1] 邱均平，李小涛，董克. 图情领域可视化研究的发展、演化与创新[J]. 图书情报工作，2014，58(13)：125－131.

[2] SWANSON D R. Undiscovered public knowledge[J]. The library quarterly，1986，56(2)：103－118.

[3] LI D，HAN J，SHI X，et al. Knowledge representation and discovery based on linguistic atoms[J]. Knowledge－based systems，1998，10(7)：431－440.

[4] 波普尔. 猜想与反驳——科学知识的增长[M]. 傅季重，纪树立，周昌忠，等译. 上海：上海译文出版社，1986.

[5] 荣毅虹，梁战平. 基于文献的发现[J]. 情报学报，2002，21

(4):386-390.
[6] 崔瑞琴,孟连生. 数字信息资源整合问题研究[J]. 图书情报工作,2007,51(7):35-37.
[7] 牛奉高. 数字文献资源高维聚合模型研究[D]. 武汉:武汉大学,2014.
[8] 窦天芳,姜爱蓉,张成昱,等. WEB环境下多源数据的集成服务——以清华大学新期刊导航为例[J]. 大学图书馆学报,2010(3):80-84.
[9] 肖希明,田蓉. 国外公共数字文化资源整合的现状与发展趋势[J]. 国家图书馆学刊,2014(5):48-56.
[10] CALISTRU C, RIBEIRO C, DAVID G. Multimedia in cultural heritage manuscripts: integrating description, transcription, and image content[J]. Eurasip Journal on image & video processing, 2009, 7238(3):347-386.
[11] HERNÁNDEX F, WERT C, RECIO I, et al. XML for libraries, archives, and museums: the project COVAX[J]. Applied artificial intelligence, 2003, 17(8/9): 797-816.
[12] RAYWARD W B. Electronic information and the functional integration of libraries, museums, and archives[M]. E Higgs History & Electronic Artefacts. Oxford: Clarendon Press, 1998:207-226.
[13] 马文峰,杜小勇,卢晓惠. 基于知识的资源整合[J]. 情报资料工作,2007(1):51-56.
[14] Simfinder: a flexible clustering tool for summarization[EB/OL]. [2016-02-22]. https://www0.comp.nus.edu/~kanmy/papers/simfinder.pdf.
[15] 赵新勇. 基于多源异构数据的高速公路交通安全评估方法[D]. 哈尔滨:哈尔滨工业大学,2013.
[16] 刘亚东,彭舰,张达平. 基于智能的网页信息提取系统的研究与设计[J]. 四川大学学报(自然科学版),2009(4):957-962.
[17] 郑建明. 数字文献资源的整合与服务——以江苏省高校文献资源保障体系建设为原型的个案研究[J]. 大学图书馆学报,2007(5):6-9.
[18] 智立方. 知识发现系统[EB/OL]. [2016-02-28]. http://zlf.cqvip.com/help/about.html.
[19] 廖崇粮. Web信息自动抽取技术的研究[D]. 成都:电子科技大学,2012.
[20] 和延立,杨海成,何卫平,等. 信息集成与知识集成[J]. 计算机工程与应用,2003(4):38-41.
[21] BUCCELLA A, CECHICH A. An ontology approach to data integration[J]. Journal of Computer Science and Technology, 2003, 3(2):62-68.
[22] STUDER R, BENJAMINS V R, FENSEL D. Knowledge engineering: principles and methods[J]. Data and knowledge engineering, 1998, 25(1/2):161-197.
[23] 徐少坤. 地理空间元数据可视化研究与实践[D]. 郑州:解放军信息工程大学,2013.
[24] 柯青. 网络环境下异构信息检索标准体系研究[D]. 武汉:武汉大学,2004.

作者贡献说明:
曲建升:提出研究命题及思路,修订最终版本;
刘红煦:设计研究方案,撰写论文。

The Standardization of Heterogeneous Information in Knowledge Discovery: A Case Study of Resources and Environment Literature

Qu Jiansheng[1] Liu Hongxu[1,2]

[1]Lanzhou Information Center, Chinese Academy of Sciences, Lanzhou 730000

[2]University of Chinese Academy of Sciences, Beijing 100049

Abstract: **[Purpose/significance]** This paper aims to research the integration of heterogeneous information in the process of data integration. Based on the accuracy requirements of literature information extraction in literature Meta-synthesis system, we proposed a heterogeneous information standardized approach by the case of resource and environment subject. **[Method/process]** Using the self-discipline ontology of resources and environment subject, we put forward the idea of heterogeneous information standardization process to guide to putting up literature comprehensive integration platform supporting comprehensive integration of human-computer interaction and information integration, through the analysis of standardizing of heterogeneous information in the geographic space, time and attribute extraction in resources and environment subject. **[Result/conclusion]** Finally, the paper realized the task of data preparation phase in literature Meta-synthesis, according to knowledge extraction and processing of documents from different sources in different data formats. The standardization of heterogeneous information is only the starting point of knowledge discovery process, and we will focus on the statistical analysis and visual display of standardized information to completely implement knowledge discovery process of literature meta-synthesis.

Keywords: knowledge discovery heterogeneous information standardization domain literature