信息资源

基于耦合关系的馆藏数字资源语义化深度聚合研究

赵蓉英 柴 雯 （武汉大学中国科学评价研究中心；武汉大学信息管理学院 湖北 430072）

摘要 文章以计量分析中的耦合关系为例，探讨了耦合关系的语义特性及其对数字资源深度聚合能力，提出了基于耦合关系的馆藏数字资源语义化聚合模型，并进行实证研究。结果表明：利用耦合关系对馆藏数字资源进行深度聚合能很好地揭示馆藏数字资源的语义特性，可以从多个角度对文献资源进行有效组织，为用户提供资源推荐和个性化知识服务，提升图书馆知识服务的能力和效率。

关键词 耦合关系 馆藏数字资源 语义化 深度聚合

An Analysis of Library- collected Digital Resources Semantic Depth Polymerization Based on Coupling Relationship

Zhao Rongying Chai Wen

（Research Center for Science Evaluation；School of Information Management，Wuhan University，Hubei，430072）

Abstract This paper proposed a semantic integration method for library digital source utilizing coupling relationship and proved it. The results show that: coupling relationship can be a good method for revealing semantic features of digital resources, organizing knowledge from multiple angles and providing users recommendations and personal service to enhance the capacity and efficiency of library service.

Keywords coupling relationship, library digital collections, semantic, in- depth integration

20 世纪 90 年代以来，知识经济和知识管理在全球范围内普遍兴起，知识作为社会竞争中一种重要的战略资源和经济资源受到了人们前所未有的关注和重视。针对馆藏数字资源的多元化，如何将多源馆藏数字资源快速存取、有序组织形成结构化和有序化的知识资源体系，进而深度挖掘和有效利用数字化、网络化信息资源，提升用户服务质量，是图书馆适应知识经济时代的必然要求。随着人工智能、语义网等新兴技术的发展和应用，未来图书馆的发展方向必然具备知识化和智能化的特点。实现馆藏资源深度聚合的首要条件就是发现资源之间的语义关联，有效的、深层次的语义关系揭示，有助于基于语义的馆藏资源深度聚合，将不同属性的知识单元进行多维度多层次关联，形成庞大的知识体系。本文以此为切入点，研究如何利用耦合关系从语义层面上深化馆藏数字资源的知识组织，加强知识间的语义关联，提升图书馆知识服务的能力和效率。

1 馆藏数字资源语义化深度聚合研究现状与问题

根据数字资源的加工程度不同，图书馆数字资源的整合可以分为三个目标层次：数据整合、信息整合和知识整合[1]。语义化聚合正是知识整合层次上的数字资源聚合，是未来图书馆的发展方向[2]。贺德方指出，当前基于语义的聚合方式有三种：基于概念及概念关系的聚合、基于引证关系的聚合、基于科研本体的聚合。这一划分方法从语义的角度重新理解了当前馆藏数字资源的聚合模式，涵盖传统分类法、本体和引文分析三种聚合方法，为馆藏数字资源的语义化聚合提供了整体的、宏观的思路。

（1）传统分类法包括分类表、叙词表、学科导航等聚合方法已经十分常见，这种方法虽然也能够在一定程度上表达资源的相关性，如属于同一学科、同一门

本文系教育部人文社会科学基金项目“馆藏数字资源语义化深度聚合的理论与关键技术研究”（编号：13YJA870023）、国家社会科学基金重大项目“基于语义的馆藏资源深度聚合与可视化展示研究”（编号：11& ZD152）的研究成果之一。

类，但是也仅仅只揭示了资源之间浅层次的语义关联，没有办法深入资源主题，无法进行深度揭示资源内容上的相关性。

（2）本体构建是语义网技术的核心内容，可以实现数字资源语义化深度聚合，但是无论是从实例还是顶层本体出发，当前图书馆数字资源本体构建应用都局限于某一研究领域或者学科，不同专家对不同领域的概念和关系的认知难以统一使得本体具有不可重复性，同时由于研究者众多，研究者之间缺乏交流与沟通，不同专家可能针对同一学科或者领域建立了本体，使得本体具有重用性。本体的不可重复性和重用性使得对整个馆藏资源构建统一的、广泛认可的本体成为难题。针对本体的种种缺点，有学者提出了计量本体[3]，试图通过计量本体简化本体构建方法，实现馆藏数字资源的语义化深度聚合。

（3）基于引证关系的资源聚合主要是借助引证关示资源之间的关联，从而将相关资源聚合成一个相互关联的体系，如利用文献耦合关系通过引文对文献进行聚类，从而揭示学科知识结构、主题、研究热点。除了引证关系，计量分析中的很多方法也可以满足资源深度聚合的要求，在引证关系的基础上，我们还可以对信息计量学中的各种方法和规律进行分析，讨论其语义特性和聚合效果，丰富馆藏数字资源语义化聚合模式。

针对以上研究的现状与问题，本文以耦合关系为例，探讨耦合关系的语义特性和聚合效果，扩展语义化聚合的方法体系。

2 耦合关系与馆藏资源语义化深度聚合

2.1 耦合关系的基本理论

耦合是指主体之间各因素的相互关联[4]。在信息计量学领域中，耦合这一概念最早源于美国学者 Kesler 提出的“文献耦合”[5]。在对文献的计量研究中不难发现，文献之间存在着引证和被引的关系，以引文分析为基础，文献耦合对两篇或者多篇文献通过共同引用的参考文献确立相关关系，这种关系的强弱与相同参考文献的数量呈正相关。随着耦合关系研究的深入与拓展，耦合关系不仅仅局限于文献之间的引用关系，它揭示的是一类普遍存在的主客体之间的相关关系，因此可以将“文献耦合”的概念予以推广，利用耦合概念来反映著者、学科、期刊、国别、机构、时间等多种特征单元的相似性关系[6]。耦合分析使得从外表看来不相关的主体之间错综复杂的相关关系被显现出来，形成耦合网络，利用社会网络分析、因子分析、聚类分析、SPSS、SAS 等分析方法和工具，我们可以探索主体特性，分析主体间相关关系，研究主体群的结构、历史、现状和未来趋势。

根据耦合主体的不同，耦合关系可以分为文献耦合、作者耦合、期刊耦合和学科耦合[7]，它们具体所揭示的关系、特性和应用如表 1 所示。我们知道，图书馆包含了种类丰富、规模庞大的数字资源，既包括将印刷资源转化为数字形态的数字化资源，也包括图书馆购买的数据库、安装在馆内的镜像资源以及从网络渠道获取的虚拟馆藏在内的数字资源[8]。这些数字资源以图书和论文为主，构成了规模庞大的学术信息资源库。利用耦合关系，我们可以对图书、期刊论文、硕博士论文等数字资源构建多角度、深层次的关联关系，从而使得馆藏数字资源体系中的内容关联更丰富，揭示的知识更全面。

表 1 耦合关系类型

2.2 耦合与馆藏数字资源语义化深度聚合的关系

馆藏资源语义化是将语义网的思想引入图书情报领域，语义网的倡导者正是万维网的发明者 Tim Berners- Lee，他提出用一种更容易被机器“理解”的表示方法来描述网络信息资源，同时采用智能技术来利用这种表示方法所提供的便利，形成更加“智能化”的网络[9]。简言之，基于网络信息资源的有效组织，语义网运动旨在让机器能“理解”资源的知识单元并能利用一定的规则和技术自主“推理”，得到新的知识。在图书情报领域，语义主要表示为对文献中蕴含知识的多维度揭示，即从多个角度将文献的知识单元和知识单元之间的联系明确地表达出来。馆藏资源语义化聚合就是从“理解”的角度出发，通过能够表征资源内容的知识单元建立资源之间的语义相关性，从而达到馆藏数字资源的深度聚合，优化馆藏数字资源的知识组织结构。

耦合本质上是一种交叉共现关系[10]，是利用客体共现来反映主体间的相关关系，如作者关键词耦合就是通过关键词共现反映作者之间的合作关系。共现关系以邻近联系法则和知识结构及映射原则为理论基础，使共现关系能够表达资源之间的相似性，根据相似性理论，现象上的相似可以反映出本质上的相似，而相似性是进行聚合的基础，因此，资源间的耦合关系能够反映资源的聚合情况。从耦合的类型可以看出，两个主体一般从引文或者关键词的角度来建立耦合关系：一方面，文献耦合（引文耦合）是最基本的耦合关系，从引证的角度出发，具有耦合关系的文献具有一些相似特征：共同引证和追溯某一历史背景；共同继承某些科学论断和经典著作；共同商榷和研究某一值得争论的问题；共同引证某些实验数据或统计资料；同属某一学科或专业；属交叉学科或边缘学科等[7]，这些相似特征使得具有相同引文的文献之间从内容上具有一定的关联性，相同引文越多，内容的关联程度越深，因此通过引文来确定的耦合关系具有语义关联特性。另一方面，关键词是文献知识的高度总结，是作者凝练出来能够表达其研究成果核心内容的词汇，通过关键词所建立的耦合关系可以认为是从文献资源的知识单元出发，构建知识单元之间的相关关系，达到

揭示数字资源语义关系的目的。由以上分析可知，从耦合客体的知识特性可以看出，耦合关系是具有语义特性的，它满足了数字资源语义化聚合的基本要求。

3 基于耦合关系在馆藏资源语义化深度聚合模型

耦合关系能够对馆藏数字资源进行语义化聚合，本节针对馆藏数字资源，以图书和论文为主构建基于耦合的馆藏数字资源语义化聚合模型。根据馆藏数字资源的类型，本文选取文献耦合、作者耦合和期刊耦合为主要聚合方法来建立馆藏资源语义化聚合的理论模型，该模型分为四个层次，如图 1 所示。

（1）数据层：主要进行数据的收集和预处理。首先从不同的数据库中采集数据，获取期刊、论文、图书等数字资源。由于所采集的数据具有字段格式不统一、字段名不同、不便于耦合分析等问题，因此需要进行标准化处理，并针对即将进行的耦合关系建立相应的数据库，包括关键词表、主题词表、引文表、作者表和期刊表等。特别需要说明，由于仅以文献的关键词作为关键词耦合的数据来源，会使得耦合矩阵中存在大量的 0，耦合向量相似性低，难以达到良好的聚类效果。为了深度揭示内容关联，数据标准化过程中对关键词进行了有益的补充，对摘要和题名进行切分词处理，利用文献普查得到的停用词表，去除诸如“热点研究”、“进展研究”等含义宽泛，缺乏针对性的词汇，取具有实际意义的词汇补充为关键词，从而加深耦合向量间的相似性，深度揭示数字资源之间隐蔽的内容关联。

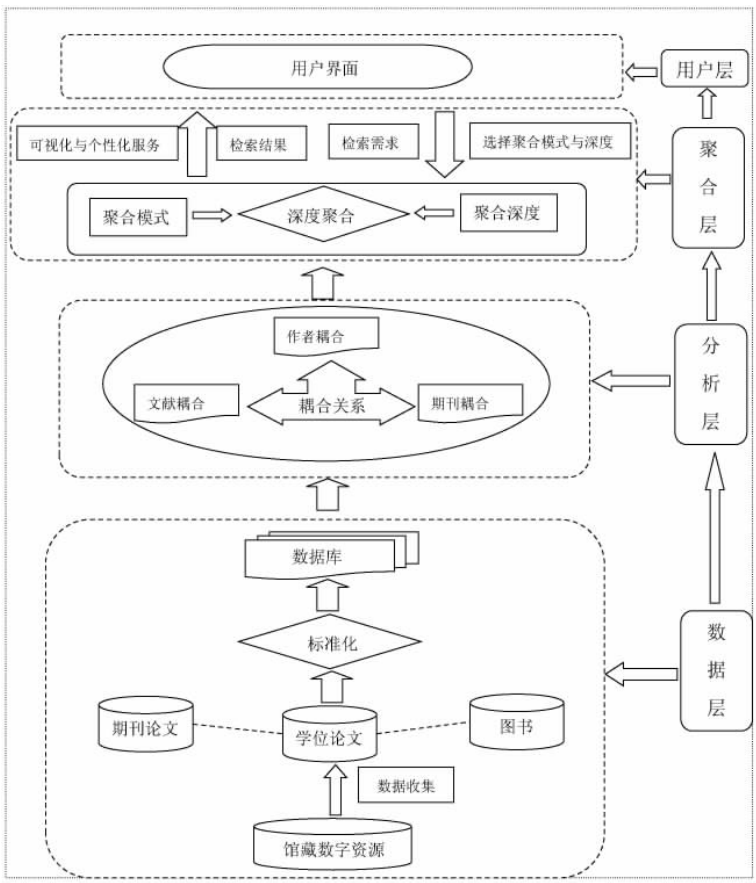

图 1 基于耦合关系的馆藏数字资源语义化聚合模型

（2）分析层：主要负责耦合关系的生成。这一层直接利用标准化数据库，借助耦合分析方法和统计分析工具，建立文献耦合、作者耦合和期刊耦合关系，形成一个语义关系集合，生成耦合矩阵甚至耦合网络。

（3）聚合层：这一层是进行语义化聚合的关键步骤。一方面，这一层为分析层提供了数字资源深度聚合的进一步指导，利用分析层所提供的语义关系，选择聚合模式和聚合深度，生成数字资源聚合结果；另一方面，聚合层面对用户的检索需求，需要将检索出的聚合结果可视化地展现给用户，并根据聚合结果生成资源推荐和个性化服务的信息，向用户提供准确、恰当的知识服务。

（4）用户层：主要负责用户需求的收集与结果输出。这一层与聚合层是紧密连接的，用户制定的检索策略会直接反应给聚合层，根据用户选择的聚合方式和聚合深度，准确得到用户需要的聚合内容，向用户返回可视化检索结果，并向用户进行资源推荐和个性化智能服务。

4 实证研究

本节以 CNKI 细胞学领域 15 311 篇论文为数据来源，选取其中被引频次高于 40 的论文（共有 171 篇）为分析对象。笔者首先根据所下载的论文题录信息构建标准化数据库，以文献关键词耦合为例，需构建论文表，包括题名、作者、基金、年份、关键词等字段。由于作者赋予的关键词个数较少且不规范，为了准确表达论文的核心内容，本文对题录中的题目和摘要部分信息进行切分词处理，去除指向性不强、指代内容过于宽泛的词汇，取剩余词汇作为关键词的有益补充。然后根据文献关键词耦合关系，建立耦合矩阵，并借助 SPSS 工具，对耦合矩阵进行层次聚类分析，得到聚类结果树状图，如下页图 2 所示。

从图中可以看出，选择不同的聚合深度，即从不同聚类步骤得到的馆藏数字资源聚合结果是不同的。以虚线所在位置为例，图中部分高被引论文被划分为 5 个簇。通过对簇内文献进行分析，得到聚合结果（见下页表 2）。从表 2 中可以看出聚类簇中的主要关键词明显体现出了簇主题，簇成员依据主题的相似性聚合在一起。同一个簇的成员之间可以进行资源推荐，图书馆也可以根据用户的定制需求，满足用户个性化的知识推送服务，定期对某些主题的新增文献进行推送。

5 结语

通过理论研究我们可以认识到，将计量分析的方法引入到馆藏数字资源语义化聚合研究中是有意义的。在馆藏数字资源的传统分类体系之上构建资源的耦合关系，使得同一学科或者领域的数字资源可以根据主题聚集在一起，使得传统分类体系得到进一步深入与细化，数字资源从内

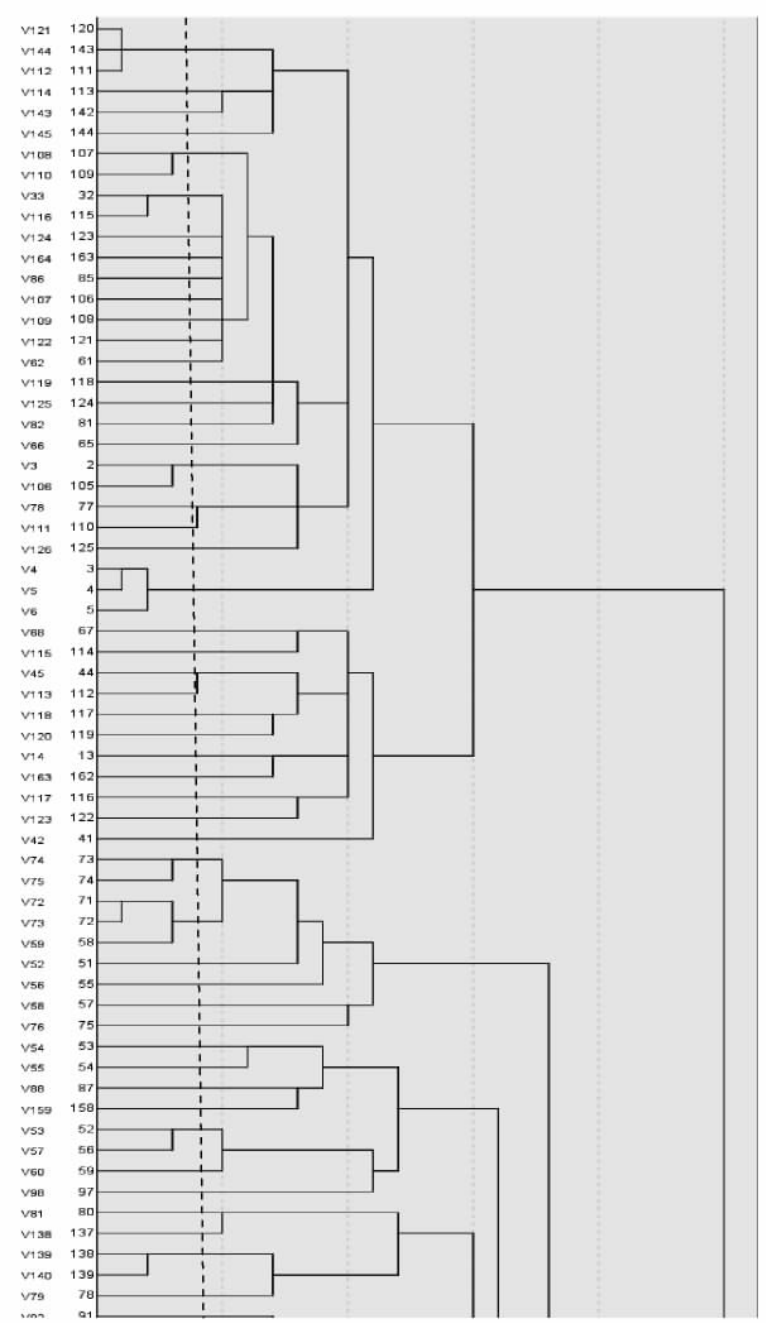

图 2 细胞学高被引论文聚类图(部分)

表 2 细胞学高被引论文聚类结果(部分)

容上密切相关, 形成统一的语义化聚合体系。但是在实证研究中, 笔者从宏观角度考虑整个图书馆的数字资源语义化聚合体系构建,分析出以下几点问题:

(1) 不同的学科或者领域, 研究主题的集中与分散情况是不相同的, 新兴学科或者领域相对于传统学科和成熟的研究领域, 其主题较分散, 主题之间的相关性也不明显。以作者文献耦合为例, 新兴领域资源的耦合分析结果中容易出现节点分布散乱、关系微弱的情况, 甚至存在孤立点, 无法对所有资源都完全归类, 影响到查全率。反之, 一个成熟的领域或者学科资源的耦合分析结果中, 一个权威或者核心作者可能与多个作者联系紧密, 一篇文献也可能与多篇文献相联系,形成一个庞大的聚类结果, 即使用户在检索时设置了一个较高的检索阈值,也有可能检索出庞大的文献量,过大或者过小会使得检索结果数量过少或过多, 在用户不知道资源主题分布的情况下, 需要用户多次的微调才能得出满意的结论,增加用户检索负担。

(2) 耦合关系类型众多, 其中文献耦合体现的是文献之间长久的稳定的相关关系, 而作者耦合关系则会随着时间的变化而变化, 图书馆资源数量庞大, 随着资源数量和类型的更新, 不断维护和建立耦合关系势必会消耗巨大的成本。

(3) 关键词规范问题。尽管通过题名和摘要对作者给出的关键词进行了有益补充, 但是通过切分词得到的关键词同作者给出的关键词一样, 存在随意性较大、近义词同义词等问题, 使得聚合结果不准确,影响信息服务质量。

参考文献

[1] 黄传慧,李 娟. 我国图书馆数字资源整合研究[J] . 图书与情报, 2009(4) : 66- 69, 82
[2] 贺德方, 曾建励. 基于语义的馆藏资源深度聚合研究[J] . 中国图书馆学报, 2012(4) : 36- 40.
[3] 邱均平,余 凡. 基于计量分析的馆藏资源语义化理论研究[J] . 中国图书馆学报, 2012(7) : 71- 78.
[4] 胡昌平. 论文献耦合[J] . 情报学刊, 1986(10) : 23- 28.
[5] Kessler M. Bibliographic coupling between scientific papers[J] . American Documentation, 1963(14) : 123- 131.
[6] 邱均平. 信息计量学[M] . 武汉: 武汉大学出版社, 2007.
[7] 罗式胜. 耦合类型与分析[J] . 图书情报知识, 1985(1) : 42- 47.
[8] 刘晓娟. 图书馆数字资源整合[J] . 图书馆理论与实践, 2007(1) : 63- 65.
[9] Grigoris Antoniou, Frank van Harmelen. A Semantic Web Primer [M] . The MIT Press, 2008.
[10] 邱均平,王菲菲. 基于共现与耦合的馆藏文献资源深度聚合研究探析[J] . 中国图书馆学报,2013(3) : 25- 33。

[作者简介] 赵蓉英, 女, 1961 年生, 武汉大学信息管理学院教授, 博士生导师,武汉大学中国科学评价研究中心副主任。
柴 雯,女,1990 年生,武汉大学信息管理学院硕士生。
收稿日期:2014- 04- 09