图书情报工作
第57卷 第9期 2013年5月

经典古籍注疏文献的知识网络研究与设计*

马创新 陈小荷 曲维光

[摘 要] 注疏文献中蕴含着丰富的知识,为了能够利用计算机分析经典古籍和注疏文献中的信息,实现知识的自动重组和聚类,分析注疏文献中存在的问题,提出使用结构化的知识表示方法组织经典古籍和注疏文献中的知识。并结合经典古籍注疏文献知识网络的基础框架结构,探讨经典古籍注疏文献知识网络中的知识组织方式和应用价值。

[关键词] 古籍数字化 知识网络 知识表示 知识组织
[分类号] G255.1

1 引 言

中华文化博大精深、源远流长,在先秦,孔子、孟子和其他诸子百家就开创了中国历史上第一次文化学术的繁荣时期。这个时期出现了儒家、道家、法家、墨家等流派,同时出现了《论语》、《孟子》、《左传》等对后世影响深远的作品,这些经典古籍是中华民族精神文明的伟大结晶,具有极高的文学价值和历史价值。经典古籍对于中华文明的传承发挥了极大的作用,历代学者对它们的研究不胜枚举。

语言文字的意义是随着时间发展变化的,经典古籍流传久远,后人难以看懂,所以在汉代就出现了对经典古籍中字、词、句的含义做解释的"注"。南北朝时期,因为语言的发展变化,有些前人所作的"注"到这时阅读起来也有了意义方面的障碍,于是就出现了对前人的注文再作注的"疏","疏"既解释原典也解释前人的注解。"注"和"疏"被统称为"注疏文献"[1]。

注疏文献中蕴含着文字、音韵和训诂等方面的丰富知识,在研究某一经典古籍时,通过分析它的注疏文献可以得到大量的有用信息。但在各种注疏文献中存在一些问题,它们影响了计算机对注疏文献的处理和分析,削弱了注疏文献的使用价值,这些问题有:

- 注疏文献是半结构化的。注疏文献中的"经"、"注"和"疏"是合在一起的,这三类内容之间并无通用的区分标志,注释语句中的"注释对象"和其他成分也无明显的区分特征。
- 注疏文献中的"解释类成分"也具有了一定的元语言思想。元语言,"就是描述语言的语言,它通过定义一套描述文档结构与含义的语法标记,使人或计算机能够利用这些标记快速准确地找到并理解文档中包含的特定语义信息。"[2]但在注疏文献中,"解释类成分"和"被释成分"使用了同一种语言,在形式上两者之间无法自动区分,因而无法准确找到特定信息。
- 由于知识的传承与积累,经典古籍的多种注疏文献之间是有联系的,但孤立地研究某一部注疏文献,无法充分挖掘出注疏文献之间丰富的联系。

因此,要想利用计算机挖掘注疏文献中的丰富知识,将训诂学家们的研究成果充分应用于经典古籍的研究,就必须要做基础性的工作,即对注疏文献做前期的处理和加工,把注疏文献从半结构化变为全结构化,添加在形式上不同于"被释成分"的元语言,并且在多部注疏文献之间构建起固定的联系。现今在这方面已经有了一些研究,比如:文献[3]研究经典古籍与其注疏文献句子对齐以及注疏文献中注释语句的自动分析方法;文献[4]介绍构建《论语》与其注疏文献对齐语料库的设计思路、基本方法和应用价值,目的就是在《论语》与其各部注疏文献之间建立起一种固定的联系;文献[5]介绍使用结构化元语言 XML 描述《论语》与其注疏文献对齐语料库中的知识,把注疏文献转化成结构化的 XML 格式,并且在原文中添加有意义的标志信息。

*本文系国家社会科学基金重大项目"汉语史语料库建设研究"(项目编号:10&ZD117)和江苏高校重点研究基地重大项目"先秦文献词汇知识挖掘"(项目编号:2010JDXM023)研究成果之一。
[作者简介] 马创新,南京师范大学文学院博士研究生,E-mail::machxin@126.com;陈小荷,南京师范大学文学院教授,博士生导师,博士;曲维光,南京师范大学计算机科学与技术学院教授,博士生导师,博士后。
收稿日期:2013-03-06 修回日期:2013-04-21 本文起止页码:124-128 本文责任编辑:徐健

基于此研究背景和研究基础，本文探讨如何利用知识网络这一结构化知识表示方法组织经典古籍和注疏文献中的知识。

2 知识网络

知识表示方式能够影响到知识在使用过程中的完备性、共享性和有效性。在人工智能和信息技术领域存在着多种知识表示方式，比如：基于逻辑的表示方式、产生式表示方式、框架表示法、面向对象表示法、基于本体表示法、知识网络表示法等，其中后 4 种表示方式属于结构化的知识表示方法，都具有良好的结构性和层次性[6]。此外，基于框架表示法还具有良好的继承性，面向对象表示法具有良好的封装性和模块性，基于本体表示法具有共享性和重用性的优点。

具有良好推理性和灵活性的知识网络，是由多学科相互结合、交叉渗透而形成的研究领域，它是实施知识管理的重要工具和平台，同时也是有效开发和利用知识资源，进行知识创新的重要途径[7,8]。“知识网络目前还没有明确的定义，它是一个集合概念，指的是知识、信息及知识间联系有关的一类网络。”[9]一般认为，知识网络是由众多的知识节点与知识关联构成的集合。其中，知识节点是知识单元的存储单位，是在认识上可以相对独立存在的各种知识单体形态，比如：论文、学科、概念或者词语等。知识关联是指构成知识网络的知识节点与知识节点之间的联系，通过知识关联将具有同一、隶属、相关关系的单元知识，按照一定的需要有序地联系在一起，形成序列化或结构化的知识集合[10]。

构建经典古籍注疏文献的知识网络就是指利用计算机技术、信息技术等新兴技术手段，再结合当代计算语言学方法，对蕴含在经典古籍和注疏文献中的知识进行多元组合，使其成为一个结构化的知识集合和立体的信息网络。

3 基础框架结构

经典古籍注疏文献的知识网络基础框架结构（见图 1）是由一个“经典古籍索引库”和多个“经典古籍与其注疏文献对齐语料库”所构成的。“经典古籍索引库”把多部经典古籍中的词语联系起来，通过倒排索引表可以查找到词语所在的目标文献及其位置。“经典古籍与其注疏文献对齐语料库”通过把经典古籍中的每一组句子，与其注疏文献中引文（即引用经典古籍的内容）的相应句子对应起来，能够在具体某部典籍与其多部注疏文献之间建立起联系，也能够使多部注疏文献，通过典籍这个中间枢纽，相互之间建立起多向的联系。各部经典古籍既存在于“经典古籍索引库”中，也存在于“经典古籍与其注疏文献对齐语料库”中，是联系两者的纽带。

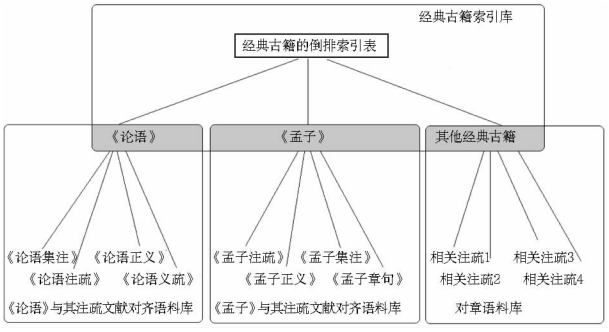

图 1 经典古籍注疏文献的知识网络基础框架

“经典古籍索引库”与“经典古籍与其注疏文献对齐语料库”有机结合在一起，就能在各部经典古籍之间、经典古籍与注疏文献之间以及各部注疏文献之间构建起全面的联系，如图 2 所示：

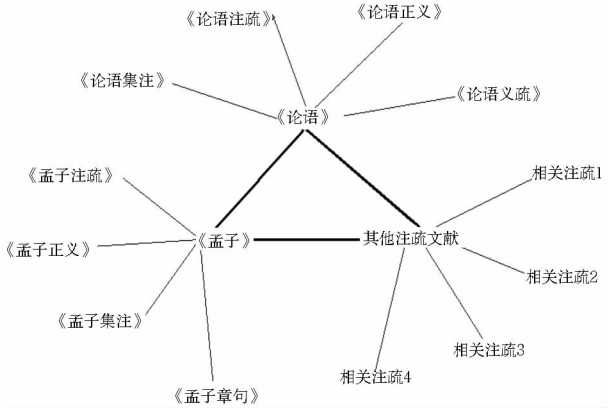

图 2 经典古籍注疏文献之间的联系

3.1 构建经典古籍索引库

对选定纳入知识网络的经典古籍建立基于词语的倒排索引，索引表的格式示例为：

诸：{(2, 3)}

祈祷：{(0, 1), (0, 4), (1, 1), (2, 1)}

战栗：{(0, 0), (0, 3), (1, 2), (2, 0)}

其中“诸：{(2, 3)}”表示的是“诸”存在于第三个文档里，而且在第三个文档中的位置是第四个词语。

“经典古籍索引库”能够把各个经典古籍中以线性方式存在的“词语的序列”作为“经”，把不同经典古籍中“相同词语的联系”作为“纬”，在各部经典古籍之间构建起相互联系的网络[11]。

3.2 分别构建各部经典古籍与其注疏文献的对齐语料库

注疏文献具有相对固定的行文体例和半结构化特

征，一般注疏文献的行文方式是先引用原文中的一句或几句话，其注文和疏文跟在所要注释的文字后面。可以把注疏文献看作是由多个信息块组成的序列，各个信息块均是由“引文和相应注释”组成。信息块中的“注释”又是由多条注释语句组成，每条注释语句都至少有一个“注释对象”，注释语句中的注释对象有多种类型，它可能是“字”，或是“词”，或是“短语”或是个句子。在大部分情况下，注释对象都是一个词语。

在各部经典古籍与其注疏文献的对齐语料库中，把注疏文献中的“引文”与经典古籍原文对应起来，并对与“引文”处在同一个信息块中的“注释”做多角度分析，找出每条注释语句的“注释对象”和“注解目的”，并把“注释对象”与“引文”中相应的“被解释成分”关联起来。为提高工作效率，可以设计相关算法，利用计算机完成经典古籍与其注疏文献的自动句子对齐和注释语句的自动分析，并把计算机处理的结果存入数据库中，然后再以人工校对。限于篇幅，关于注疏文献中注释语句的自动分析方法请参见文献[3]，关于构建经典古籍与其注疏文献对齐语料库的详细介绍请参见文献[4]，此不赘述。

在一部注疏文献中实现了注释语句分析，就可以很容易地找到引文中某个字、词或短语的注释；如果在一部经典古籍与其多部注疏文献之间实现了句子对齐，同时这多部注疏文献又各自实现了注释语句分析，那么在这多部注疏文献的所有注释之间就建立起了联系，通过这种联系，可以把指定注释对象在这多部注疏文献中的不同注释汇集起来，以便于综合比较、辨析正误；如果再结合“经典古籍索引库”，就能把位于不同经典古籍的注疏文献中的注释语句联系起来，通过这种联系，可以找到指定注释对象在相关的不同注疏文献中的注释。

4　知识组织方式

4.1　知识节点

在经典古籍注疏文献的知识网络中，是以“资源－属性－属性值”的形式描述各个节点信息的[12]，节点的资源有两类：一类是经典古籍中的词语，另一类是注疏文献中“引文部分”的词语。每个节点都有三个属性，分别是：①资源标识属性（RID）：表示节点在知识网络中的位置，它的值具有唯一性。②资源注释属性（RNO）：当某个节点的词语资源来自经典古籍时，这个节点资源注释属性的值为空。当某个节点的词语资源来自注疏文献中的“引文部分”时，如果有注释语句与这个词语相关联，其注释属性的值就是与其词语资源相关联的注释语句；如果没有注释语句与之相关联，其注释属性的值为空。③资源注释类型属性（RNT）：当某个节点的资源注释属性不为空时，本属性表示注释语句的类型，如标音、释义等。

4.2　知识关联

在经典古籍注疏文献的知识网络中，知识节点之间是通过“相邻关系”和“等同关系”两种方式联系在一起的。

4.2.1　相邻关系　可以把各部经典古籍看作是由“词语”链接而成的一维的线性序列，知识网络中节点的资源就是这些词语，当两个节点的资源处在同一线性序列之中并且具有相邻关系时，这两个知识节点之间就存在一条边。注疏文献中的各个“引文部分”也被看作是由“词语”构成的线性序列，当两个节点的资源处在同一个“引文部分”之中并且具有相邻关系时，那么这两个节点之间存在一条边；并且在一部注疏文献中，如果 A 和 B 是相邻的两个“引文部分”，当一个节点的资源是“A 的最后一个词语”，而另一个节点是“B 的第一个词语”时，这两个节点之间也存在一条边。

4.2.2　等同关系　如果两个节点的资源相同，并且它们存在于不同的经典古籍中时，那么这两个节点之间存在一条边。如果两个节点的资源相同，并且一个节点的资源存在于一部经典古籍中，另一个节点的资源存在于这部经典古籍的注疏文献中时，那么这两个节点之间存在一条边。

如果两个节点的资源相同，但是它们都存在于同一经典古籍中，或者都存在于同一注疏文献中，或者分别存在于不同的注疏文献中时，那么在这两个节点之间不存在因等同关系而连接的一条边。如果两个节点的资源相同，但是一个节点的资源存在于一部经典古籍中，而另一个节点的资源存在于其他经典古籍的注疏文献中时，那么在这两个节点之间也不存在一条边。

基于上述规则，与每个节点相连的多个边中，因等同关系而连接的边会有一条或者多条，因相邻关系而连接的边一般情况下有两条，而当节点资源是文献的第一个或者最后一个词语时，就只有一条因相邻关系而连接的边。

4.3　各种知识节点的具体关联方式

图 3、图 4 和图 5 中的点表示知识网络中的节点，线段表示边，空心点表示该节点中的资源存在于经典古籍中，实心点表示该节点中的资源存在于注疏文献中，粗线段表示节点之间的联系是因等同关系，细线段

表示节点之间的联系是因为相邻关系，虚线段表示中间内容省略。在表 1、表 2 和表 3 中，用 m 表示在经典古籍中的词例（tokens）数目，用 n 表示在注疏文献中“引文部分”的词例（tokens）数目。

4.3.1 两个节点资源都存在于经典古籍中时的关联方式 表 1 和图 3 结合起来展示了当两个节点中的资源都存在于经典古籍中时的 4 类关联方式。

表 1 两个节点资源都存在于经典古籍中时的 4 类关联方式

方式	两个节点中的资源	经典古籍	两个节点之间的最短路径
1	相同	同一	至少通过一条边（图 3A），至多通过一个中间节点和两条边就能够形成联系（图 3B）
2	相同	不同	总是通过一条边就能形成联系（图 3C）
3	不同	同一	至少通过一条边（图 3D），至多通过 m－2 个中间节点和 m－1 条边就能形成联系（图 3E）
4	不同	不同	至少通过一个中间节点和两条边就能形成联系（图 3F），在极端情况下，无法形成联系

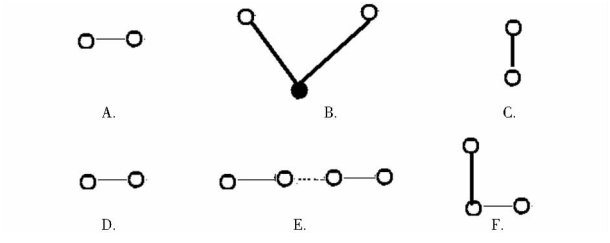

图 3 两个节点资源都存在于经典古籍中时的知识关联

4.3.2 两个节点资源分别存在于经典古籍和注疏文献中时的关联方式 表 2 和图 4 结合起来展示了当一个节点资源存在于经典古籍中，而另一个节点资源存在于注疏文献中时的 4 类关联方式。

表 2 两个节点资源分别存在于经典古籍和注疏文献中时的 4 类关联方式

方式	两个节点中的资源	经典古籍与注疏文献	两个节点之间的最短路径
1	相同	相关	总是通过一条边就能够形成联系（图 4A）
2	相同	不相关	总是通过一个中间节点和两条边就能形成联系（图 4B）
3	不同	相关	至少通过一个中间节点和两条边（图 4C），至多通过 m－1 个中间节点和 m 条边就能形成联系（图 4D）
4	不同	不相关	至少通过两个中间节点和三条边就能形成联系（图 4E），在极端情况下，无法形成联系

4.3.3 两个节点资源都存在于注疏文献中时的关联方式 本网络中心位置的节点资源都来自于经典古籍，而注疏文献中的词语都作为外围节点的资源，因此，当两个节点资源来自注疏文献中时，它们必须以网络中心位置的节点作为中间节点才能联系到一起。表 3 和图 5 结合起来展示了当两个知识节点中的资源都存在于注疏文献中时的 6 类关联方式。

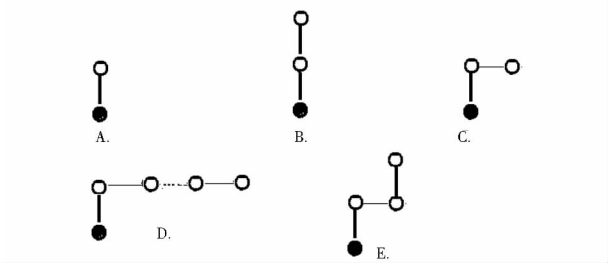

图 4 两个节点资源分别存在于经典古籍和注疏文献中时的知识关联

表 3 两个节点资源都存在于注疏文献中时的 6 类关联方式

方式	两个节点中的资源	经典古籍	注疏文献	两个节点之间的最短路径
1	相同	同一	同一	至少通过一条边（图 5A），至多通过一个中间节点和两条边就能形成联系（图 5B）
2	相同	同一	不同	总是通过一个中间节点和两条边就能形成联系（图 5C）
3	相同	不同	不同	总是通过两个中间节点和三条边就能形成联系（图 5D）
4	不同	同一	同一	至少通过一条边（图 5E），至多通过 n－2 个中间节点和 n－1 条边就能形成联系（图 5F）
5	不同	同一	不同	至少通过两个节点和三条边（图 5G），至多通过 m 个中间节点和 m＋1 条边就能形成联系（图 5H）
6	不同	不同	不同	至少通过三个中间节点和四条边就能形成联系（图 5I），在极端情况下，无法形成联系

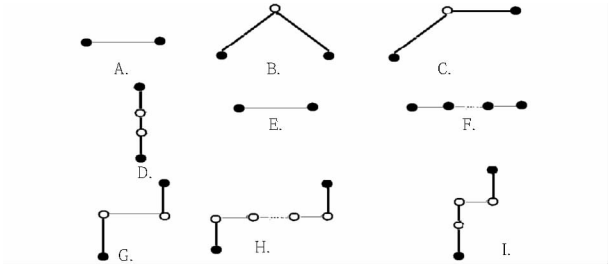

图 5 两个节点资源都存在于注疏文献中时的知识关联

5 应用价值

经典古籍注疏文献的知识网络具有广泛的应用价值，能够为语言研究者提供经典文献研究的新平台，其应用价值主要体现在以下三个方面：

5.1 知识重组

古籍文献中的知识原本是以文字符号的形式存在的一维的线性序列，文献之间没有显性的联系。本网络把文献中的词语作为节点的资源，并且加上了属性信息，以等同关系和相邻关系连接知识节点，把多个一维的线性序列组合成为一个多维的知识网络，把多部文献之间原有的隐性联系变为显性联系，构成了结构化的知识集合，通过知识重组实现了知识创新[13]。

5.2 知识聚类

在本网络中，因等同关系连接的知识节点会把同

形词聚集在一起;因相邻关系连接的知识节点会把固定的词语搭配聚集在一起;分析节点的资源注释属性,可以把同一注释对象的多个注释语句聚集在一起。如果把本网络中的各类信息综合在一起,可用于大规模的词语聚类和词义聚类[14]。

5.3 知识挖掘

基于本网络,可以考察古籍文献中的词例(tokens)、词型(types)、常用词、常用词语搭配在网络中的分布规律,考察各部古籍的用词特点以及训诂术语的使用特点,等等。

6 结　语

经典古籍的知识管理可以分为"录入与存储"、"网络传播与信息检索"、"基本信息标注与系联"、"语义标注与知识检索"共4个方面的内容,其中前两个方面的内容属于"表层知识管理",后两个方面的内容属于"深层知识管理",表层知识管理主要解决经典古籍的数字化存储、网络传播和全文检索的问题,而深层知识管理深入到"内容和意义"层面研究古籍文献,包括研究古籍著录和描述的元数据标准、古籍内部知识元的标注问题、知识元之间的联系方法以及古籍之间的联系方法等。

经典古籍的知识管理研究发展到今天,已经表现出由"表层知识管理"向"深层知识管理"发展的趋势。表层知识管理研究起步较早,至今已取得丰硕成果,一些疑难问题基本上得到解决,而深层知识管理研究是建立在对古籍内容本身有着较为深入理解的基础之上的,牵涉的学科领域更广泛,问题更复杂,难度更大,要求研究者具备更高的技术水平和更全面的学术素养。

经典古籍注疏文献的知识网络属于深层知识管理的研究范围,它的构建需要经过理论探索、框架设计、语料库建设、知识节点划分和知识链接等多个阶段,其建设工程浩大、技术复杂。本文初步探索经典古籍注疏文献知识网络的基础框架结构、知识组织方式和应用价值,以为下一阶段的深入研究打下坚实的基础。

参考文献:

[1] 许威汉. 训诂学读本[M]. 上海:上海交通大学出版社,2010:48-84.
[2] 胡佳佳.《说文解字》语料库的XML标注设计[J]. 社会科学论坛,2011(7): 214-223.
[3] 马创新,陈小荷,曲维光. 注疏文献中的注释语句自动分析[J]. 计算机科学,2012(10):220-223.
[4] 马创新,陈小荷,曲维光,等.《论语》与其注疏文献对齐语料库的构建[J]. 现代教育技术,2012(7):109-113.
[5] 马创新,陈小荷. 基于XML的《论语》与其注疏文献对齐语料库的知识表示[J]. 图书情报知识,2013(1):107-113.
[6] 王珏,袁小红,石纯一,等. 关于知识表示的讨论[J]. 计算机学报,1995(3):212-224.
[7] 温有奎,赖伯年. 网格技术将推动知识管理革命[J]. 情报学报,2004(1):124-128.
[8] 赵蓉英. 知识网络研究(II)——知识网络的概念、内涵和特性[J]. 情报学报,2007(3):470-476.
[9] 刘向,马费成,陈潇俊,等. 知识网络的结构与演化——概念与理论进展[J]. 情报科学,2011(6):801-809.
[10] 赵蓉英. 论知识网络的结构[J]. 图书情报工作, 2007,51(9): 6-10.
[11] 宋继华,王宁,胡佳佳. 基于语料库方法的数字化《说文》学研究环境的构建[J]. 语言文字应用,2007(1): 132-138.
[12] 戴维民. 语义网信息组织技术与方法[M]. 上海:学林出版社,2008:71-75.
[13] 姜永常. 基于知识网络的动态知识构建:空间透视与机理分析[J]. 中国图书馆学报, 2010(4):115-124.
[14] 常娥,黄建年,侯汉清. 古籍智能整理与开发系统构建研究[J]. 情报资料工作,2009(4):43-47.

Study and Design on Knowledge Network of Classical Ancient Books and Commentary Literatures

Ma Chuangxin[1]　Chen Xiaohe[1]　Qu Weiguang[2]
[1]College of Liberal Arts, Nanjing Normal University, Nanjing 210097
[2]College of Computer Science and Technology, Nanjing Normal University, Nanjing 210097

[**Abstract**] The commentary literatures contain a wealth of knowledge. In order to analyze the information in classical ancient books and commentary literatures with computers, achieve automatically restructuring and clustering of knowledge, and analyze the problems in commentary literatures, this paper proposes to organize the knowledge in classical ancient books and commentary literatures with structured knowledge representation method. Combined with the basic framework structure of knowledge network of classical ancient books and commentary literatures, this paper discusses knowledge organization mode and application value of the knowledge network.

[**Keywords**] digitization of ancient book　knowledge network　knowledge representation　knowledge organization