Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

知识组织系统构建中对既有资源的利用方式分析*

□ 张运良 张兆锋 闫莹莹 许德山 / 中国科学技术信息研究所 北京 100038

摘要：知识组织系统的构建是一项艰巨而复杂的工作，而利用既有资源，尤其是词表和语料，则在一定程度上能够减少这一任务的工作量。文章据此研究了对既有资源利用方式的四个相关问题。首先是从既有知识组织系统中提取出所需局部的方式及注意事项。其次是跨语言利用外文知识组织系统，生成当地语言知识组织系统雏形。再次是从选词、相关词推荐和词间关系验证等角度分析语料库在知识组织系统构建过程中的应用。最后探索了建设中的知识组织系统的利用方式，并提出了两个循环模型。了解和掌握对既有资源的利用方式，能够促进更好更快地构建知识组织系统。

关键词：知识组织系统，既有资源，语料库，利用方式，循环模型
DOI：10.3772/j.issn.1673—2286.2013.11.006

引言

知识组织系统包括叙词表、词系统、本体等不同的类型。根据其领域和知识结构的不同，可以用于文献标引、专利分析、科技监测、情报分析等信息和知识服务，在面对海量信息资源的条件下，对知识组织系统的需求也在不断增加[1]。但是，构建知识组织系统是一项艰巨而复杂的工作，需要大量具体领域和知识组织系统方面专业人士的参与。在这种情况下，如何能够以较低的成本，更快更好地建设知识组织系统则变得尤为重要。一方面，既有的不同类型的知识组织系统或多或少能够提供一部分在待建知识组织系统仍然适用的知识。另一方面，与待建知识组织系统相关的语料资源，能够保证知识的准确性。不同类型的知识组织系统内容不尽相同，对于知识组织系统构建最为重要的几类知识包括词条、词条定义、翻译、属性以及词条之间的关系。在对既有资源利用方面，本文结合工作实践，重点分析了对既有同语言知识组织系统、既有跨语言知识组织系统以及语料库系统的利用方式。同时，本文提出知识组织系统建设中对已经建成部分的利用，并提出两种循环利用模式。

1 利用既有同语言知识组织系统

知识组织系统是来自现实世界的知识体系，而知识体系是一脉相承的，所以总能找到一些相关的知识组织系统。从既有知识组织系统的数量上看，可能存在一部或者多部，一部的情况多对应相对比较局限的狭小领域，如顶级机构[2]；多部的情况对应建立知识组织系统是综合性的领域，如工程技术，或者交叉性新兴领域，如新能源汽车领域[1]。

对于只有一个知识组织系统来源的情况，处理相对比较简单，仅仅从中抽取出需要的内容即可。主要有三种抽取方式：1）按照词族抽取；2）按照范畴抽取；3）按照子网络抽取。

如图1中（A）为一个叙词表的局部的示意。其中圆圈表示词条，双圈特别表示族首词。而单箭头表示层级关系，从上位词指向下位词，不同颜色的单箭头联系起不同的词族，在这一叙词表局部中包含三个词族。紫色的双箭头表示在局部范围内的相关关系，而黑色的虚线表示词条对应的其他相关关系和用代关系。按照词族抽取，可以仅仅抽取词条以及词族范围内的上下位关系，也可以此为基础，进一步扩展抽取出一定范围内的关系和

* 本文系国家自然科学基金项目“面向特定情报分析应用的知识组织系统快速构建关键问题研究”（编号：71203208）、国家“十二五”科技支撑计划课题“面向外文科技文献信息的超级科技词表和本体建设”（编号：2011BAH10B01）、中国科学技术信息研究所重点工作项目“汉语科技词系统建设与应用工程”（编号：ZD2012-3-2）的研究成果之一。

关系词，通常扩展一两层即可。因为扩展必然涉及外部词族，如果不停扩展下去，就有可能把原有的知识组织系统的全部或者大部涵盖，失去了抽取的意义。如图1中（B）表示自左上方的族首词开始逐层扩展的结果，可以发现只要扩展4层即可以包含图1（A）中全部词条。如果从右上方的族首词开始，扩展6层，也能包含局部的全部词条，如图1（C）所示。

范畴也是对于词条的另外一种划分依据，如对于图1中（A）所示的叙词表片段，可以根据范畴分为图2（D）的方式，在图中以不同颜色表示不同范畴。对于这种情况，也可以按照类似词族的方式进行扩展。第三种方式，是目前相对理想的方式，但是划分更加复杂，需要将原有的知识组织系统抽象为一个复杂的图，然后利用复杂网络相关理论，将其划分为若干的不交叉的子网或者社区，如图2中（E）所示，根据初始的种子词，找到一个划分，从中抽取若干的子网能够包含种子词（也可以限定一定比例的阈值），从而把需要的子网切割出来。对于这种方式，无需作逐层的网络扩展。虽然当前没有将社区检测技术应用于知识组织系统的子网划分，但是相关的研究和方法可以借鉴[3,4]。

对于单一来源的词条和知识，还需要逐条审核，以去掉不合理和明显已经过时的局部知识。对于多个来源的情况，第一步的抽取是类似的，但是还有集成融合的处理，这种处理，也可以有不同的处理方式。一是全部吸收，然后检验排查逻辑错误，在这方面已经有较为成熟的研究[5,6]。另外一种办法是先吸收全部词条，然后重新建设关系知识。

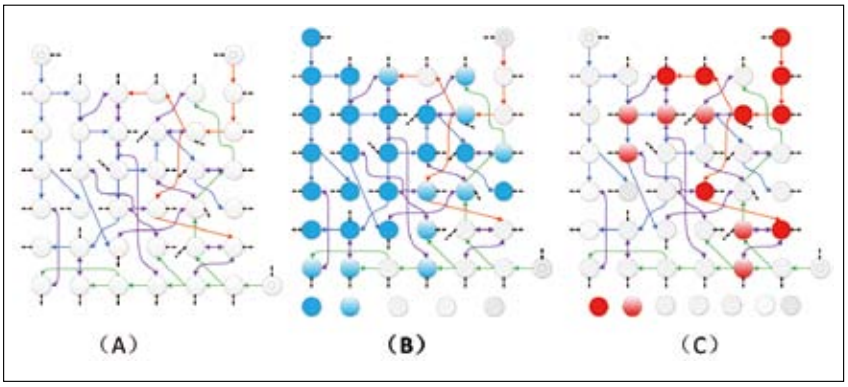

图1 从既有同语言知识组织系统中提取词族并逐层扩展示意图

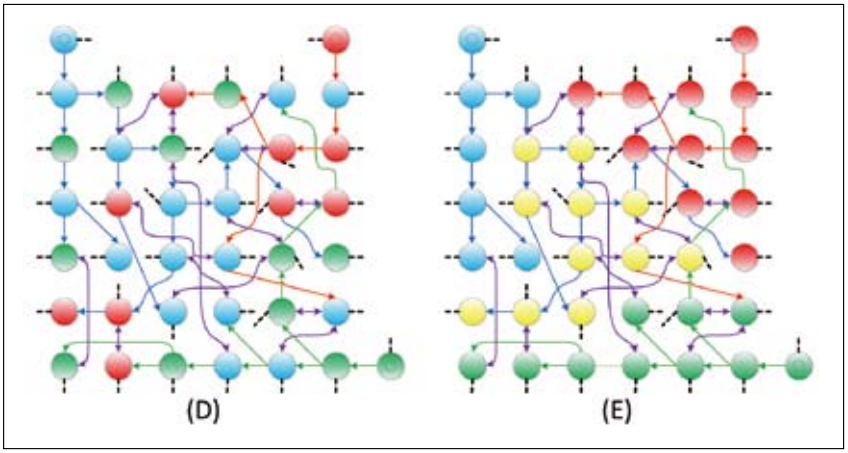

图2 对于叙词表局部进行范畴划分和社区子网划分的结果示意图

此外，也有一些领域没有相对完善的知识组织系统。如百科全书，可以将条目及其英文翻译作为词条的基本信息，条目对应的目录范畴作为初步的分类，条目的解释作为词条的定义，相关的实践尝试证明上述方法是可行的。

2 利用既有跨语言知识组织系统

由于相当数量的知识组织系统存在词条的双语或多语对应，因此可以用某种语言的知识组织系统来生成第二种语言的知识组织系统的雏形。对于仅仅是单语的情况，也可以先邀请专家或者结合语料进行翻译的过程，相对重新建设，其成本较低。

相应的构建流程如图3所示，针对来源词表V1切割为概念表（concept）和关系表（relation），目的是将两表转化为另外一种语言的词表V2，对于其余的如定义、分类和属性等知识，转化方法和规则暂时还不成熟，将在后续工作中进一步研究。V1概念表包含CL1（当前知识组织系统的语言L1的词条）和CL2（待建设知识组织系统语言L2的词条）两个字段。关系表包含CL1A、REL和CL1B三个字段，分别表示以L1语言表示的两个词条及它们之间的关系，自左向右来解释。一般叙词表中包含用代关系、

层级关系和相关关系三类，更为复杂的关系，往往也可以归纳为以上三种。对于每个CL1对应的CL2可能是多条，也可能局部是缺失的。对于缺失的需要先进行完善补充，再将1对多的拆分为1对1的形式。最后对于没有对应翻译的词条和对应的关系进行删除，以此为基础进行V2的构建。

首先构建V2的概念表，以CL2为主键进行合并，将多条CL1作为其翻译，然后进一步将对应同一个CL1的多条CL2词之间建立用代关系，需要确定一个为优选词。目前采用从同义词集合中选取关系数量最多的一个作为优选词，其余的作为可替代的词条，并将关系命名为alterLabel。然后将所有附加在非优选词上的其他关系都转移到优选词上，剩余的关系类型包括related（相关）、narrower（下位）和broader（上位）三类，形成V2的关系表。下一步需要在关系表中查重去掉重复的关系，然后查找CL2A和CL2B都相同，但是REL不同的数据，并根据以下原则进行处理：1）narrower/broader之一如果出现，则优先选用这个关系，否则就选用related关系；2）如果narrower和broader的关系同时出现，则选用related关系。在这一过程中，希望尽量减少人工的判断，主要是由于原来的词系统都是领域专家和知识组织系统专家共同建设和审定而成，如果完全依赖一两个人进行人工判断，很难保证准确性。而按照以上的原则，则可以相对保证其一致性，同时，其速度也较快。

3 对既有语料资源的利用

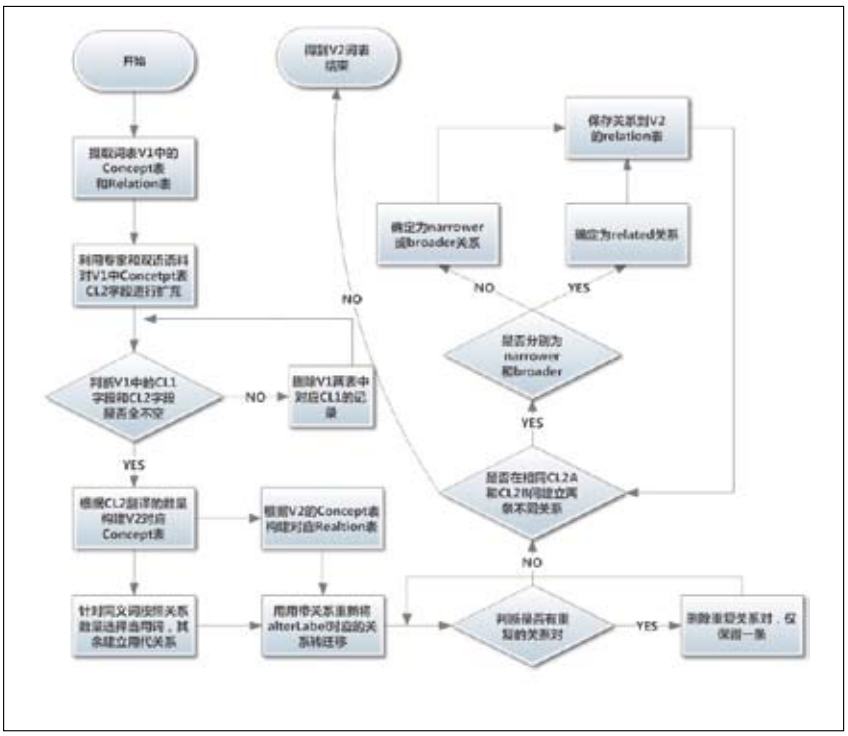

图3 跨语言利用知识组织系统构建知识组织系统流程图

语料资源往往结合语料库平台使用，在本文以中信所综合语料库辅助汉语科技词系统建设为例进行介绍。该平台开发的核心思想是利用领域相关的期刊论文、会议论文、学位论文和专利等科技文献进行计算分析，提供对选词和关联关系构建的统计支持。当然，本语料库平台也可以基于企事业单位自有资源进行针对性分析，此过程可能需要对相关资源的元数据作一定的转换处理。

该平台包括三项前台功能和四项后台功能。前台功能是关键词词频统计、关键词共现和语料全文检索。关键词词频统计，主要从文献中提取关键词，并对词频作分析，分析可以从所属领域、语料类型、出现位置、时间范围、出现频率等几个角度进行分析，从而实现初步的选词功能，如图4所示。关键词共现则是在选定的词条基础上（包括机器筛选的和人工修订完善的）进行两两共现分析，从而支持关系构建。共现类型区分为关键词共现、句内共现和文献内共现，可以根据需要为不同的共现类型赋予不同的权重。全文检索基于Lucence构建，主要用于构建过程中临时的共现分析，这是因为知识工程师在知识构建过程中，总会引入一些已经筛选好的词条列表中没有的词，这些词与既有的词的共现信息并没有预先计算，需要以全文检索来补充。后台功能是领域范围管理、语料管理、专业词典维护和数据计算。领域范围管理用于管理语料库平台中多个领域的增加、删除和修改。语料管理主要用于浏览、增加和删除各领域的语料。专业词典维护可以将人工修订的筛选词导入系统，使得系统能够在这一基础上进行共现计算。数据计算主要分为三个子功能，分别是索引建立、词频统计和共现计算，此外，还

图4 中信所综合语料库平台关键词词频统计功能截图

有一些计算辅助功能。

在整个建设过程中，语料库平台得到广泛的使用。

首先是选词阶段，目前在实践中文献来源主要是来自万方数据的期刊论文、学位论文、会议论文和专利数据。根据实际情况，可以补充应用中可能用到的用户的数据，这样的效果会更好。目前选词主要有5个原则：

1）高频词优先，低频词尽量不选用。根据不同领域的情况以及选词要求，也可能会保留部分低频词。虽然理论上来说高频词也有一些实际上接近通用词且不宜选用，但是在初始处理上已经做到尽量选用作者关键词或者既有词表词库作为选词基础，所以一定程度上可以避免这个问题。此外，这只是一个初选集，后续还有人工审核互动过程，可以进一步排检。

2）关键位置上的词优先。针对目前涉及的文献，关键位置一般包括标题和关键词部分，在这些位置上的词相对更重要。

3）时间靠后优先。也就是同等情况下，近期出现频次较高的，相对来说较为重要，并且有逐步变得更加重要的趋势。

4）用户自有资料优先。一般来讲在用户自有资料中的频次比在通用文献中的频次更重要，因为以后可能大量处理类似的自有资料。

5）选词参考数值。具体的频次数值、年代数值没有统一参考数值，主要由于不同领域所能得到的基础文献和资料的数据差异较大，所以无法给出统一的参考数值。但有一个参考标准，可以按照希望选择出来的词按照1.5-2倍选取，这一过程，可以通过反复尝试和在检索结果中进行再检索得以实现。

核心词实际上有三个来源：一是语料库中的统计数据（当然领域专家还会删除一些词条，所以这部分是专家审定后保留的），二是领域专家审定过程中补充添加的词条，三是知识工程师在加工过程中补充的词条。

语料库平台在建设过程中的作用，主要是共现分析，分析经过专家审定后的所有词条，计算共现频次，以适当的形式展现给知识工程师参考。此外，还利用了全文检索功能，为知识工程师新增词条，构建相关关系提供部分语料支撑。保证新增的知识也有依据，如果试图在两个词条间建立了关联，在资料或者文献中没有共现过，则需要重新重点评估其准确性。

4 利用知识组织系统已建成部分的循环模型

知识组织系统的建设是一个循序渐进的过程，建设过程中包含一定的顺序流程。这个顺序是循环的，每一类型知识的改变，都可能改变知识组织系统其他类型的知识。一般知识组织系统包含的内容都可以归纳为五个要素，分别是词条、定义、翻译、关系和属性。首先根据词条，人工添加对应的定义、翻译、关系和属性知识，通过这些知识反过来能够进一步丰富词表，增加词条或者修订词条。而利用翻译[7]和定义[8]，一方面可以发现新的关系和属性，也可以发现既有关系和属性中矛盾冲突之处，从而有可能对存在错误的关系、属性作修订，或者对翻译和定义作修订。所以在五个要素之间存在添加知识、词条丰富、知识发现和检验校对等四种类型的关系。具体的循环模型如图5所示。此外，还可以进一步引入外部语料，从而在上述循环模型基础上得到更大的循环模型，即引入领域语料库，从词条、定义、翻译、关系和属性等的建设，都能得到语料的印证，同时建设好的知识组织系统可以反过来作用于语料库，用于筛选确认选择语料的合

理性，可以直接计算知识组织系统在每篇语料上的覆盖程度，也可以对每篇语料进行标引，再以标引词进行计算。经过如此反复循环，最后将实现语料库和知识组织系统的全面融合，见图6。两个循环模型分别称为内部循环模型和内外循环模型。

5 结语

本文结合工作实践，探索了知识组织系统构建中对既有资源的利用方式问题，先后分析了同语言知识组织系统资源、跨语言知识组织系统和语料库等外部既有知识资源的利用方式。同时，本文提出了在知识组织系统内部，利用既有知识组织系统建成部分进行知识组织系统建设的循环模型，并将语料库融合其中，形成了内外循环模型。集成既有的平台，并针对知识组织系统构建的循环模型完善有关环节，将是下一步工作中的研究重点。

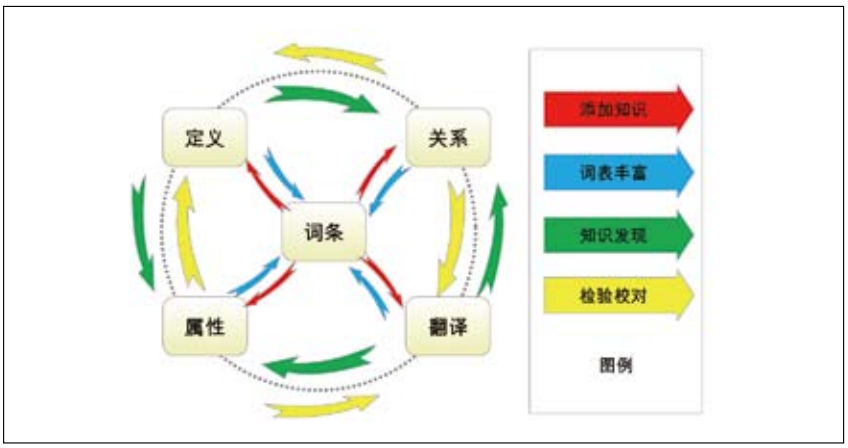

图5 知识组织系统构建的内部循环模型

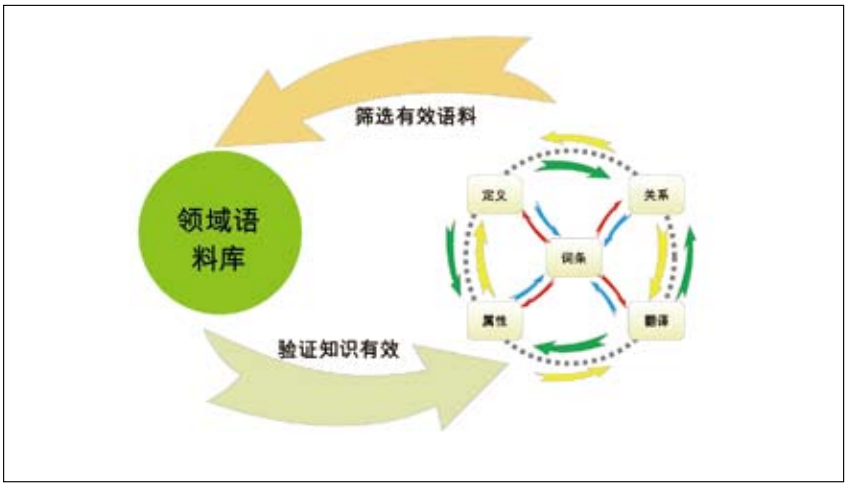

图6 知识组织系统构建的内外循环模型

参考文献

[1] 贺德方,乔晓东,朱礼军,等.汉语科技词系统(新能源汽车卷)[M].北京:科学技术文献出版社,2012.
[2] 杨奕虹,李雅萍,张立丽,等.机构多层级词表的编制及在文献计量评价与科研绩效管理中的应用[J].数字图书馆论坛,2013(6):57-63.
[3] LANCICHINETTI A, FORTUNATO S. Community detection algorithms: a comparative analysis [J]. Physical review E, 2009, 80(5): 056117.
[4] JIA GUANBO, CAI ZIXING, MUSOLESI M, et al. Community Detection in Social and Biological Networks Using Differential Evolution [A]. Learning and Intelligent Optimization,Lecture Notes in Computer Science, 2012: 71-85.
[5] 徐硕,乔晓东,朱礼军,等.机构多层级词表的编制及在文献计量评价与科研绩效管理中的应用[J].数字图书馆论坛,2010(8):55-58.
[6] 吴雯娜,王星.叙词表融合方法研究[J].中国图书馆学报,2012(4):110-118.
[7] 张远良,乔晓东,朱礼军,等.基于术语翻译信息的同义关系快速构建方法研究[J].图书情报工作,2013,57(8):109-113.
[8] 张远良,梁健,朱礼军,等.基于术语定义的科技知识组织系统自动丰富关键技术研究[J].现代图书情报技术,2010(7/8):66-71.

作者简介

张远良（1979-），男，博士，副研究员。研究方向：为知识组织、知识工程、自然语言处理、文本自动分类。E-mail:zhangyl@istic.ac.cn
张兆锋（1979-），男，在读博士，助理研究员。研究方向：专利分析、数据挖掘、信息可视化。
闫莹莹（1981-），女，中国科学技术信息研究所，硕士。研究方向：知识组织，自动标引。
许德山（1979-），男，博士，中国科学技术信息研究所信息技术支持中心助理研究员，研究方向：知识组织、文本挖掘、语义Web。

The Utilization Pattern of Existing Resources in the Construction of Knowledge Organization Systems

Zhang Yunliang, Zhang Zhaofeng, Yan Yingying, Xu Deshan /Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: It is complicated and difficult to construct a knowledge organization system nowadays. To utilize the existing resources, especially vocabularies and corpus, will decrease the work to some extent. Four problems about the utilization pattern of existing resources in the construction of knowledge organization systems are studied. The first is the patterns and attentions of extracting the useful parts from existing knowledge organization systems. The second is the patterns of constructing a prototype of a knowledge organization system in some language by use another one in foreign language. The third is the three patterns of term selection, related terms recommendation and verification of relation between terms with a domain corpus. At last, the patterns of using already constructed parts of knowledge organization system itself, and two circulation models are proposed. It will lead to better and more rapid construction of knowledge organization systems to know and master the utilization patterns of existing resources.

Keywords: Knowledge organization systems, Existing resources, Corpus, Utilizationpattern, Circulation model

（收稿日期：2013-10-09）