Science Innovation
2017; 5(1): 9-14
http://www.sciencepublishinggroup.com/j/si
doi: 10.11648/j.si.20170501.12
ISSN: 2328-7861 (Print); ISSN: 2328-787X (Online)

Research on Periodical Literature Knowledge Organization Based on Knowledge Element

Wang Yu, Li Xiuxiu

Faculty of Management and Economics, Dalian University of Technology, Dalian, China

Email address:
ywang@dlut.edu.cn (Wang Yu), 1434002120@qq.com (Li Xiuxiu)

To cite this article:
Wang Yu, Li Xiuxiu. Research on Periodical Literature Knowledge Organization Based on Knowledge Element. *Science Innovation*. Vol. 5, No. 1, 2017, pp. 9-14. doi: 10.11648/j.si.20170501.12

Received: January 16, 2017; **Accepted**: February 10, 2017; **Published**: March 3, 2017

Abstract: With the fast growing of internet technology, people can contact a big sum of information, the right information people needed are submerged into rubbish information. So it is essential to analysis the deep into the literatures, and to realize the reasonable expression,extracting,classification and linking of the literature knowledge for the mass literatures. In this paper, we regarded the knowledge element as the smallest unit of journal literature knowledge structure. Proposed a five knowledge organization model of journal literatures and introduced lexical chain as the topic presentation of knowledge element. For the extracted knowledge elements, HNC field concept was used to organize the same knowledge elements together. Then based on the characteristics of knowledge element and HNC theory, this paper analyzed the logical and semantic linking, and proposed three knowledge element semantic linking method, thus different knowledge elements were combined together to form a complex semantic net in order. Lastly, knowledge element database were built for the knowledge elements in different fields, and achieved the knowledge organization.

Keywords: Knowledge Organization, Knowledge Element, Hierarchical Network of Concepts (HNC), Lexical Chain

基于知识元的期刊文献知识组织研究

王宇，李秀秀

管理与经济学部，大连理工大学，大连，中国

邮箱
ywang@dlut.edu.cn（王宇），1434002120@qq.com（李秀秀）

摘要：网络通讯技术的飞速发展使得可以接触的信息呈爆炸式增长，急需的知识通常被实际不需要的信息垃圾淹没。针对海量期刊文献有必要深入文献内部，实现对文献知识的合理表示、抽取、分类和链接。本文把知识元作为期刊文献知识结构的最小组成单元，提出期刊文献的知识元五元组描述模型，引入词汇链作为知识元主题的表述，利用HNC的领域概念归类方法对知识元的主题领域进行了划分，并基于知识元的属性和HNC理论，分析了知识元的逻辑关联和语义关联，提出了词汇链、主题领域、HNC概念关联式三种知识元语义链接方法，最后针对不同领域粒度下的知识元建立了相应的知识元库，实现了知识组织。

关键词：知识组织，知识元，HNC（概念层次网络）理论，词汇链

1. 引言

随着网络信息技术的高速发展，信息资源也呈爆炸式增长，传统的知识服务模式通常是以文献或是信息为单元进行的服务，面对信息资源泛滥、知识匮乏的境地，这种服务模式俨然满足不了用户需求，用户更多的是希望可以准确定位到文献内部的知识，而不仅仅停留在文献表层，这就需要我们通过各种方式对各种信息资源进行知识的收集、整理，使得信息资源知识化、网络化，能够主动向用户提供满足他们个性化需求的知识化服务。

在信息资源管理中，文献资源是一种重要的资源形式。文献是知识的载体，是知识组织和呈现的主要方式。随着数字化信息资源的出现以及数字图书馆的构建，电子期刊文献成为主要的信息资源形式，国内著名的CNKI、万方等期刊文献网络数据库，收录了各大期刊论文、硕博论文、外文文献、专利、科研成果等，科研工作者可以通过文献检索、文献阅读，了解相关领域研究内容、科研动态和最新热点等。期刊文献是科研工作者进行科研工作不可或缺的重要资源之一。

在传统的信息服务模式下，期刊文献的管理大多基于文献的物理特征，如文献主题、作者、机构、篇名、关键词、摘要、参考文献等外部特征，通过物理特征的关联为用户提供所需文献，人们可以通过对检索到的文献进行阅读梳理获取内部知识。这种知识组织方式是对文献这一知识载体的组织，而非知识本身。有文献统计[1]，文献的更新数量以每20个月翻一倍的速度增长，在信息资源丰富，知识匮乏的状况下，如何有效地对文献知识进行组织管理，为用户提供便捷的、可利用的知识变的越来越重要。这就需要摆脱传统的以文献和逻辑信息为组织单位的模式，深入文献内部构建适合知识组织和管理的知识单元。知识元具备完整的知识表达，并且具有可组合、可链接、独立单一等特性，是构成知识的基本单元[2]。从知识结构的角度出发，每一篇文献都可以看成是一个知识单元，知识元通过不同的排列组合和知识关联形成知识单元。知识元是实现知识组织的核心，以知识元的形式对文献内容进行表示，利用知识元之间的链接关系组织、关联，为用户呈现深入文献内部的知识[3-4]。姜永常[5]提出了用引文关联的方法来提取文献知识元的思想，根据句型结构定义不同的规则来抽取三元组。提取出来的不同知识元之间存在着逻辑依存关系，知识元链接就是将这些具有关联关系的、构成知识的基本单位链接到一起，无数的知识元链接可以构成知识网络[6]。知识元作为一种工具，不仅可以通过知识元抽取、标引，对知识进行有效组织和集成，而且可以通过知识元挖掘和链接解释知识内涵，为知识创新提供切入点[7]。吕颖[8]将知识链接与网络知识服务结合起来进行研究,阐明了知识链接在网络知识服务中的应用模式。

国内对于知识组织的研究主要集中在图书馆情报领域，陈烨,赵一鸣[9]分别从知识描述与揭示、知识单元互联、知识序化3个方面分析了关联数据在知识组织中发挥的作用。高俊峰[10]提出一种基于语义标签的数字文献资源组织方法,为新技术标准下的数字图书馆知识服务工作的开展提供了解决方案。刘术华,牛现云[11]分析了移动阅读环境下公众对知识需求的“碎片化”趋势和特点，并且对“碎片化”知识选取和组织给出了建议,为图书馆进行知识组织与服务提供了借鉴和参考。但是，随着计算机软件、网络通信等技术的发展，知识组织从传统的以文献为单位的组织转向以信息为单位的组织，从物理层次的信息组织发展为认知层次的知识单元组织，知识组织的形式也从传统的文献管理、信息管理到如今的知识管理。陈述年[12]提出知识组织是由文献单元深入到知识单元的文献整序的过程。刘淼[13]对期刊文献设计了知识元的表示方式，并构建了期刊文献的知识仓库。总之，以知识元作为知识表示和知识组织的主要基元已经得到大家的广泛认可。

本文首先针对期刊文献的特点以及基于HNC理论[14]对知识元的表示方式做了研究，提出了期刊文献知识元的五元组表示形式，然后通过将期刊文献的知识元划分为10大主题领域对知识元进行了分类，并把知识元的链接分为逻辑链接和语义链接，提出了三种知识元语义链接方法，最后通过建立知识元库实现了期刊文献知识组织。

2. 期刊文献的知识元表示

知识表示不仅仅是对客观事实的表述，更要表示知识之间存在的关联性。通过对知识的合理化表示，使得知识更容易被识别和理解；对知识关联性的表示，也使知识之间更加融会贯通。为此，把期刊文献的知识元定义为五元组形式，如表 1 所示。

表1 文献知识元组成项描述。

基本属性	属性描述
来源	文献知识元具体来自的期刊文献信息
导航	文献知识元在期刊文献中上下文信息
名称	文献知识元所表达内容的集中概括，在这里用主题词表示
内容	文献知识元描述的知识信息，在这里用主题句表示
主题领域	文献知识元所描述主题的归属领域

知识元的五元组表示是文献内容的结构化表示，是对文献内容和知识关联的描述。以下是对知识元五元组基本属性的详细说明：

（1）来源：文献知识元具体来自的期刊文献信息，是知识元和文献单元之间的直接关联，知识元还可以与文献互逆导航，可以查看文献内的知识元信息，也可以通过知识元的检索查找直接相关的有价值的期刊文献。

（2）导航：知识元导航这一属性，用以表示知识元在期刊文献中的上下文信息，例如对文献中知识元的查看，可以根据知识元的段落排序，使知识元以在原文中出现的顺序呈现，复原文献原来的知识结构顺序。

（3）名称：名称属性是对知识元所描述的主题句的集中概括，这里是用主题词来表示的。

（4）内容：知识元是知识的载体，知识元的内容属性表示的是知识点的具体内容，这里是对知识元所属主题内容的具体描述，通过主题句表示。主题句和主题词具有对应关系，主题句是从包含主题词的句群中产生的，是对文献某一主题内容的描述。

（5）主题领域：主题领域是文献知识元所描述主题的归属领域，对主题领域进行划分可以使具有相同主题领

域的知识元汇集在一起，同文献的知识元以及不同文献的知识元因共主题领域而相互关联，对主题领域的划分也为知识元检索提供了重要的导航信息。

3．知识元主题领域划分

对知识元“领域主题”的判定也是知识分类的过程，知识分类是知识组织的重要方式，知识元的主题领域划分是在以文献为单位的知识分类体系下的一种细化，为知识元库的构建提供标准，也为知识元的检索提供了重要的知识导航信息。HNC理论的领域概念吸收了传统的分类体系的思想，可以把知识元的类别归入到领域概念层次树中，然后根据需要设置分类层次。

HNC理论在基层概念空间中，以人类活动为主题划分10大领域类，共300多个高层分类，它们是10大领域类下的二到三级子类，每个高层分类都可以延展细分，领域之间也可以组合。HNC的高层领域划分[15]如表2，形成了封闭的领域类别空间。

表2 HNC高层领域。

编号	领域	概念基元符号
1	心理活动及精神状态	71,72
2	人类思维活动	8
3	专业及追求活动	a,b
4	理念活动	d
5	第一类劳动	q6
6	业余活动	q7
7	信仰活动	q8
8	本能活动	6m(m=0,1,…,6)
9	灾祸	3228n(n=8,9,…,b)
10	状态	503,50k(k=8,9,…,b)

表2中列出的是HNC领域的高层概念节点，HNC的语义网络是树状的分层结构，树的每个节点代表一个概念，即概念基元，每棵树的概念节点形成一个概念聚类，每颗子树的概念节点形成一个子类。上表的高层概念节点还可以向下分类为二级子类节点、三级子类节点等，例如HNC的专业活动a领域的二级子类有$a0$（一般专业活动）、$a1$（政治）、$a2$（经济）、$a3$（文化）、$a4$（军事）、$a5$（法律）、$a6$（科技）、$a7$（教育）、$a8$（卫生），图1是专业活动a领域的概念树。

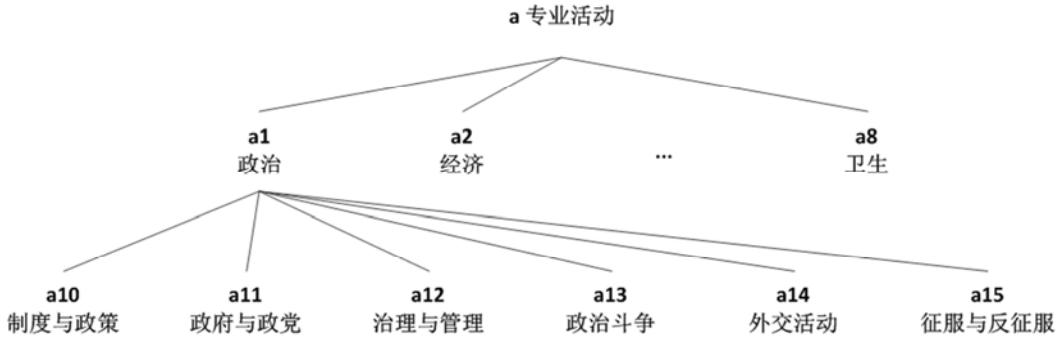

图1 概念层次领域树。

图1是a领域的三级概念树，可以根据需要从概念树的根概念向下逐级概念延伸，作为领域分类类别，底层的概念更为具体，描述也更为详尽。

知识元的主题领域划分主要利用词汇链中包含领域信息的概念节点并结合概念关联式进行的。首先利用文献[16]提出的词汇链生成算法，对词汇链进行构造、合并和优化，建立词汇链基本信息统计表，其包含的字段为w，hnc，freq，(l1，l2，…，lk)，rel，relWord，其中w为词汇链中的词汇，hnc是词汇映射的HNC语义符号，freq为该词的词频，(l1，l2，…，lk)为词汇出现时所在的段落、句子位置，relWord是该词汇链在入链时与其相似度最大或是具有概念关联的词，rel是与其相似度的值。然后从信息表中获取该链的hnc字段，作为领域判定的特征词，在这里先选取含有领域信息的概念节点进行一级压缩，然后对于剩下的不含有领域信息的概念节点，利用概念关联知识库中的与其相关联的基元对其进行替换，以对词汇链中的概念节点进行二级压缩。知识元的概念关联式有10种关联关系，如果某个概念基元有多个概念关联式，则对于关联式的选择必然有优先级问题。总的原则是尽量选择包含领域信息的关联概念来替换原来的不含有领域信息的概念，并且按照强关联>强交式关联>包含、属于>源、流关联的顺序。

经过概念的两级压缩，如果词汇链压缩之后不含有任何的领域信息，则该词汇链所表达的主题领域被划分为一般领域，也就是无领域类别，否则词汇链将被压缩成仅含有领域基元的多个组成项，多个组成项根据概念树向上归并，寻找共同的父类节点，如果父类节点的领域粒度太大不满足分类需求，则利用词汇链组成项的权重值确定优先级，在词汇链关键词选获取中已对每一项的主题贡献度按其权重排序，利用每一项的权重值选择优先级，对于高优先级的多个领域按照领域概念树进行向上归并到指定的领域层级。

用以下的一小段文字进行说明，“针对我国群众对关注身体健康的热潮，尤其是对老年人和婴幼儿的健康，提出需要加强药品在其生产过程中的规范性，医用器具的安全性，以及药品使用的合理性；身体健康不仅要注重营养补充，更不能忽视平日的锻炼，提高免疫力；不存在医疗流派歧视，重视西医例如B超、手术等高级的医疗手段的同时，也不能忽略中医的按摩、草药等方式。”

上述小段文字由于每个词汇的词频较低，这里选切割后的所有符合条件的词语参与相似度计算，首先按照文献[16]算法进行词汇链抽取，这里为了强调区分度我们取相似度阈值为0.65，则抽取的强词汇链为L1：{健康，营养，免疫力，锻炼，按摩}，L2：{药品，医疗，医疗器具，医疗流派，西医，B超，手术，医疗手段，中医，草药}，词汇链中的词汇映射HNC知识库，这里可以忽略其他的概念基元，只选取存在领域信息的概念基元，例如“药品”的符号化为ws422+ga823\1，则对其领域概念基元提取之后“药品”的领域基元为ga823\1，词汇链L1两级压缩之后为{a819,a81ab,a81a}，则词汇链L1的领域项向上归并共同的父类节点是a81（保健）；词汇链L2的两级压缩之后的领域项为{ga823\1, ga823\2, a827\k, a827\1, a829, a82te22, a827, a823}，则词汇链L2的领域项向上归并的共同父节点是a82（医疗），所以词汇链L1和L2的三级领域粒度划分为保健和医疗两大类，则由这两条词汇链产生的知识元都相应的分别归入保健和医疗两类中，如果选择二级粒度，再向上归并，则可划分为a8（卫生）类。

4．知识元的语义链接

知识之间的链接关系是知识组织的必要环节，否则知识单元将相互脱离，成为难以被利用的知识孤岛。知识链接就是利用一些规则或是关联技术将不同的知识节点相互关联，使其结构化、有序化，同时揭露知识之间本质上的联系，建立起领域知识之间的关系网，用户通过这个关系网查阅一个知识的同时，能够通过这种网状关系准确获取所需知识。

知识关联从知识组织的角度出发，将处理加工的分散知识重新整合的过程，知识单元之间的关联从不同的出发点可以分为不同的类型，我们从知识内部和外部把知识关联分为两种：一种是元数据项的关联，是一些文献内容以外的信息的关联，例如作者关联、机构关联、学科领域的关联、关键词的关联等形式，主要是同一作者可以关联对多篇文献单元联，同一机构、同一学科分类下的文献单元相互关联，以及利用分类—主题词表等形式的关联。另一种是基于知识元内容的链接，主要是针对文献内部的知识点之间的关联，知识元的链接也分为显式链接和隐式链接，即逻辑上的关联和语义上的关联，显式链接可以从知识元的逻辑关系中获取，隐式链接是知识元深层的关联，基于相关联性分析的语义层面的链接。

属性“来源”关联知识元与文献单元，“导航”属性是知识元所来源的文献内部的段落信息，可以关联知识元之间的上下文关系，知识元的这种逻辑链接可以直接在知识元抽取的过程中完成。知识元的语义链接分为三类：

（1）同一文献内同主题链接

“主题”属性是指知识元所归属的文献主题，同一文献内的同主题在这里用词汇链表示，通过词汇链的形式把文献内同一主题的不同的知识元关联，如图2所示。

在图2中P1表示的是文献单元，LC_i表示文献中抽取的词汇链，KL_i表示的是知识元的名称，KW_i是词汇链中的相关联词汇信息，文献所表达的的主题被分割成多个子主题，每一条词汇链即是对子主题的表达，每条词汇链又由一系列相关联的词汇信息组成，从词汇链中抽取主题词作为知识元的名称，一篇文献中的知识元通过词汇链而相关联，形成对某一主题的表达。基于词汇链的同一主题的链接是对文献深入分析的结果，脱离了文献的形式表达，以词汇链作为中间层形成了知识元-知识元关联以及知识元-文献单元关联。

（2）不同文献同主题领域链接

文献内词汇链所表述的主题把同一文献不同的知识元关联在一起，而不同文献的知识元关联，需要通过不同文献的词汇链主题领域的分割进行关联，从而把主题下的知识元关联，词汇链主题领域的分割把文献内部以及文献与文献之间的知识元相互链接在一起，如图3所示。

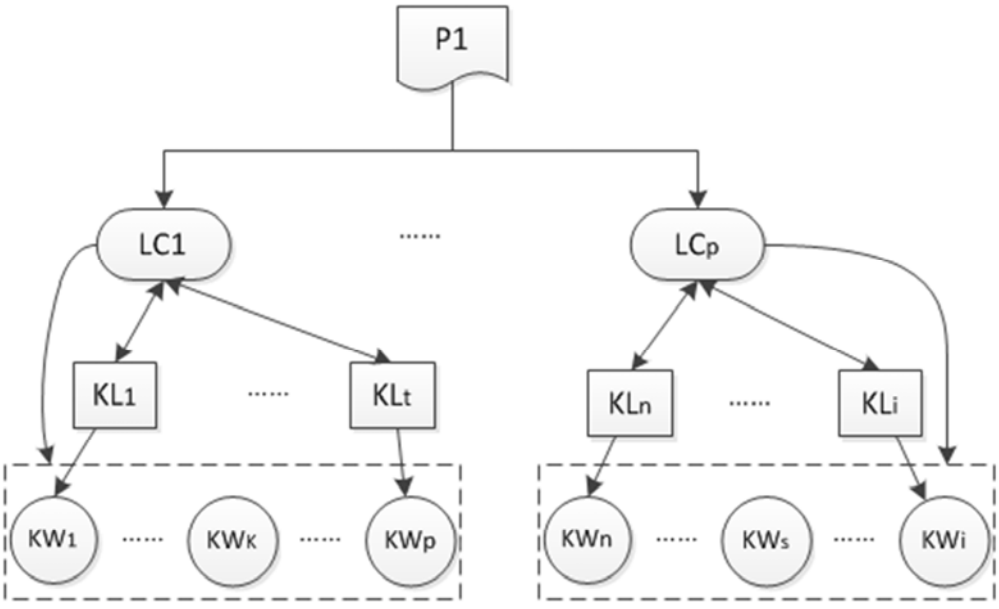

图2 词汇链对知识元的链接。

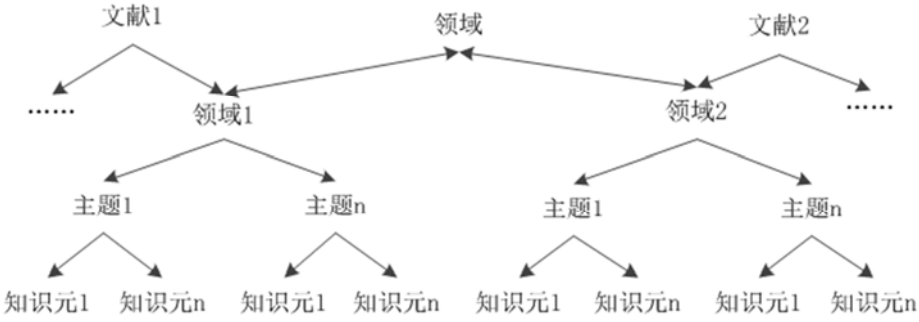

图3 同主题领域的链接。

（3）HNC概念关联式链接

HNC概念具有一种天然的对语义关联表达的优势，概念符号设计的出发点就是语义联想，重要的准则有同行优先准则、交式关联和链式关联，例如b类概念（追求活动）与a类概念（专业活动）的关联性很强，两者互为因果；d（规约性活动）类概念与7（心理活动及精神状态）特别是72（精神状态）强交链式关联。HNC概念关联知识库定义10种关联类型[17]，例如强关联、包含、属于等，具体可参考表3。

表3 10种逻辑关联类型。

关联类型	符号	示例
强关联	≡	a219\10*b\25≡a42
强交式关联	＝	3099＝107a
强流式关联	＜＝	j112＜＝53
强源式关联	＝＞	7103^e46d01＝＞a60
包含	%＝	q701e22%＝a72^e21
属于	＝%	a228i\3＝%a59a
对应	:＝	a11e1ne223:＝a109
等同	＝:	a15＝:a13\1d01
定义	::＝	73228::＝(7322,183,d22)
虚设	＝＝	a103e22＝＝a143

知识元主题词映射概念关联知识库，可以发现与其强关联或是属于、包含、源流等形式的关联关系。HNC概念关联是脱离知识元的形式化描述，从语义层面发现知识元之间的隐式关系，全方面的表述知识元的静态和动态关联，促进多科学多领域的知识发现和增值创新，把握知识发展脉络。

5．知识元库构建

知识元的表示、抽取、整理、关联是对文献知识结构化、有序化的过程，知识组织的最终目的是服务知识化。本节对获取的知识元及其关联知识进行合理化存储，构建知识元库，为用户直接获取文献内部知识提供便利。

知识元库是由一个个独立的、完整的、唯一的知识元组成的数据库，每一个知识元都是一条可理解的信息的知识化表示。在知识元表示及抽取、知识元主题领域划分、知识元链接等步骤实现之后，对已经归类的知识元分别进行知识元库的构建，从而将知识元永久化存储。

利用关系型数据库SQL Server 2012对知识元存储，知识元数据表是知识元库的基本表，知识元表结合关联表和一些辅助信息表构建知识元库。表4是知识元数据表，存储了知识元的基本信息：知识元的编号、文献来源、上下文信息、知识元名称、知识元内容、主题和主题领域。

表4 知识元数据表。

字段名	类型	说明
KE_id	数字	知识元的编号
KE_source	文本	文献来源
KE_context	文本	在文中的上下文信息
KE_name	文本	知识元名称
KE_content	文本	知识元内容
KE_domain	文本	知识元主题领域

知识元的辅助信息表最重要的是文献词汇链表，它记录了每篇文献的主题信息和主题领域，是知识元检索的重要导航信息。表5是文献词汇链表，它记录了词汇链编号、文献来源、词汇链内容和主题领域。

表5 词汇链表。

字段名	类型	说明
LC_id	数字	词汇链编号
LC_source	文本	词汇链文献来源
LC_content	文本	词汇链内容
LC_field	文本	词汇链主题领域

知识元数据表、词汇链表、文献表和其他的辅助信息表构成了文献的知识元库，词汇链表为知识元提供了主题领域信息，不同文献中的知识元由同主题领域而关联，可以通过对主题领域的搜索，获取领域下的知识元信息，直接用来学习。也可以通过对知识元的检索，发现知识元在不同领域下的存在状态和研究点，促进知识发现和知识融合。

6．结论

随着信息技术和互联网技术的发展，信息资源也在成爆炸式增长，如何从海量的信息资源中获取人们所需要的知识，成为一个亟待解决的问题。合理的知识组织方式是问题解决的关键，只有通过知识组织把信息资源知识化，才能为用户提供知识化的服务。知识的组织管理随着发展的需要，也从以文献为单位的管理发展到现在以知识为单位的组织管理，以知识元的形式作为知识的基本单位是近几年情报学领域专家学者所广泛倡导的。本文针对期刊文

献，以知识元作为基本的组织单位，以HNC理论作为主要工具，对期刊文献的知识组织进行了探索，提出了知识元五元组的表示方式，引入词汇链把知识元与文献联系起来，并基于HNC的领域概念树把主题领域划分到不同的领域粒度，描述了由词汇链、主题领域、HNC概念关联式对知识元之间的语义链接作用，最后通过构建知识元库来进行知识组织和检索。

随着研究的深入，本文会进一步利用HNC理论的句子语境理论和语义块理论对句子语义分析，改善主题词和主题句的抽取；同时在构建知识元库时不仅仅利用知识元的链接关系，还可以利用其他文献相关信息使得知识元库相互关联，形成知识仓库，以便为用户提供知识检索、知识推送、知识挖掘等更为多样化的丰富的知识服务。

参考文献

[1] 蒋玲.面向学科的知识元标引关键技术研究[D].武汉：华中师范大学.2011。

[2] 温有奎，徐国华.知识元链接理论[J].情报学报,2003,22(6):665-670。

[3] 温有奎.基于"知识元"的知识组织与检索[J].计算机工程与应用,2005,41(1):55-57,91。

[4] 姜永常,杨宏岩,张丽波等.基于知识元的知识组织及其系统服务功能研究[J].情报理论与实践,2007,30(1):37-40。

[5] 姜永常.基于知识构建的数字图书馆知识服务研究[D].哈尔滨:黑龙江大学,2007。

[6] 司莉,李月婷.我国三大全文数据库知识链接方式比较分析[J].图书馆建设,2013,(4):39-41,46。

[7] 于秀慧,李宝山.基于知识元的知识管理[J].山东图书馆学刊,2013,(1):14-17。

[8] 吕颖.知识链接及其在网络知识服务中的应用研究[D].湘潭：湘潭大学,2015。

[9] 陈烨,赵一鸣,姜又琦.基于关联数据的知识组织研究述评[J].情报理论与实践,2016,(02):139-144。

[10] 高俊峰.基于语义标签的数字文献资源知识组织方法研究[J].现代交际,2016,(01):1。

[11] 刘术华,牛现云.移动阅读环境下图书馆知识组织与服务模式研究[J].图书馆杂志,2016,(04):27-30，36。

[12] 陈树年,李青华,朱连花.近年来我国信息组织研究进展及趋势[J].图书馆建设,2006,(03):62-67。

[13] 王宇,刘淼.一种基于知识元的期刊文献知识仓库构建[J].情报理论与实践,2013,(08):91-94。

[14] 黄曾阳.HNC(概念层网络)理论—计算机理解语言研究的新思路[M].北京:清华大学出版社,1998。

[15] 缪建明,张全,赵金仿等.基于文章标题信息的汉语自动文本分类[J].计算机工程,2008,34(20):13-14,17。

[16] 王宇,伍力慧.基于HNC理论的中文文本词汇链构造方法[J].情报杂志,2016,35(2):182-187。

[17] 池哲洁,张全.基于HNC概念关联性的领域判定研究[J].中文信息学报,2013,27(6):45-50。