DOI:10.12154/j.qbzlgz.2021.05.002

领域知识组织理论基础及方法分类简述*

曹思源[1,2]　马海云[1,2]
([1]南京大学信息管理学院　江苏　210023；[2]江苏省数据工程与知识服务重点实验室　南京　210023)

摘　要：[目的/意义]互联网环境下人们之所以能够快速便捷地从网络中获取自己需要的信息和知识，是因为所能检索到的信息和知识来源于组织有序的系统，而知识组织则是形成有序系统的关键。文章通过对知识组织方法的梳理和分类，明确了知识组织的核心任务以及其在情报工作乃至知识检索领域的重要作用，以期使得知识组织在大数据时代中仍旧能够得到情报工作人员的重视，得以继续发展。**[方法/过程]**梳理知识组织发展的脉络，依据功能的不同将知识组织方法分为基础知识架构类知识组织方法、分类体系构建类知识组织方法和关系网络建立类知识组织方法，叙述这些知识组织方法的理论基础，并阐释每类方法在大数据领域知识组织环境下的作用和功能。**[结论/结果]**为了在新环境下最大化地发挥其效用，领域知识组织的发展可以做出逻辑表达语言专业化、更加个性化并注重知识的实时更新、自动逻辑推理和人机交互，更加智能化等方面的调整，以便能够更好地迎接大数据时代的挑战。

关键词：领域知识组织　主题组织法　分类组织法　元数据　语义网　本体　关联数据

A Brief Introduction to the Theoretical Basis and Method Classification of Domain Knowledge Organization

Cao Siyuan[1,2]　Ma Haiyun[1,2]
([1] School of Information Management, Nanjing University, Jiangsu, 210023;
[2] Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023)

Abstract: [Purpose/significance] The reason why people can obtain the information and knowledge which they need quickly and conveniently in the Internet environment is that the information and knowledge they can retrieve come from an organized system, and knowledge organization is the key to forming an organized system. By sorting and classifying knowledge organization methods, this paper clarifies the core tasks of knowledge organization and its important role in information work and even in the field of knowledge retrieval. So that knowledge organization can still receive the attention of information workers and continue to develop in the newly arrived era of big data. **[Method/process]** We sort out the development of knowledge organization, classify knowledge organization methods into basic knowledge architecture organization method, classification system construction organization method and relationship network establishment organization method according to their functions. Then we describe the theoretical foundations of these knowledge organization methods, and explain the roles and functions of each method in the knowledge organization environment of big data. **[Conclusion/result]** In order to maximize its utility in the new environment, the development of domain knowledge organization can make adjustments in terms of specialization of logical expression language, more personalization and focus on real-time updating of knowledge, automatic logical reasoning and human-computer interaction, and more intelligence, so as to be able to better meet the challenges of the big data era.

Keywords: domain knowledge organization　subject organization method　classification organization method　metadata　semantic Web　ontology　linked data

* 本文系国家社会科学基金重点项目“大数据环境下领域知识加工与组织模式研究”(项目编号:20ATQ006)的研究成果之一。

1 引言

知识组织作为图书情报学领域的核心研究内容，是用户进行信息检索、资源开发利用的基础，无序、混杂、分散的信息不仅不利于检索，而且所能提供的价值是大打折扣的，更难以被人们转化为结构化的经验，进而也难以被转化为知识。所以，根据一定的科学原理和规则对信息、知识进行包括整理、提取、排列和序化在内的组织，使之形成一个有序的、结构化的系统，才能在最大程度上提升知识的利用效率和价值。因此，对知识的组织可以被视为是一个知识增值的过程。回顾人类整理利用信息资源的历程[1]，从最初的对文献资源的物理管理即对纸质文献的分类、排架；到对主题有了认识并加以提取，编纂不同领域内的主题词表，处理细粒度细化到了信息点、知识单元；再到注重信息、知识之间的关联，构建具有复杂关系的网络；手工操作转向机器操作，人工理解推理转变为计算机理解并推理，知识组织随着载体的改变、认知的提升和技术的进步不断发展，也越来越智能化。

本文以各种知识组织方法功能上的差异为区分标准，将现阶段知识组织领域的知识组织方法整体上划分为基础知识架构类知识组织方法、分类体系构建类知识组织方法和关系网络建立类知识组织方法三大类。首先阐释这些知识组织方法的理论基础，然后对其功能和在领域知识组织中的应用进行简述。同时在对这些方法进行分类叙述的基础上构建领域知识组织系统模型，自底向上把系统的构建划分为语料预处理及术语抽取、知识关联和可视及系统化三个阶段。

其中基础知识架构类的方法作为知识组织领域内的基础架构，起到了提供知识来源和知识释义的作用，作为一种知识组织方式，它不仅可以标引知识，用户也可将其作为知识来源工具和检索浏览工具，用户在检索（词表）时，不仅可以获取与检索目的直接相关的知识，还能提取到与检索目的间接相关联的知识。第二类分类体系构建类方法是建立并提供了一个完整的、逻辑层次分明的分类体系，用户可浏览分类体系所反映出来的类表，发现知识的上下位从属关系。第三类的关系网络建立型方法其要素在于关联，以语义为核心和桥梁，将知识关联起来，破除知识孤岛的弊端，让隐性知识显性化，构造领域内的知识结构。

2 知识组织的理论基础

2.1 符号学理论

语言作为一种符号系统是知识组织系统的基础，无论是自然语言还是人工语言，有了这种符号才能进一步去构建功能完备的知识组织系统。以主题知识组织系统和分类知识组织系统为例，这两种组织系统下的符号语言分别为主题语言和分类语言即分类号。在学者Pierce对符号理论的论述中，存在着符号（sign）、对象/客体（object）和代表物（interpretant）这三个内在相互作用的部分，其中符号可视为一种指代，对象或者是客体表明了被指向的客观存在，代表物则是建立了人所理解的符号与对象之间的联系[2]。在这种语境下的符号是客观实体与其代表的意义之间的桥梁，符号也就是通过“代表意义”这样的一种方式与实体之间建立了联系[3]。此外，符号学中的三个层次——语法、语义和语用有机地把知识组织系统的结构串联了起来。第一层次的语法（syntax）研究词或句子的特征，这一层次表征了知识内容的基本单元，这些基本单元构筑了它们所代表的实体，例如在主题知识组织系统中体现为合乎主题词遴选规则的单个条目以及单条目之间的组配规范，在分类系统中体现为从宽泛到细化的类目层次划分依据和排列规则。第二层次的语义（semantics）表示概念的意义，集中关注概念的意图、是怎样与实体进行关联的并且同时注重该概念下的特定实体或项目，语义层次的构筑建立了对基本知识内容单元的解释，并将基本单元和其特征、属性进行映射。例如在主题组织法中，主题词本身和其属性项、族项被对应起来，而且正式主题词也会参见非正式的主题词或者是相关主题词；在分类组织法中，这样的一种映射即是把分类号和实体进行联系，声明该分类号指代的是哪个实体。第三层次的语用论（pragmatics）关注概念是如何在使用的过程中被测定、其意义是怎样被取决的，这一层次的含义不仅仅局限于语法和语义，它突出表现了该概念是如何在具有上下文的情境下被用作交流的，主题法中的上位词、下位词、“参见”参照、“见自”参照和族性项目等都突出了上下文语境下符号的语用。

2.2 分类学理论

分类作为一种基本的逻辑思维形式，是人们认识事物的一种基本方法，分类学直接为分类组织法提供了思想理论基础。知识分类体系是建立在人类对外部世界的探索和发现的基础之上的[4]。分类将事物或实体按照其特征和属性加以区分和类聚，并将区分的结

果按照某种特定的次序进行组织序化，并将区分所得到的结果赋以特定的、规范的、体系化的标识符即分类语言。分类的过程首先是依照事物的主要属性和特征进行区分，首次分类所得到的类目下集中了大量具有该属性的事物，其次在具有该属性的集群中再依照次一级的属性进行区分，依此类推，直到无法再进行细化。经过这样一种层级式的划分可得到粒度由粗到细的树形结构，这样的一种结构符合人们逻辑思维和认识新事物的习惯，既保证了逻辑的连贯性，又便于理解。分类这样的一种思想，不仅展现了每个独立的信息单元，而且也将其和其他信息单元以所有和归属这种层级方式联结了起来，逐层分类展示了各个知识单元之间的逻辑联系，也是根据族性原则对知识进行划分和处理的依据。

2.3 复杂网络分析理论

知识并不是孤立存在的，科学是一个统一的内在的整体，正是由于知识之间的联系和交叉，学科的知识结构才得以存在、人类社会的知识体系才得以不断被完善、不断被扩充。行为科学中的复杂网络分析理论为这种联结式的关系分析提供了洞见。复杂网络分析关注真实世界中的个体、个体之间的关联以及这些关联的模式和此种关联所带来的影响。知识组织系统中的知识单元对应到复杂网络中的个体，关系(relation)则代表个体与个体之间的联结(linkage)也就是知识之间的关联，复杂网络节点、边以及结构的复杂性会产生不同的影响和作用[5]。Price[6] 在 *Little Science, Big Science* 一书中也曾用科学分子来表示科学知识单元，他认为应该用科学的、严谨的方式看待科学分子之间的相互作用，并以科学的方法研究科学结构和科学发展动态[7]。关联的分析使得分析的层面不再局限于个体上，上升至社群层面的分析摆脱了“局部知识”的局限，在一定程度上缓解了个人知识的非对称性和差异性[8]。同样，作为复杂网络分析数学基础之一的图论为网络图的呈现和系统的可视化呈现提供了理论背景。图论将科学知识节点和知识的联系即边以严密且具有逻辑的数学方式表示出来，并以严密的数学化形式分析网络的拓扑结构，能以高效的方式分析复杂网络的基本属性、提升预测方法的性能[5]。此外，图论还为网络拓扑结构的可视化提供了理论基础，可视化把知识单元和单元之间的联系以一种符合人类视觉习惯的方式呈现出来，图论根据关联来对对象按照数学结构进行模式上的配对，基本要素为点(nodes/vertices/points)和联结点的边(edges/links/lines)，采用数学的方法，将各种类型的关系进行建模和映射，反映真实的世界和知识系统。

3 知识组织方法分类

在学者Hodge对知识组织系统的整体叙述中，他根据知识组织系统的特性例如结构、复杂性、词/术语之间的关系和历史作用的不同，将知识组织系统划分为了Term Lists、Classification and Categories 和 Relationship Lists 三大类[9]。以学者Hodge的划分方法为启发，本文拟从基础知识架构、分类体系构建和关系网络建立三个角度来对知识组织方法进行分类简述，如图1所示。

3.1 基础知识架构类知识组织方法

基础知识架构类的知识组织方法是组织领域知识的基础，一方面，这类方法承担着知识来源的作用，能够在知识服务系统中为用户提供知识释义[10]；另一方面，它们有着系统化的架构，能对领域知识进行有序化的梳理。这类组织方法通常包括主题组织法和元数据组织法。主题词法是以能够反映文献或知识主题内容的主题词为组织体系，并将主题词按照一定逻辑顺序规则所排列的组织方法；元数据组织方法则是对信息、知识进行的再描述，是关于数据的数据、关于信息的信息、关于知识的知识，元数据亦被称为描述记录，是通过对知识进行不同侧面的细化描述，并将这些对于知识的再描述数据按照一定逻辑准则排列，形成一个结构完整的组织体系；元数据一词，开始主要指网络资源的描述数据，用于网络信息资源的组织，随后，其逐步扩大到各种以电子资源形式存在的信息资源的描述数据[11]。这些方法所对应的一系列知识组织工具主要有领域主题词表和元数据方案等，它们往往是由该领域

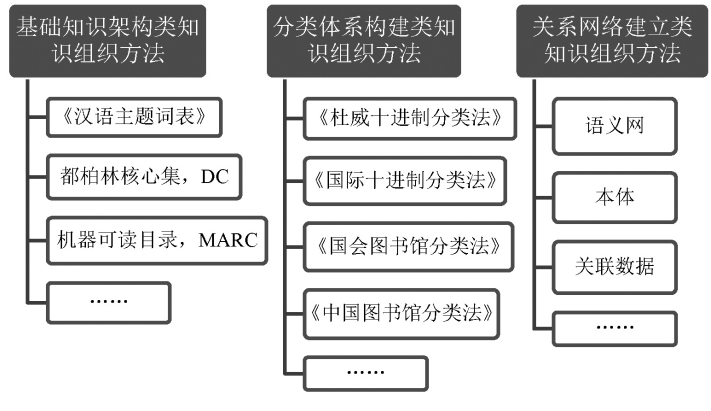

图1　知识组织方法分类及架构

的专家经过商讨、严格的筛选而编纂制订出来的，所反映的是已经成熟稳定且标引程度高、质量高的信息和知识，无论是用户或者是知识组织系统的构建者还是图书情报工作人员均可在查询条目时获取到相应的知识释义，明确此知识条目的外延和内涵。

以《汉语主题词表》为例，它的条目中列出了该叙词的上位类和参照词，用户在通过字顺查找到能够清晰表达自己检索意愿的叙词后，可以清晰地了解到该叙词的外延和内涵以及它的同义词，不仅提高了信息检索过程中的查全率和查准率，还完善了用户的知识结构。元数据以都柏林核心集（Dublin Core Element Set, Dublin Core）的结构为例，15个基本的数据单元完成了对领域知识的描述、定位、搜寻、评价和选择的一系列完整处理与控制，将领域知识以结构化、规范化的方式呈现出来，不仅有助于将领域知识规范为结构分明的体系，而且也为后续检索、开发利用领域知识打下了基础。与都柏林核心集相似的还有机器可读目录（Machine Readable Catalogue, MARC），MARC把文献的不同特征用不同的字段来表示，并赋予每个字段以独特的编号即标识符，不仅格式伸缩性强、适应面广泛，而且著录内容详尽，可检索的字段数量多，能够被计算机所理解。

基础知识架构类这样一种系统化、结构化的组织方法在知识组织演化过程中起到突破性作用的一点就是它们把对知识进行组织的细粒度细化到了知识单元，不再局限于以篇为单位的文献组织，对知识的组织深化到了某一类事物、某个词，可直接获得细节上的知识，不再需要人为重新阅读文献并从文献中总结出想要的知识。这样的一种细化在知识组织的个体维度上完善了以知识基因、作品（work）和知识元为基础的知识单元理论[8]。传统意义上的主题概念不再是知识组织的唯一标准，任何一个信息点、知识点都可以作为组织点，成为检索的入口。

作为传统的知识组织手段，这类知识组织方法在保证了其核心功能和价值的基础上与现代的知识组织自动化技术相结合，术语的自动抽取[12]和标引[13]、基于机器学习[14]和知识分类体系的自动分类以及网络环境下主题词表[15]和分类树[16]的应用等都为传统的知识组织工具带来了新的发展，在不同领域的知识环境下，结合领域的百科全书，构建领域词典和领域内知识分类体系，这些新的发展都使得来源类知识组织工具在保有本身高质量内容、严密的逻辑体系和高度结构化特点的基础上更广泛的为用户所接触，发挥了更大的价值。

3.2 分类体系构建类知识组织方法

分类指依据实体的属性或特征对实体进行区分和类聚，并将区分所得的结果按照一定次序进行排列、序化和组织的行为。分类包含了分类和归类两个概念，分类是根据事物的属性和特征对事物加以区分，建立类别体系；对于这样一个分类体系，任何事物都可以在找到一个属于自己的类目，这就是归类的过程。分类体系构建类知识组织法是从内容的角度来组织和揭示知识，它的实施也是一个分类的过程，首先根据知识所固有的属性包括外在和内在特征，对知识进行区分和类聚，建立一个层次分明的知识分类体系；然后再根据这个知识分类体系，找到适合每一个知识单元的类目，赋予他们分类代码和相应的语词形式的类别标识（即用各种分类号作为基本标识来表达信息概念）[17]，并根据不同类别的分类代码的某种次序进行排列组织。分类体系的构建目的在于为用户导航信息资源和领域知识，将分散的、扁平化的知识组织成符合人类认知习惯的树形层级结构，不仅能够把知识细化了，也清晰地表现了知识的上下位的从属关系，为用户提供了领域知识集合的全景结构图[18]，便于用户从族性的角度认知领域科学知识并提升用户在浏览获取知识时的体验。

典型的分类体系构建类知识组织方法有着严密的逻辑层次结构，按照知识内容特征的相互关系进行组织，并采用一定的标记符号作为排序工具，《杜威十进制分类法》（*Dewey Decimal Classification*, DDC）、《国际十进制分类法》（*Universal Decimal Classification*, UDC）、《国会图书馆分类法》（*Library of Congress Classification*, LCC）和《中国图书馆分类法》这几种分类法便是分类体系构建类方法在知识组织中的主要体现。它们按照用户从学科的角度来检索文献的习惯，首先将人类的知识体系按照学科划分为若干大类，并赋予相应的标识符，之后再在每个大类下进行细分，层层展开，最终反映为科学性和实用性结合的类表。网络环境下的知识组织主要体现为将传统文献分类法网络化的学术资源分类和利用分类思想的自编分类系统，以分类目录著称的雅虎目录（Yahoo! Directory）在构建分类体系的过程中并没有以传统的分类法为参照，更多以事物为核心，直接面向对象，以便于理解的语词而不是较难理解的分类号来作为类目的代表，同样也形成了一个由类目、子类目构成的可浏览的目录索引等级结构。

3.3 关系网络建立类知识组织方法

知识量的大幅增长、实体之间的关联日益复杂（实体关联类型膨胀）[19]让知识之间的关联变得不可忽视起

来，关系网络建立型的知识组织方法通过把孤立的数据和信息关联起来以达到知识呈现的目的，关联是反映知识实体与实体之间的语义关系。在这类知识组织方法中，每个实体都需要被定义，并拥有明确的概念，概念之间的关系必须明晰，整体是一个完备的逻辑系统，有着自己的应用规则[20]，用户可以从一个实体出发，顺着链路查询到与出发点相关的实体，也可以从关系入手，得到具有完备语义关联的知识。包括语义网(Semantic Web)、本体(Ontology)、主题地图(Topic Map)以及关联开放数据(Linked Open Data)等在内的、新型的以关系网络为基础的组织方法的重点在于强化知识的内容特征，是基于概念内容的，能够整合不同来源的信息资源和知识[21]，并实现知识的连通和共享。更近一步，关系网的构建也可与深度学习、向量机和聚类算法等结合，跨越领域的壁垒，不仅将领域内的知识进行连接，而且也将不同知识组织系统之间的知识进行连接，把分布式的资源集中起来利用，最终采用图形界面来呈现出可视化的结果，在面向语义的同时也更多地面向知识概念本身。

关系网络构建型的知识组织方式能够灵活地与传统的知识组织手段相结合，以主题地图(Topic Map)为例，三要素之一的topic即主题，它的表示方式为受控的标引词，受控标引词的来源为领域的主题词表或领域词典；经过关系化Association，主题之间的关联被显式地表达了出来，加以主题地图采用图形技术，主题、关系和链接的信息资源被表示为某个领域范围内的知识结构，用户不仅能够直接浏览整体的知识结构[21]，更能在浏览知识结构的基础之上进行知识单元的检索。

语义网(Semantic Web)与主题地图类似的地方在于其同样叙述了实体的概念和关联，但语义网的创造性的特点在于它更注重语义，核心在于为包括知识在内的实体和它们之间的联系添加能够被计算机所理解的元素即元数据，在计算机充分理解语义和逻辑关系的基础上，人与计算机之间才能实现无缝交互，对领域知识的自动组织才有可能成为现实，从而将一个个现存的信息孤岛发展成一个巨大的数据库。本体(Ontology)这个来源于哲学领域的概念在当今开放语义的网络环境下，突破了树形结构固有的类/族的概念，让知识之间的关联不局限于等级层次、隶属关系，而是形成了跨等级层次、跨分支的多元多维关系[22]，本体同样也可和主题词表[23]、分类体系[24]和概念格[25]相结合，形成领域本体，将异构的数据进行整合，实现资源和概念的共享，在连接和共享的过程中甚至有可能形成跨领域的宏观关联，即所谓的“超领域”和将知识关联扩展至单一领域本体之外的本体映射(Ontology Mapping)[26]。这样一种跨越领域式的关联从根本上改变了知识关联网络的结构，使网络的构建摆脱了传统的树形模式，关联种类更加丰富、概念更加明晰，同时也具备了推理能力，在使用上发挥了更大的潜力。

开放关联数据(Linked Open Data)以数据的语义为基础，连接互联网上不同的数据，这个不同可以是形式上的不同，也可以是内容上的不同，但所被连接起来的数据必定都是有语义上的关联，开放关联数据以统一的标识和框架规则来实现数据之间的关联[27]。关联数据的实现过程中必须借助资源描述框架(RDF)的方式，用唯一资源标识符(URI)来标识资源，以描述(Description)这一RDF核心要素表明特定URI的概念和URI与URI之间的关联。对关联数据较早采用应用的例子有美国国会图书馆将国会分类号和杜威十进制分类号进行了互关联[28]，在表面上看似只是两套不同分类系统之间人工控制语言即分类号的相互对照、连接，但本质上这是两种不同知识体系、不同认知论之间的关联，因为同一个实体，在此分类体系下所属的类目在另一分类体系中由于划分制订者不同的认知则是可能会属于一个完全不同的类目的，所以形式上的关联在关联数据这里也是本质上的关联。由于关联数据与资源描述框架相结合，所以关联数据的可视化可分为数据模式的可视化和数据的可视化两个层面[29]，而且也能为领域知识管理原型系统[30]和领域知识组织系统[31]的构建提供了思路。

对知识的提取是后续进行知识组织的首要步骤，传统的组织方式中，提取信息和知识并对其进行标引的往往是领域中的专家或是该领域中富有实践经验的研究员和图书情报人员的工作，一切的工作均由人工手动完成，当利用计算机对信息进行自动抽取时，自然语言处理工具必不可少，这些自然语言处理工具主要包括哈工大语言技术平台(Language Technology Platform)、中文自然语言处理工具包FudanNLP/FNLP和主要致力于英文自然语言处理的Natural Language Toolkit等。对文本进行分词、词义消歧，对命名实体进行识别和语词标注等具体的处理工作为后续语义的关联、资源的整合提供了内容来源和基础的结构框架。在对领域知识进行了识别和抽取之后，对知识进行分类或聚类也是构成知识组织系统框架的重要一部分，也是

认知知识的重要基础。与人工标引、分类有所不同，基于计算机的文本自动分类需要事先有确定的分类体系也就是有训练集的存在，用户或操作人员对输入数据格式进行控制，之后选择合适的特征即可进行分类工作[10]。目前能够实现分类功能分类器常用的算法有取类别样本平均值，以此来构造质心向量的Rocchio算法；基于概率统计的朴素贝叶斯算法和本质为解决二次线性规划问题的支持向量机（Support Vector Machine，SVM）模型。在对聚类工具的研究中，常用Cluto或Weka来进行高维和低维数据的聚类，并对所聚合到的类进行特征分析。聚类这一过程同分类不同的点在于，聚类组织事先是没有给定特定参照的，也就是没有固定的分类体系，是根据不同的算法对实体特征的识别来判断它们之间的相似程度，并把相似程度高的实体归为一类。如在医学领域的方面，Patterson等[32]就利用Cluto对临床记录的描述文本生成词项-文档矩阵，并进行聚类。对知识关系进行呈现也就是一定意义上的可视化对理清领域知识的结构和演化进程方面起到了重要作用，关联的可视化以一种直观的方式展示了知识之间的深层潜在联系。如CiteSpace根据引用、被引用和共现的关系将文献联系起来，结合Web of Science, CSSCI等引文索引进行可视化分析；Unicet支持对社会网络中的隐含知识进行分析，挖掘它们之间的深层次语义；同样面向语义网的、能够进行本体构建的工具Protégé不仅为领域本体的研究提供了技术手段，更能对知识进行处理进而面向知识服务[33-35]。

4 领域知识组织系统框架构建

4.1 语料的预处理和术语抽取

4.1.1 语料预处理

作为知识组织系统框架构建的基础，语料预处理是进行术语抽取的前提。语料库往往是领域内的文集，文集信息往往是非结构化文本，不同于西文文本以空格作为分隔，中文文本行文中不存在空格，一句话往往表达的是一个完整的意义，如果不进行预处理的话往往是不容易从非结构化的文本中提取出来主题和术语的。目前对语料的预处理方式主要包括分词和词性标注、语料的格式转换以及降噪等[12]对自然语言的处理。国外已有一些较为成熟和著名的自然语言处理平台，例如GATE(General Architecture for Text Engineering)、UIMA(Unstructured Information Management Architecture)和NLTK(Natural Language Toolkit)，这些工具注重体系的构建和系统的结构[36]，但缺乏对中文语言的分析。目前针对中文的分词技术主要包括基于词典匹配的算法和基于统计的算法。

基于词典匹配的分词算法是较为传统的一种技术，其原理可以概括为将待分析的语料与领域词典中的条目进行比较，也被称为基于规则匹配的分词技术。匹配算法可根据匹配词长短分为最大匹配法、最小匹配法、逐词匹配法和最优匹配法，以及根据匹配方向的不同分为正相匹配法、逆向匹配法和双向匹配法。由于词所能表达的信息量和词的长短呈正相关，所以基于词典匹配的分词技术大多采用最大匹配算法[37]。最大匹配算法以长词为优先选项[38]，首先选取词典中最长词的长度作为Max_Length，将其作为第一次取字数量的长度，然后在词典中进行扫描，当匹配到词典中的词时，便进行切分，若未匹配到，则将词串从右向左逐字递减，每递减一次，则再在词典中重新匹配一遍，循环往复，直至语料库中的所有语料都被切分完成。基于规则的分词方法依赖于领域内的高质量词典，即把领域词典作为规则来对所要切分的语料进行预判。虽然最大匹配的算法便于理解，实现起来较为容易，但它高度依赖于领域的高质量词典，并且在分词的过程中容易产生长词被忽略、匹配次数较多和歧义词串处理效率低等问题[39]，所以基于统计的分词算法被提出并更多地被用作后续的分词技术。基于统计的分词算法以统计学为基础，其核心主张是“词是稳定的组合”，所以成词的可信度是与字字之间相邻的概率呈正相关的，当字与字之间相邻的频率达到或超过了一定的阈值时，便可认为这些相邻的字可成词。常见的分词统计模型包括隐马尔可夫模型[40]（Hidden Markov Model, HMM）、条件随机场模型（Conditional Random Field, CRF）、最大熵模型（Maximum Entropy Model, MEM）和神经网络模型[41-42]。基于统计的算法依赖于语言训练模型，从而实现对未知文本的切分，其步骤可概括为首先对句子进行词切分并找出所有的分词结果，然后再对初步分词的结果运用统计学计算概率，最终找出概率最大的分词结果。

4.1.2 术语抽取

术语抽取是自然语言处理中一个重要的方面，在语义关联、本体构建和知识图谱建立的过程中都发挥着重要的作用，更是知识组织系统框架构建的关键一步。与分词方法的思想类似，对术语抽取的主流方法分为基于规则的抽取方法、基于统计的抽取方法和规则与统计相结合的抽取方法。基于规则的抽取方法同样需要领域词典的协助，归纳出词典中术语构成的规

则、提供语法模式列表，然后将语料库中已经经过分词处理的语料与领域词典进行匹配，并将匹配到的词语作为术语收入术语词表(Term List)中，对于未匹配成功的语料或是词典中没有的术语，则通过构建规则模版的方法来进行识别[43]。虽然基于规则的抽取方法由于有专业词典的辅助准确性较高，抽取出的术语可被用作为主题词，但该方法的局限性在于可移植性不高，不同领域的语料特点不同，所参照的规则也就不同，在制定规则时需要高质量的领域词典和较强的领域知识背景[44]。基于统计的术语抽取方法以术语在语料库中分布的统计特性为基础，建立统计模型，根据计算的结果来确定较为准确的种子词，然后在此之上进行扩展，来获取最终的术语[43]。常用的统计算法有词频(Frequency)、词频-逆文档频率(Term Frequency - Inverse Document Frequency, TF-IDF)[45]、领域相关性(Domain Relevance)和领域共识(Domain Consensus)、互信息(Mutual Information)和最大似然值等。相较于基于规则的抽取方法，基于统计的方法灵活性较高、适应性强，不局限于某一特定的领域，但是该方法与原始语料库的质量高低有着较强的线性关系，当语料库的质量较低时，抽取到术语的准确性还有待提高。

为了克服规则和统计这两种单一抽取方法的弱点，结合规则和统计的抽取方法先利用语法规则获取待检验的术语列表，然后再根据统计特性对术语列表进行筛选和过滤，这样在一定程度上既保持了较高的准确性因为参照了规范性较高的术语词典，又保持了统计抽取方法的灵活性和强适应性。一般较多采用C-value方法和NC-value方法，C-value方法综合考虑了术语候选词的长度、词频和嵌套信息，所以在长术语抽取的方面效果较好[46]；NC-value方法基于C-value方法，首先计算候选术语的C-value评分，形成“语境词列表”和其权重值，根据C-value评分和语境词列表权重值来综合计算候选词的NC-value值，由此来确定所要抽取的术语，NC-value方法在抽取高频术语的方面表现较好[46]。

4.2 知识关联

知识关联关系，抽象地说是知识概念之间的逻辑关系[47]。人类的知识系统是一个有机的、内部互相关联的整体系统，内部的知识并不是孤立存在的，也不是简单的堆积和罗列的，而是相互作用互相联系的。无论是传统知识组织方式的主题组织法、分类组织法还是领域本体等，其基本依托都是知识之间的关联。术语作为描述领域知识的词汇，凭借其单义性、规范性和领域性常作为领域知识的基本单元，术语之间的关联关系体现了知识之间联系的紧密程度和频繁度[48]。知识关联实际上是一种动态的把知识之间隐性关联标注为显性关联的一种计算行为[49]，术语之间关联规则的挖掘常采用算法来解决，这种挖掘是一种基于规则的机器学习算法，它的目的是利用一些度量指标来分辨数据库或者是语料库中存在的强规则。常采用的且最具代表性的Apriori算法[48]是利用逐层搜索的迭代方法找出术语库中项集的关系，以形成规则，该算法以先验知识或者是假设为基础，使用支持度来作为判断是否为频繁项集的标准，之后再采用频繁项集的先验性质，扫描整个术语库，找出符合最小支持条件的项，将其记为频繁项集，然后再利用该频繁项集作为标准找出下一个频繁项集，如此循环往复，直至无法再找出频繁项集为止。

4.3 可视并系统化

4.3.1 建立知识网络图

在已经建立的知识关联的基础上，可将知识关联可视化并建立起整体化的系统，利用特征项之间的关系，构建领域知识网络。共现分析、共词分析等的分析方法可以探索出同一特征集之间或不同特征集之间的关系，包括同项共现和异项共现网络[50]。加以结合可视化，各知识元(节点)之间的关系被直观呈现，基于图形化的分析更便于用户或分析人员去了解领域知识网络的特征结构、网络的规模大小(如平均最短路径、网络直径)、内部结构的属性(如密度、聚集性等)和各知识节点的计量属性(如中介中心性、接近中心性) 等信息[50]。进一步对知识网络进行聚类分析，可采用基于划分的K-means聚类算法[51]来实现对知识点的聚合，分析领域内的知识子结构、知识社团的分布情况与内部结构，这些分析进而反映出领域内的具体研究方向，同时也可帮助使用该知识组织系统的用户以层级的、逐步深入的方式进行浏览，构建完备的隐性知识结构。

4.3.2 步骤

首先对术语进行向量化，对其特征进行描述，构建实体-属性-关系三元组，术语作为实体(Entity)在框架中是被唯一标识的，并作为关系网络中的节点，代表社会网络中的一个行动者，它可以与其他一个或者是若干个行动者之间产生联系；属性(Attribute)即为实体的属性，也可以理解为对实体的描述，描述可以产生实体的特征，反应网络的特征结构；关系(Relation)即为实体与实体之间的联系，本质为知识的关联，可采用知识

关联挖掘算法来进行隐含关系的挖掘。其次将实体-属性-关系三元组映射为知识网络中的节点和边，映射过程可表示为(E,A,R)->G=(V,E)。在构建了知识网络之后进一步采用复杂网络的方法进行分析，网络的直径确定了该领域知识结构的规模；节点的度数中心度确定该节点代表的知识元在领域中的重要程度；分析边的权重反映关联的强弱；分析节点的中介中心性(Betweenness Centrality)判断该知识元在联系其他知识方面起到的作用大小等。最后采用基于划分的K-means算法进行多层次聚类分析，分析领域知识网络中知识社团属性。采用此种无监督的算法将之前已经向量化的术语按照预先设定的类目个数逐层划分为团体，每个团体都是整体知识网络中的一个小的知识结构，有着自身的特征结构和属性但同时又与其他结构之间存在若干关联。

5 结语

从基础知识架构类知识组织方法到关系网络建立类知识组织方法的演化既是知识组织沿时间、历史脉络的发展，更是人们对待知识的态度和认知上的进步。从文献篇章到知识单元，从知识孤岛到互联、复杂的知识库，人们的认知方式是从线性转变到树形结构再到网状结构。知识组织的方式是同时受到知识载体、认知方式、用户需求和技术水平多方面的影响的，其功能也在不断拓宽，从对词单(Term List)的标注到分类体系的形成再到对数据的挖掘、关联语义交流和系统的互操作，这架构起人与机器、机器与机器之间进行语义交换的桥梁[52]。

现如今，海量、异构且动态变化的数据使得领域知识的组织将面临更艰巨的挑战，为了适应大数据环境并发挥其应有的作用，知识组织的发展趋势也要顺应以下几个方面进行适当的调整。(1)逻辑表达语言专业化。可将关系网络类知识组织方法与基础知识架构类知识组织方法相结合，关系网络中每个作为节点的实体可由领域词表/词单中的条目来表示，利用规范的科学语言[53]、术语对知识进行专业、规范且富有逻辑的表达，专业化的规范语言有助于将难以编码的隐性知识显性化，知识也会以一种更便于理解和直观的方式被呈现出来。(2)更加个性化并注重知识的实时更新。在大数据时代海量数据的衬托下，个性化数据的推送和知识的动态更新是一种更为高效的手段。知识组织工具可根据个性化的需求动态组织知识资源，并完善自身，成为一个开放的动态有机系统，根据耗散结构理论与外界不断进行数据、知识乃至关系的交换[54]，使自身保持在一个动态的、高效的状态。(3)自动逻辑推理和人机交互，更加智能化。手工对于知识处理的效率是会随着信息量和知识量的大幅增长而下降的，知识组织与计算机相结合，实现了从传统信息处理向大数据处理的转变。依托大数据技术，知识得以被高效处理，在此基础上，未来的知识组织工具应更注重对语义的挖掘和关联，以语义为核心的处理才能够让机器理解文本，不需要人为的干预便可自行进行逻辑推理，达到专家系统的智能化程度。

参考文献

[1] 成 全, 罗 栋, 钟晶晶. 知识组织的理论缘起及演进路径探析[J]. 图书馆论坛, 2014, 34(11): 26-34.

[2] Stanford Encyclopedia of Philosophy. Peirce's Theory of Signs [EB/OL].[2021-07-10]. Peirce's Theory of Signs (Stanford Encyclopedia of Philosophy). https://plato.stanford.edu/entries/peirce-semiotics/.

[3] Atkin A. Peirce's final account of signs and the philosophy of language[J]. Transactions of the Charles S. Peirce Society, Indiana University Press, 2008, 44(1): 63-85.

[4] 周 宁, 吴佳鑫. 信息组织[M]. 第3版. 武汉: 武汉大学出版社, 2010.

[5] 李东巧, 陈 芳, 韩 涛, 等. 基于二模复杂网络的隐性知识发现方法研究——以潜在药物靶点挖掘为例[J]. 图书情报工作, 2020, 64(21): 120-129.

[6] Price D J D S. Little Science, Big Science[M]. Columbia: Columbia University Press, 1965.

[7] 关 鹏, 王曰芬, 曹嘉君. 整合主题的学科知识网络构建与演化分析框架研究[J]. 情报科学, 2018, 36(9): 3-8.

[8] 王 琳. 知识组织理论的三维结构[J]. 图书馆学研究, 2012(24): 2-8.

[9] Hodge G. Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files[M]. Washington DC: The Digital Library Federation Council on Library and Information Resource, 2000.

[10] 谢 靖, 钱爱兵, 韩 普, 等. 面向知识服务的知识组织工具: 现状与未来[J]. 现代图书情报技术, 2013(9): 8-14.

[11] 马费成, 宋恩梅. 信息管理学基础[M]. 第2版. 武汉: 武汉大学出版社, 2011.

[12] 李智杰, 曾 文, 乔晓东. 知识组织系统构建技术研究[J]. 情报理论与实践, 2017, 40(1): 115-120.

[13] 李千驹, 李思达, 刘建毅. 一种基于知识组织的关键词自动标引方法[J]. 情报科学, 2016, 34(11): 107-110, 139.

[14] 李育嫦. 传统知识组织系统的重构及其在网络环境下的应用[J]. 情报杂志, 2011, 30(7): 114-118.

[15] Shiri A A, Revie C. Thesauri on the Web: current developments and trends[J]. Online Information Review, Bradford: Mcb Univ Press Ltd, 2000, 24(4): 273-279.
[16] 于海涛, 高一波, 杨一平. 基于知识树的领域知识组织和应用[J]. 计算机应用研究, 2008(11): 3246-3248,3252.
[17] 孙建军, 柯 青, 陈晓玲, 等. 信息资源管理概论[M]. 第2版. 南京: 东南大学出版社, 2008.
[18] 罗鹏程, 陈 翀. 从大众分类到层次式资源组织体系——利用聚类信息构建标签树[J]. 图书情报工作, 2013, 57(22): 120-125,59.
[19] 张运良. 大数据服务中知识组织的挑战及应对[J]. 图书情报工作, 2020, 64(4): 88-94.
[20] 司 莉, 周李梅. 近年来国外英文知识组织系统研究现状和发展趋势[J]. 图书馆论坛, 2010, 30(6): 220-226,208.
[21] 王忠红. 知识组织工具的发展和趋势[J]. 图书情报知识, 2009(6): 97-102.
[22] 滕广青, 贺德方, 彭 洁, 等. 结构与秩序:知识组织领域中结构主义思想的演进[J]. 情报理论与实践, 2015, 38(4): 6-10.
[23] 何 琳. 基于知识组织资源仓库的领域本体构建研究[J]. 图书馆杂志, 2011, 30(12): 59-62,112.
[24] 白 华. 大众分类本体与知识组织系统融合研究[J]. 图书馆学研究, 2016(10): 53-59.
[25] 毕 强, 鲍玉来. 数字图书馆知识组织体系构建的发展路径——概念格与本体的互补融合[J]. 华中师范大学学报(人文社会科学版), 2011, 50(5): 130-136.
[26] Choi N, Song I-Y, Han H. A survey on ontology mapping[J]. Sigmod Record, New York: Assoc Computing Machinery, 2006, 35(3): 34-41.
[27] 肖 强, 郑立新. 关联数据研究进展概述[J]. 图书情报工作, 2011, 55(13): 72-75,134.
[28] 高 斌. 网络发展背景下的知识组织新思考[J]. 图书情报导刊, 2020, 5(1): 26-32.
[29] 曲佳彬, 欧石燕. 关联数据可视化研究进展分析[J]. 图书与情报, 2018(4): 51-61.
[30] 董 坤. 基于关联数据的高校知识资源语义化组织研究[J]. 情报理论与实践, 2016, 39(3): 91-95.
[31] 曾子明, 周 知, 蒋 琳. 基于关联数据的数字人文视觉资源知识组织研究[J]. 情报资料工作, 2018(6): 6-12.
[32] Patterson O, Hurdle J F. Document clustering of clinical narratives: a systematic study of clinical sublanguages[J]. Annual Symposium Proceedings. AMIA Symposium, 2011, 2011: 1099-1107.
[33] 王 昊, 苏新宁. 基于CSSCI本体的学科关联分析[J]. 现代图书情报技术, 2010(10): 10-16.
[34] 王 昊, 谷 俊, 苏新宁. 本体驱动的知识管理系统模型及其应用研究[J]. 中国图书馆学报, 2013, 39(2): 98-110.
[35] 张 璇, 黄香玲, 李 鸿, 等. 基于CSSCI的本体研究热点与前沿的可视化分析[J]. 情报科学, 2013, 31(5): 151-154,160.
[36] 刘 挺, 车万翔, 李正华. 语言技术平台[J]. 中文信息学报, 2011, 25(6): 53-62.
[37] 杨 涛. 中文信息处理中的自动分词方法研究[J]. 现代交际, 2019(7): 93-95.
[38] 杨贵军, 徐 雪, 凤丽洲, 等. 基于最大匹配算法的似然导向中文分词方法[J]. 统计与信息论坛, 2019, 34(3): 18-23.
[39] 戴上静, 石 春, 吴 刚. 中文分词中的正向增字最大匹配算法研究[J]. 微型机与应用, 2014, 33(17): 15-18.
[40] 吴 帅, 潘海珍. 基于隐马尔可夫模型的中文分词[J]. 现代计算机(专业版), 2018(33): 25-28.
[41] 王 星, 于丽美, 陈 吉. 融合字根信息的卷积神经网络中文分词方法[J/OL]. 1-8 [2021-03-19]. http://kns.cnki.net/kcms/detail/21.1106.TP.20210319.1020.014.html.
[42] 王 星, 李 超, 陈 吉. 基于膨胀卷积神经网络模型的中文分词方法[J]. 中文信息学报, 2019, 33(9): 24-30.
[43] 季培培, 鄢小燕, 岑咏华. 面向领域中文文本信息处理的术语识别与抽取研究综述[J]. 图书情报工作, 2010, 54(16): 124-129.
[44] 袁劲松, 张小明, 李舟军. 术语自动抽取方法研究综述[J]. 计算机科学, 2015, 42(8): 7-12.
[45] 彭 博. 主题-知识关联的网络文物信息资源知识推荐方法研究[J/OL]. 情报科学, 1-7. [2020-10-23]. http://kns.cnki.net/kcms/detail/22.1264.G2.20201022.1509.012.html.
[46] 张 雪, 孙宏宇, 辛东兴, 等. 自动术语抽取研究综述[J]. 软件学报, 2020, 31(7): 2062-2094.
[47] 滕广青. 基于频度演化的领域知识关联关系涌现[J]. 中国图书馆学报, 2018, 44(3): 79 - 95.
[48] 阮光册, 夏 磊. 基于词共现关系的检索结果知识关联研究[J]. 情报学报, 2017, 36(12): 1247-1254.
[49] 李旭晖, 凡美慧. 大数据中的知识关联[J]. 情报理论与实践, 2019, 42(2): 68-73,107.
[50] 张发亮, 刘君杰, 周 沫. 领域知识结构基础理论及构建研究[J]. 情报杂志, 2018, 37(2): 188-193.
[51] 王 昊, 邓三鸿, 苏新宁. 我国图书情报学科知识结构的建立及其演化分析[J]. 情报学报, 2015, 34(2): 115-128.
[52] 贾君枝. 面向数据网络的信息组织演变发展[J]. 中国图书馆学报, 2019, 45(5): 51-60.
[53] 孙 兵. 知识组织工具的发展趋势浅析——基于分类表、叙词表和知识本体的比较研究[J]. 图书馆学刊, 2009, 31(11): 86-88.
[54] 苏新宁. 知识组织的科学理论阐释[J]. 图书与情报, 2013(6): 1-7.

[作者简介] 曹思源,女,1997年生,南京大学信息管理学院硕士研究生。
马海云,女,1995年生,南京大学信息管理学院博士研究生。
收稿日期:2021-04-30