情 报 学 报 ISSN 1000 - 0135
第 29 卷 第 2 期 331 - 341，2010 年 4 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol.29 No.2，331 - 341 April 2010

doi:10.3772/j.issn.1000-0135.2010.02.019

基于本体的 CSSCI 学术资源网络模型构建及其应用研究[1)]

王 昊 苏新宁
（南京大学信息管理系，南京 210093）

摘要 具有语义描述能力的知识组织方式本体机制的提出和发展，为改善 CSSCI 信息服务提供了新的契机。本文在分析 CSSCI 数据和服务现状的基础上，提出了基于本体构建 CSSCI 学术资源网络模型的解决方案，即通过本体的面向对象的知识结构来组织 CSSCI 中的学术资源，以达到提高 CSSCI 信息服务质量的目的。在完整地阐述 CSSCI 本体概念模型的建立和基于概念模型的 CSSCI 数据语义标注过程的基础上，提出了专门用于 CSSCI_Onto 的评价模型，认为可以从正确性、合理性、有效性三个方面分阶段实现本体评价。最后通过具体的实践应用（包括建立基于 CSSCI_Onto 的知识检索服务平台和实现基于本体的引文分析）验证了该解决方案具有可行性和有效性。

关键词 本体 知识组织 学术资源网络模型 概念模型 语义标注 知识检索 引文分析

Research on Construction and Applications of the CSSCI Academic Resources Networks Model Based on Ontology

Wang Hao and Su Xinning
(*Information Management Department of Nanjing University, Nanjing* 210093)

Abstract It is the ontology which has the ability for describing semantics as a mode of knowledge organization has been introduced and developed, which offers a good and new chance for improving CSSCI information services. On the basis of analyzing present situation of CSSCI data and services, the solution on constructing the CSSCI academic resources networks model based on ontology is put forward, which organizes the CSSCI academic resources with object-oriented knowledge structure of ontology in order to improve quality of CSSCI information services. With entirely discussing the process of CSSCI ontology conceptualization construction and CSSCI data semantic indexing based conceptualization, putting forward the evaluation model specially for CSSCI_Onto which thinks that ontology could be evaluated by stages from three aspects of correctness, rationality and effectiveness. At last, practical applications including establishing the knowledge retrieval services platform based on CSSCI_Onto and realizing citation analysis based on ontology are developed, which verify the feasibility and effectiveness of the solution.

Keywords ontology, knowledge organization, academic resource networks model, conceptualization, semantic indexing, knowledge retrieval, citation analysis

收稿日期：2008 年 11 月 6 日

作者简介：**王昊**，男，1981 年生，博士，讲师，主要从事信息智能处理与检索、本体学习技术及应用、学科评价和引文分析等方面的研究。E-mail:ywhaowang810710@sina.com。**苏新宁**，男，1955 年生，教授，博士生导师，主要从事信息智能处理与检索、基于信息技术的情报分析与评价等方面的研究。

1） 本文系“江苏省研究生培养创新工程-科技创新计划”人文社科项目“基于本体的学术资源网络模型研究”（项目编号：CX07B-252r）研究成果之一，并得到南京大学人才引进科研启动基金的资助。

1 引 言

随着信息爆炸时代的到来，互联网创始人 Tim Berners-Lee 提出了语义网(Semantic Web)的概念[1]，用以解决当前互联网缺乏语义理解这一问题。本体(Ontology)作为一种有效的知识组织方式，被纳入语义网体系，用于在网络资源上融入计算机可以理解的信息，达到资源的语义理解。本体是解决语义层次上网络信息共享和交换的基础。

中国社会科学引文索引(Chinese Social Science Citation Index, CSSCI)自20世纪90年代末诞生以来，以其规范、权威的检索和分析服务得到了使用者的一致认同。然而 CSSCI 信息检索服务的简单化(基于关键词匹配的检索方式)和直线型的知识组织方式，使得用户在检索时很难获得准确、全面的查询结果。此外，要求更加精确、并能发现隐含知识的引文分析方法，也对传统的基于数理统计的 CSSCI 引文分析服务提出了更高的要求。而数据挖掘技术作为一种信息处理手段，只能在现有资源结构的基础上实现专深方向的剖析，不具有对学术资源间关联进行语义表达的能力，因此无法提供个性化、专业化的用户服务。

CSSCI 从全国4000多种期刊中精选出400～500种人文社会科学精品期刊作为来源期刊，收录所刊载的论文及其相关的关键词、作者、机构、期刊、学科、被引文献等学术资源。其来源数据以关系数据库的形式存储，包括来源文献表、来源文献作者表、被引文献表、期刊载文表、来源刊表以及字典表6个关系，具有庞大的数据量和丰富的实例知识。笔者收集了 CSSCI 2000～2006 年共7年的实例数据，包括近56万关键词、50余万篇来源文献、160余万篇被引文献，涉及近31万名作者。现有的 CSSCI 数据结构仅从来源文献和被引文献两个角度展示相关学术知识，而对于其他重要学术资源如作者、主题、期刊、学科、机构等知识的揭示力度不够，无法满足用户快速、准确获取知识，了解学术资源之间共现关联的需要。

CSSCI 各种资源间纷繁复杂的语义关联，如合作作者、相似文献、学科引用、同被引期刊、关联学科等，对于丰富检索方式、改变引文分析模式具有重大作用，有必要对其进行充分揭示和描述。于是，本文提出了用具有知识语义描述能力的本体机制来提升 CSSCI 学术资源服务的方案，试图借助本体对知识的有效组织从本质上改变 CSSCI 原有的数据组织结构，建立基于本体的 CSSCI 学术资源网络模型(以下简称 CSSCI_Onto)，将 CSSCI 中所有学术资源联系在一起，形成一个巨大的学术资源网络，像地图一样引导用户获得任何所需的知识，包括本体库中现存的显性知识和可以经过逻辑推理获得的隐形知识，以新的语义检索服务和引文分析模式代替原有的服务模式，解决用户需要更完善的知识服务和现有 CSSCI 提供学术资源服务相对落后之间的矛盾。

2 CSSCI 本体的系统建模

2.1 CSSCI_Onto 的特点

根据 CSSCI_Onto 构建的目的以及现有数据基础，总结其特点如下：

(1) CSSCI_Onto 的目标。CSSCI_Onto 的目标不仅包括对学术概念的知识组织，即本体概念模型的建立(包括构建抽象概念层次关系和定义类属性)，更重要的是进行知识描述，即基于本体概念模型的语义标注(Semantic Annotation)[2,3]，展现 CSSCI 中各种学术资源实例之间错综复杂甚至隐含的关系。

(2) CSSCI_Onto 的目的。CSSCI_Onto 的目的是为了更好地提供知识检索服务和引文分析服务，因此，除了将现有关系数据库结构转化为面向对象的本体结构外，还需要充分挖掘实例之间的语义关联，并计算一定的量化值作为引文分析的基础，如作者间的关联度等。

(3) CSSCI_Onto 来源数据结构特点。CSSCI_Onto 构建的基础是关系型数据库，这与目前构建领域本体多以词典或叙词表为基础有所不同。

(4) CSSCI_Onto 来源数据内容特点。CSSCI 每年收录的来源文献以万计，引用文献则以数十万计，实例数量巨大。相对于实例而言，CSSCI 中的概念(类)，如作者、文献等数量少，概念间层次关系相对明确。巨大的实例数量导致实例间存在着大量的隐含语义关系，值得领域专家去挖掘和探讨。可获得的实例知识成为本体构建的重点。

2.2 CSSCI_Onto 的构建模型和实施过程

基于 CSSCI 来源数据的特点以及学术资源本体构建的目标，本文建立了如图1所示的 CSSCI_Onto 构建模型。本体构建的具体实施过程分为6个步骤。

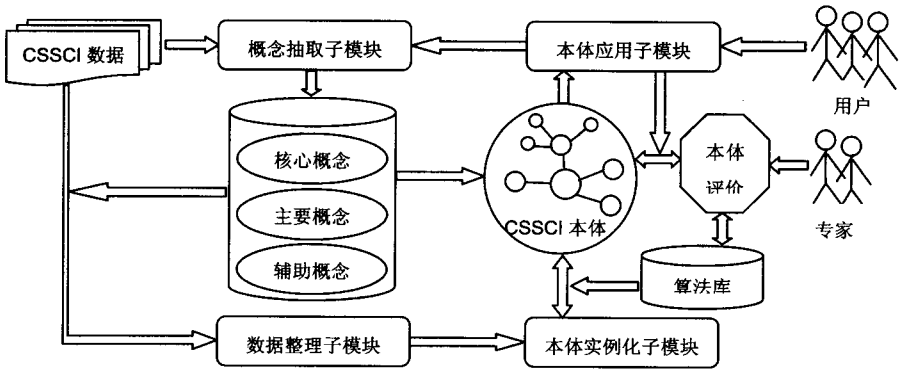

图 1　CSSCI 学术资源本体构建模型

(1)概念层次结构的建立。分析 CSSCI 元数据，从中获得各种具有考察价值的学术概念。按照一定的分类依据，基于无二义性、互不交叉、覆盖整个 CSSCI 领域的原则，抽取领域内核心概念。对核心概念进行扩展，获得主要和辅助概念，建立概念层次结构。

(2)定义概念属性。属性用于表示概念间的语义联系。属性定义实际上就是关联概念的探测过程，同时也决定了实例间存在的关系，因此可通过考察实例间存在的现实关系来归纳概念间的抽象关系，并以属性或约束的形式将关系规范化。概念属性定义和概念层次结构建立的完成标志了 CSSCI_Onto 概念模型的实现。

(3)概念实例化。根据概念模型中定义的概念层次结构和属性集合，从经过预处理的 CSSCI 原始数据中获得实例，并在算法库的支持下，为每个实例设置属性值。

(4)本体存储和展示。使用 Protégé[4] 等本体开发工具系统规划 CSSCI_Onto 的概念模型，并采用 OWL 描述，其层次结构可以使用 OWLViz、TGVizTab 等图形化插件展示。对于实例库，可采用自动生成 OWL 文件或关系型数据库等形式进行描述和存储，使用列表或图形方式展示实例间的关系。

(5)本体评价。在一般情况下，可以对照本体评价准则，如 Gruber 提出的五大准则[5]等，由领域专家和本体工程师共同实现本体评价。对于 CSSCI_Onto，笔者认为可以通过用户和领域专家从评价准则和实践应用[6]两个方面来发现问题，完善本体。

(6)本体应用。本文从基于本体的知识检索服务和引文分析服务，包括服务提供的方式、能够实现的功能以及结果可视化等方面，来探讨 CSSCI_Onto 的具体应用。

开发 CSSCI_Onto 的 6 个步骤不是采用“瀑布”模式，而是采用“循环式增量迭代”模式，即在下一个步骤开始之后如果发现问题，及时返回到上一步骤中进行适当修改，整个过程迭代(overlap)前进；而且整个过程“循环往复”，即通过领域专家和用户的评价和实践应用发现问题，继而进行本体的再开发。

3　CSSCI 本体概念模型的构建

本体概念模型的构建也称为领域知识组织，包括概念的确立、概念层次结构的建立以及基于属性的概念间关系的设置等。领域的知识组织为知识描述准备了必要的知识框架和模型作为指导和参考，其准确性和合理性在很大程度上决定了本体构建的有效性。

3.1　CSSCI_Onto 的概念抽取

CSSCI_Onto 的概念模型来自 CSSCI 原始数据结构(关系模型)的转化，而具体的实例及其关联则来自来源数据的内容(关系元组)，这是 CSSCI_Onto 构建的基本思路。因此可以遵循指定的构建方法，从 CSSCI 元数据结构中提取本体的主要学术概念。

目前建立领域本体概念模型的方法主要有 4 种[7]：①自顶向下(Top-Down)方法。即先鉴别出领域中所有综合抽象概念，然后逐步细化为具体概念。②自底向上(Bottom-Up)方法。即先确定领域中的所有具体概念，然后通过二义性处理和归纳、聚类等处理，泛化(Generalize)形成综合性的抽象概念。③混合(Hybrid)方法。即先定义顶层的综合性抽象概念和底层的特殊性具体概念，然后分别细化和泛化，逐渐关联到同一中间层概念。④由里而外(Inside-Out)或核心扩展(Middle-Out)方法。这是一种结合①和

②的混合扩展策略，但起点是中间层概念。即首先确定领域中的核心概念，然后扩展出其他同层、上层和下层概念。具体采用何种方法主要依赖于领域知识自身的特点（数据环境）以及本体开发人员对领域知识结构的理解。

在CSSCI_Onto构建中，没有现成可利用的上层本体。现有资源以关系型数据库的形式存在，具有明确的元数据结构，主要概念作为关系的主关键词或外部关键词。根据领域数据的这一特点，笔者认为可采用核心扩展的方法来构建概念模型。考察CSSCI关系模型，在领域专家的帮助下，经过慎重的识别、分析和统计，确定主题、文献、期刊、学科、作者、单位、时间7个概念作为CSSCI学术资源领域的主要概念，将其中主题定为核心概念。主要概念是CSSCI_Onto中的描述重点，其本身具有重要的实际应用意义，领域中的其他知识都是用于描述这些概念的。主要概念具有无二义性、互不相交的特点，结合辅助概念如基金、项目等，能够覆盖整个CSSCI领域的专业知识。

3.2 CSSCI_Onto概念层次结构的建立

明确了主要概念之后，即可对CSSCI_Onto雏形进行扩展，其方向主要有两个。

1）同级扩展

从CSSCI中抽取辅助概念，完成对整个领域的知识覆盖。对来源数据进行分析、总结和抽取，结合本体开发目的，可以获得地区、基金、项目、类型、被引出版社5个辅助概念。辅助概念一方面可以作为提供知识检索服务的辅助元素，例如检索某年度国家社科基金发表的所有论文等；另一方面，也是引文分析的重要对象，如统计某地区各年度的发文引文情况，可以分析出该地区学术活动的发展变化等。

关于CSSCI_Onto概念层次结构中概念关系的选择和组织问题，可以采用两种方案：①使用部分-整体关系（part-of）组织概念层次结构，即认为层次结构中的下位类是上位类的部分。这是一种面向结构的知识组织方式。②使用子-父类关系（kind-of）组织概念层次结构，即认为层次结构中下位类是上位类的子类。这是一种面向对象的知识组织方式。本文选择了第二种方案，理由是：①面向对象的知识组织方式以其封装性、继承性等特点更符合人们的思维方式，而且其稳定性和可扩展性已经在软件工程领域得到了实践的检验。②概念继承思想避免了冗余类的产生。例如在来源文献和引用文献类中，都存在年份类，两者之间不存在明显的对象差异（属性一致），同时将主要概念和辅助概念作为层次结构的最顶层概念。

2）向下扩展

根据定义好的抽象父类，逐步细化说明其下位类。概念细化实际上是一个分类过程，不同类型概念可以根据具体的数据情况（从来源和引用两个角度细化）、实际应用的需要（如根据颗粒度的不同，可将来源期刊细分为种刊和单刊，分别探讨种刊和单刊相互之间的关联；从单位中派生出机构和部门等）以及层次结构可扩展的需要等进行细化。

在概念层次结构的构建过程中，需要把握两个方面的问题：①顶层概念的领域覆盖性和无二义性，防止“知识孤岛”和多重继承现象的出现；②概念层次结构可用性和精确性之间的平衡。领域本体的目的是完整清晰地描述领域知识框架，达到知识重用和共享。然而，过于详尽的层次结构往往带来不可理解性，而且完整的程度也并不像想象中容易把握。笔者认为，对于实际应用而言，概念层次结构应该尽可能简明，易于本体工程师理解，便于概念属性定义。基于CSSCI学术概念的抽取和核心扩展，本文建立了CSSCI_Onto的3层概念层次结构。图2显

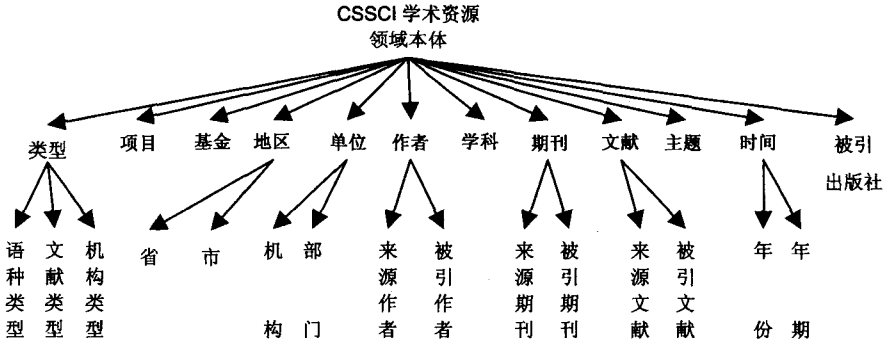

图2 CSSCI_Onto的2层概念层次结构

示了其中的 2 层结构。

3.3 CSSCI_Onto 概念属性的设置

如果说概念层次结构是本体骨架，那么其血肉就是以属性形式存在的概念间关系（或约束）。CSSCI_Onto 中的概念属性大致分为两类：

（1）数值属性（Datatype Property）。用于描述概念自身状态和结构等信息，如主题名称、文献编号等。

（2）对象属性（Object Property）。以某一概念的对象作为属性值，用于描述对象（实例）之间的关系，其取值随着对象间关系的变化或环境状态的影响而发生改变。例如主题概念中的“来源文献”属性就是一个典型的对象属性，揭示了主题与来源文献概念之间的语义关联。对象属性根据属性值来源又可以分为：①同类对象属性。如主题概念中的交叉主题属性等，用以描述同类概念间的关联。学科地图、知识地图等的绘制，都是对同类概念间关联的语义描述。Astrova[8] 在从关系型数据库中进行本体学习的研究中，也提出可通过对关系数据库元组（同类对象）的分析，得到概念间的“继承”关系。②异类对象属性。以其他类型对象作为值的属性，例如主题概念中的来源种刊、学科、来源作者等属性就是分别以期刊对象、学科对象和作者对象等作为属性值，揭示的是主题对象和期刊、学科、作者等的相互关系。

如果把处于核心地位的概念称为主体，作为属性的概念称为客体，主客体之间可以通过属性建立关系。这种关系有时是二元的，可以通过三元组 < 主体，属性，客体 > 描述。例如 < 主题，下位主题，主题 > 就是通过三元组描述了主题与主题之间的下位关联。然而这种属性关系在更多的时候是多元（N-ary）的。例如主题概念的“交叉主题”属性，其属性值不仅需要指出具体的客体，还需要说明主客体的关联度。客体和主客体关联度一起构成了属性值。此时就需要使用多元组来描述主客体关系：〈主体，属性，客体，主客体关联度〉（如〈语义网，交叉主题，RDF，0.12975〉等）。这类属性也称为复合属性。在使用 OWL 进行描述时，需要引入属性对象进行 N-ary 关系分解[9]。

根据 CSSCI 原始数据描述的概念性质，并结合 CSSCI_Onto 的实际应用的需要，在领域专家的参与下，可以分别定义概念层次结构中所有类的属性，并对属性性质，如取值范围、允许取值以及属性的基数等信息进行说明，以实现对概念和关系的明确定义。概念层次结构的确定和概念属性的设置标志着 CSSCI_Onto 知识框架定义的完成。

4 基于 CSSCI_Onto 概念模型的语义标引

CSSCI_Onto 概念模型构建完成后，即可用于指导原始数据的语义标注。CSSCI_Onto 语义标注过程，就是根据概念模型定义的知识框架在原始数据中抽取实例，并为实例的属性设置具体值的过程。鉴于 CSSCI 来源数据以及本体概念模型的特点，本文采用信息抽取的方法自动、批量地从数据库中获取实例，并针对不同类型的属性采用不同的计算方式获取属性值，实现 CSSCI_Onto 的语义标注。

4.1 CSSCI_Onto 概念属性值的获取方式

考察 CSSCI_Onto 中概念的所有属性，发现可采用 3 种方式来完成各类属性的取值。

（1）基于关系模型直接抽取。某些属性直接来自关系的字段名，包括所有数值属性和部分的异类对象属性。针对这类属性，可以直接从关系元组中抽取实例之间的依赖关系（字段值）作为实例的属性值。如来源文献的“所属学科”、“所属期刊”等属性。

（2）基于直接统计并适当辅以 TF-IDF 算法。对于部分异类对象属性需要统计属性实例相对于中心概念实例的关联次数，如来源作者概念的“年份发文量”等。为了使关联属性值更具意义，有时甚至需要基于统计值计算各年总值或实例间关联度。如主题概念的“来源作者”属性，除了统计主题与来源作者之间的关联次数外，还可以基于 TF-IDF 算法计算主题与来源作者的关联度，以便更合理地揭示主题与来源作者的关系。

（3）基于标准加权方法。在讨论同类型概念间关系时，从不同角度出发可以获得概念间的不同关联，例如作者概念之间可能存在共现主题、合作、应用、同被引以及同部门等多种关联。可以对各种关系设置不同权重，进而计算同类概念实例的综合关联度。

4.2 基于标准加权的语义关联解析

对于 CSSCI_Onto 中的主要概念，同类型概念的不同实例之间均存在一定的内容或其他形式的关联，例如“交叉主题”、“相似文献”、“关联期刊”、“关

联作者"等。而且各关联实例对中心实例的关联程度也存在差异,有必要对这种同类概念间的内在关联进行深刻揭示。

本文采用标准加权的方法来解析同类实例之间的交叉语义关联。基本思想是:同类实例之间由于不同的关联依据可能存在多种类型的语义关联,因此可以先基于多种标准分别获得实例间关联,再根据不同标准权重各异(对综合关联的贡献程度不同),计算加权平均值,作为实例间的综合关联度。本文仅以来源作者间综合关联度计算为例,说明基于标准加权算法实现同类实例间语义关联解析的方法和过程。

(1)设置同类概念实例间关联的依据标准。考察来源数据,结合领域专家意见,笔者认为可以从来源作者间的主题共现、合作、引用、同被引以及同部门5个角度,挖掘来源作者间的综合关系。

(2)根据不同关联标准分别建立描述作者实例的二元关联矩阵。即每一作者实例分别用一个属性向量来表示,向量值为作者实例与描述属性之间的关联次数。

(3)对于主题共现和同被引矩阵,可以采用机率模式[10]计算来源作者间交叉关联度。对于合作、引用和同部门关联矩阵,则可以采用TF-IDF算法计算交叉关联度。

(4)鉴于各个关联标准对综合关联度的贡献程度不同,分别设置权重。

(5)将基于5种标准获得的关联度值归一化后,根据权重计算归一值的加权平均值作为作者实例间的综合关联度。对每一来源作者,取综合关联度≥0.01且最高的25名作者作为其关联作者。

图3为基于CSSCI(2000~2006年)来源数据获得的,来自"南京大学信息管理系"的15位来源作者之间的关联云图。

4.3 基于TF-IDF的概念属性设置

在很多情况下,仅使用频次并不能完全反映对象之间的关联程度。例如主题与来源作者之间的关联,并不是说关联频次越高,作者对主题就越重要,这还与作者涉及主题的总个数有关。因此本文引入TF-IDF指数作为关联系数来描述实例之间的关联程度。

基于TF-IDF算法设置实例属性值的一般过程是:①基于CSSCI来源数据统计中心概念实例和关联概念实例的关联次数。②建立中心概念实例×关联概念实例的二元关联矩阵,以关联次数作为矩阵值。③建立适用于描述中心概念和关联概念关系的TF-IDF公式。如公式(1)即为用于计算来源种刊(sq_j)相对于主题(t_i)关联度的TF-IDF算法,式中N为主题总个数,$n(sd_j)$表示与sq_j有关系的主题个数。④将二元关联矩阵中的数值代入TF-IDF公式中,计算中心概念实例和关联概念实例之间的关联度。⑤使用三元组(关联概念实例,关联次数,关联

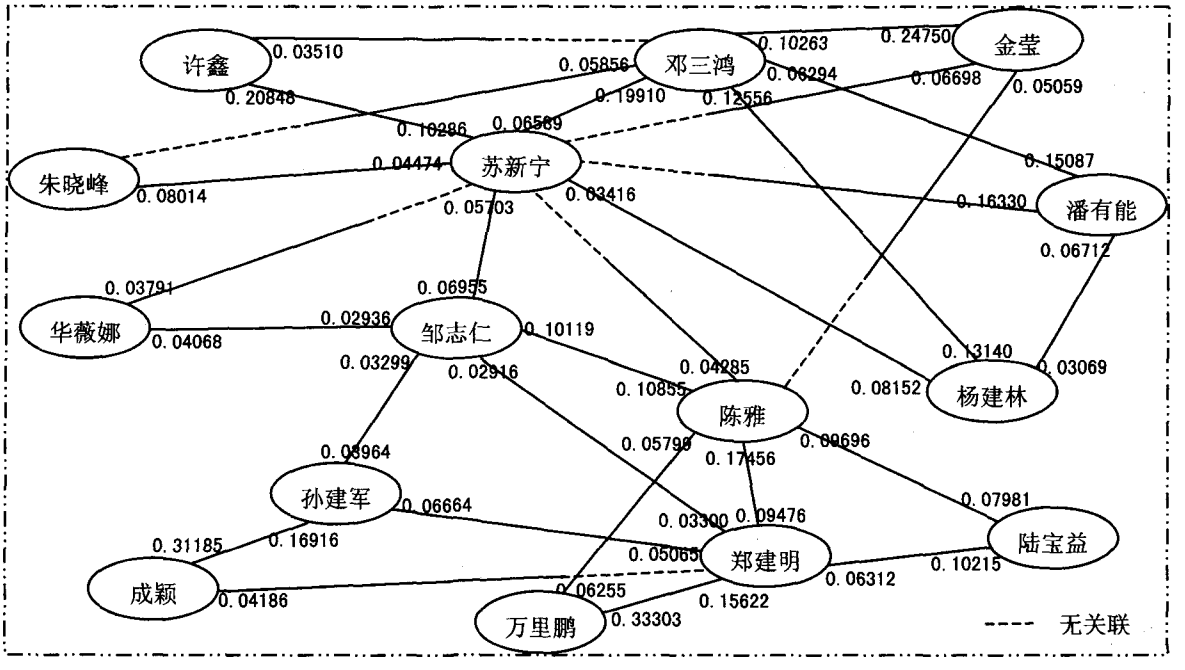

图3 "南京大学信息管理系"15位来源作者间的关联云图

度)描述中心概念实例,作为其关联属性值。

$$f(t_i, sq_j) = TF \times IDF = \frac{tf(t_i, sq_j)}{tf(t_i)} \times \frac{\log_{10} \frac{N}{n(sq_j)}}{\log_{10} N} \quad (1)$$

以主题概念的来源种刊属性设置为例。首先统计主题在来源种刊中的出现次数。一般认为某一主题在刊载文献中出现,即认为它在种刊中出现一次,而不考虑其在来源文献中具体出现的次数。然后建立主题×来源种刊的关联矩阵,矩阵值为主题在种刊中出现的总次数。将矩阵中的数值代入公式(1)中,计算来源种刊相对于主题的关联度。表 1 展示了部分主题实例的来源种刊属性值。

4.4 基于本体的 CSSCI 学术资源集成

本体实例通过对象属性和其他类型概念建立联系,而关系属性又以其他概念作为其对象属性,于是学术概念之间以属性形式串联成一个巨大的网状结构,形成知识网络。根据本体各类概念的基本结构可以将 CSSCI(2000 ~ 2006 年)中的所有知识实例化,以更合理的面向对象的方式组织学术资源。图 4 为基于本体的"语义网"主题的知识地图,图中清晰地显示了各种类型的概念实例之间的关联,以此为导航可以获得与主题"语义网"相关的学术资源信息。类似地,将 CSSCI(2000 ~ 2006 年)中的所有学术资源实例集成在一起,就形成了一个巨大的基于本体的学术资源网络,在其中能够方便快捷地获取任一学术资源实例的相关信息,以支持知识检索和引文分析服务。

建立概念模型以获得概念库,对原始数据实行语义标引可获得实例库,由概念库和实例库共同构成本体库。基于 2000 ~ 2006 年的来源数据,历时近 10 个月构建成的实验性 CSSCI_Onto 是一个 3 层概念层次结构,共包含 39 个本体类、336 个属性,其中数值属性 53 个,对象关系属性 283 个。此外,为了实现蕴涵推理,建立了含有 18 条推理规则的知识库。在 7 年的数据中,共标注出了552 566个主题,504 021篇来源文献,558 种期刊,25 个学科,215 942个来源作者,89 884个部门等实例。

5 CSSCI_Onto 的评价模型

CSSCI_Onto 主要面向知识服务,以转变现有的

表 1 主题的来种刊属性值 top5(局部)

主题 X	来源期刊 Q	频次	关联度 R	主题 X	来源期刊 Q	频次	关联度 R
…	…	…	…	…	…	…	…
语义网	图书情报工作	9	0.05405	语义检索	现代图书情报技术	6	0.10668
	现代图书情报技术	7	0.04641		情报学报	5	0.09545
	情报科学	7	0.04210		情报理论与实践	3	0.05689
	情报杂志	7	0.03889		图书馆杂志	2	0.03279
	情报理论与实践	5	0.03535		情报科学	2	0.03226
知识组织	图书情报工作	28	0.06360	数据挖掘	情报杂志	67	0.06101
	情报理论与实践	22	0.05883		情报科学	33	0.03253
	情报杂志	20	0.04203		情报学报	23	0.02683
	情报资料工作	14	0.03798		现代图书情报技术	18	0.01956
	中国图书馆学报	14	0.03648		中国统计	19	0.01864
本体构建	情报学报	2	0.12000	本体	现代图书情报技术	18	0.04924
	现代图书情报技术	2	0.11176		情报杂志	21	0.04814
	中国图书馆学报	1	0.05807		图书情报工作	17	0.04212
	图书情报工作	1	0.05062		情报学报	14	0.04112
	情报杂志	1	0.04683		情报理论与实践	13	0.03793
…	…	…	…	…	…	…	…

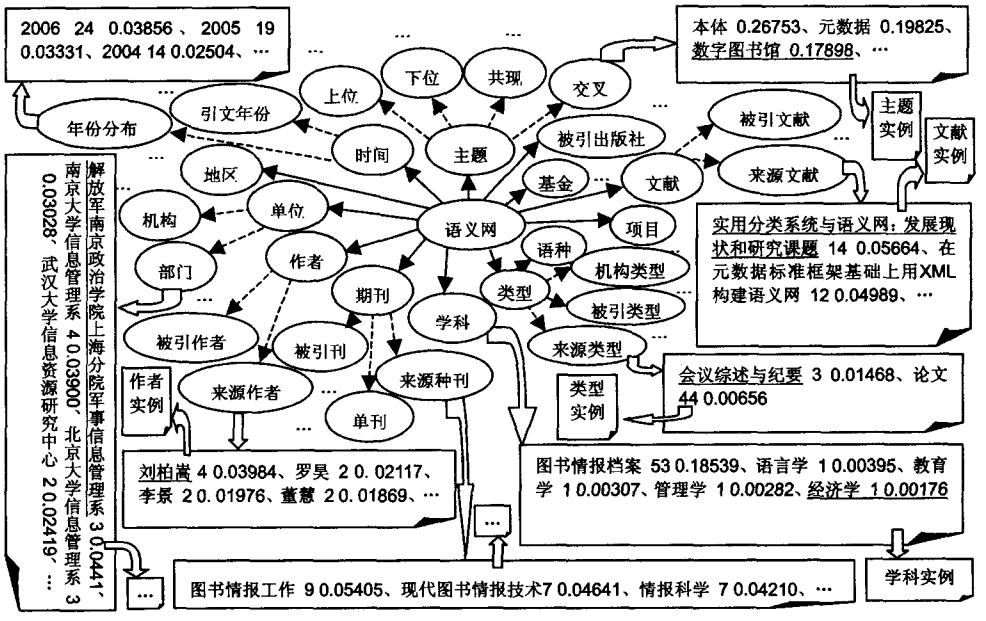

图 4 基于 CSSCI(2000～2006)的"语义网"主题知识地图

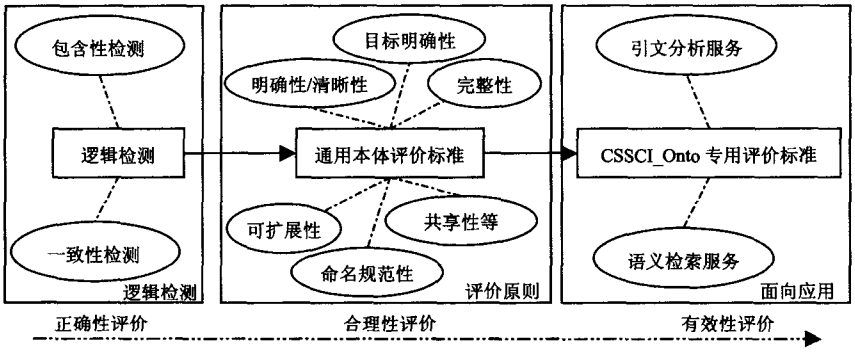

图 5 CSSCI_Onto 的评价模型

服务模式、改善服务质量、提高服务水平为主要目的。因此对该本体的评价，主要着眼于其实际应用价值。本文提出了专门用于 CSSCI_Onto 的评价方法，其基本流程如图 5 所示。

笔者认为应该从本体的正确性、合理性和有效性 3 个方面来评价 CSSCI_Onto。具体地说，①正确性评价是保证本体概念模型不存在逻辑错误和逻辑冲突，包括描述逻辑的一致性检测和包含性检测；②合理性评价是指由领域专家和本体工程师等根据通用的评价原则和标准，如明确性、完整性、共享性、可扩展性等，对建立的本体进行适用性评价，检查其对领域知识组织的合理程度；③有效性评价则是针对 CSSCI_Onto 在实际应用中的有效性和便利性，由用户来检测本体的可用性。

上述本体评价分 3 步进行，在本体构建的不同阶段实施。正确性评价在本体概念模型建立之后进行，合理性评价则在本体语义标注完成后进行，而面向应用的有效性评价则在本体的实际应用中进行。在每次评价后，都可以根据反馈来修正本体。该模型适用于面向用户服务的应用型领域本体的评价。

6 基于 CSSCI_Onto 的应用研究

本体构建是本体应用的基础，而本体应用则是本体构建的最终目的和完善途径。构建 CSSCI_Onto 是为了改善现有的 CSSCI 信息服务的质量，提高服务水平。

6.1 CSSCI 信息服务现状

当前 CSSCI 主要应用于信息检索和引文分析

服务。

(1)提供信息检索服务。渠道包括光盘检索和Web检索,基本功能是根据检索元数据项,实现来源文献检索和被引文献检索。

(2)提供引文分析服务是CSSCI建立的初衷。基于现有的数据结构,当前主要进行两个方面的引文分析研究:①基于数理统计的引文分析。即通过对来源数据的简单和回溯统计发现学科规律,评价来源期刊、作者、机构、地区等的学术水平和学术影响,探讨哲学社会科学的学科热点和发展趋势,如苏新宁[12]教授等的研究成果。②基于数据仓库和数据挖掘的引文分析。即构建引文分析数据仓库,实现OLAP,同时采用关联规则和主题聚类等算法挖掘同类对象间的潜在关联,构建知识地图,如邓三鸿[13]等的研究。

6.2 基于CSSCI_On的知识检索服务平台

现有的CSSCI信息检索系统的检索对象单一,功能简单,仅能够提供有限的服务,这与其直线型的后台数据组织方式有重要关系。笔者通过建立基于本体的学术资源网络模型,以CSSCI_Onto面向对象的结构实现知识的组织。在这种知识结构下,可以建立面向多学术资源对象、实现语义推荐检索和基于规则推理检索的知识检索服务平台(简称KRSP_CSSCI_Onto),将CSSCI信息检索服务提升到知识服务的层次。KRSP_CSSCI_Onto的主要功能包括:

(1)支持多学术资源对象的检索。CSSCI_Onto将主题、作者、文献、期刊、学科、机构等作为主要对象,为各种对象设置属性,因此可以通过各种对象的局部属性特征,检索指定的对象实例,并获得该实例的所有相关信息。

(2)能够实现基于语义关联的学术知识推荐。①从检索表达式角度来看,很多用户在检索前期往往对检索意图不明确或希望能够实现查询扩展,这就要求检索系统能够针对用户的查询词进行关联扩展。CSSCI_Onto中明确描述了同类概念实例之间的交叉关联,能够从中得到所有与查询词关联的对象,从而实现查询式的语义推荐。例如用户使用作者检索时,输入"苏新宁",此时系统将会向用户推荐本体中存在的与"苏新宁"相关的其他作者,如"许鑫"、"邱均平"、"吴鹏"等,供用户选择。②从检索结果角度来看,由于查询式构造难、用户对专业知识理解不充分等原因,用户往往并不满足于获得的检索结果,而希望能够得到更多的相关信息。同样的,基于本体中的同类学术资源关联知识,能够对检索结果进行语义扩展。例如,用户通过检索获得来源文献后,试图获得更多的相似文献以供参考,此时系统可以根据本体中定义的来源文献间关联,向该用户推荐当前结果的相似文献、引证文献、被引证文献、同引证文献、同被引文献等知识。

(3)能够实现学术资源间关系的查询和推理。在CSSCI_Onto中,学术资源之间以属性建立关联,因此实例间关系的查询实际上就是属性名称及多元属性值的显示过程。KRSP_CSSCI_Onto将学术资源之间的关系分为两个部分:一是本体中定义的关系,二是基于规则的推理关系。例如查询来源作者"杨建林"(主体)和"邓三鸿"(客体)之间的关系,首先在本体定义的关系中,可以查询两者之间存在交叉关联关系;随后搜索规则库,查询相关规则,并根据规则检索出两者之间的推理关系,如合作、被引、共引等,并可以进一步获得推理的路径。

6.3 基于CSSCI_Onto的引文分析研究

基于直接统计的引文分析无法揭示学术资源之间深层次的潜在关联,例如作者间关联、学科间关联等。数据挖掘只是一种技术手段,在现有CSSCI数据结构中直接挖掘,只能得到一些片面的结论,例如基于主题共现挖掘学科间关系等,而对于学科或作者的研究热点、研究趋势等分析,便显得无能为力。因此,本文提出了基于CSSCI_Onto的引文分析方法,在包含现有分析的基础上,试图提高CSSCI引文分析的准确性和全面性,得到一些可供参考的结论。基于CSSCI_Onto的引文分析主要包括以下几个方面:

(1)基于学术资源统计属性的学术影响力分析。在构建CSSCI_Onto过程中,有意识地为概念设置了统计属性,如基金类别概念的基金类别发文量、年度发文量等属性,便于在进行学术影响力分析时直接获得学术资源的统计数据。此外,本体概念之间的继承关系使得能够实现钻取(Drill)操作,例如通过地区(省)发文量分析地区的学术影响力。如果想进一步了解指定"省"的所有"市"的发文情况,则可以根据CSSCI_Onto中定义的省实例和市实例之间的所属关系,实现下钻(Drill down)操作。

(2)基于本体的学术资源关联分析。一方面,在CSSCI_Onto中存储了同类学术资源的相互关系,根

据这种关系可以直接获取任一实例的关联对象，例如与“图书情报档案学”最相关的学科依次是新闻出版广播学、管理学、教育学、法学和经济学，甚至可以列出其关联程度。另一方面，在CSSCI_Onto中使用数据挖掘技术(包括聚类、多维尺度分析等)，可以获得更精确、更完善的结论。基本步骤是：从CSSCI_Onto中取得同类学术资源之间的关联度，建立学术资源×学术资源的二元关联矩阵，以两者的平均关联度作为矩阵值，对关联矩阵进行聚类和多维尺度分析，得到分析结论。例如，图6揭示了“南京大学信息管理系”28位现任教师的多维关联情况，以二维平面点距离揭示教师间的两两关系。结合图7所示的聚类情况，可以将上述作者分为比较符合实际情况的六大类：一类作者以信息资源管理、文献分析为主要研究方向；二类作者以多媒体信息的处理和检索、自动化技术在数字图书馆中的应用等作为主要研究方向；三类则以竞争情报、市场调研为主要研究内容；四类作者之间的关系相对松散，主要以档案学、编辑出版为主要研究主题；五类可以分为两小类，一小类包括施云、岳泉、谭华军，以信息传播、文献评论等为主要研究主题，存在较多合作，另一小类包括徐雁和叶继元，以文献阅读、评论和期刊评价等编辑学内容为主；六类作者以信息分析、信息安全、网络版权、元数据等数字图书馆理论、情报学基础理论以及产业分析、电子商务等为主要研究内容。

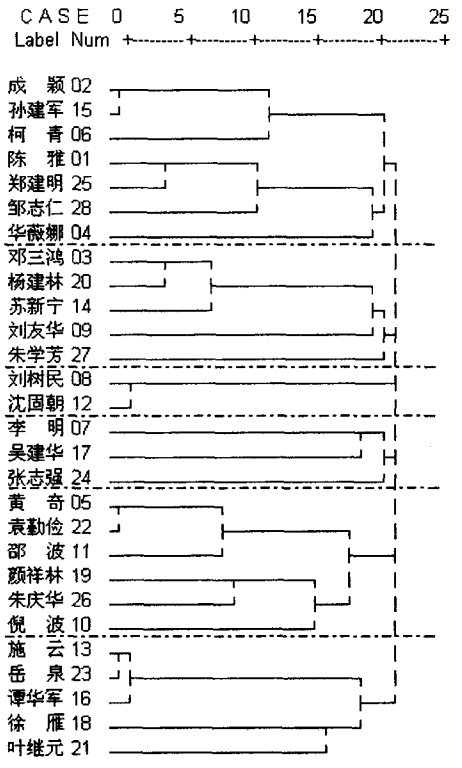

图7 基于综合关联度的学者聚类结果

(3)基于本体的学科热点分析。即根据本体中存储的主题间关联对主题进行聚类和多维尺度分析，从学科或学科中具有影响力的学者所涉及的热点主题(出现频次高)角度总结出当前学科的研究热

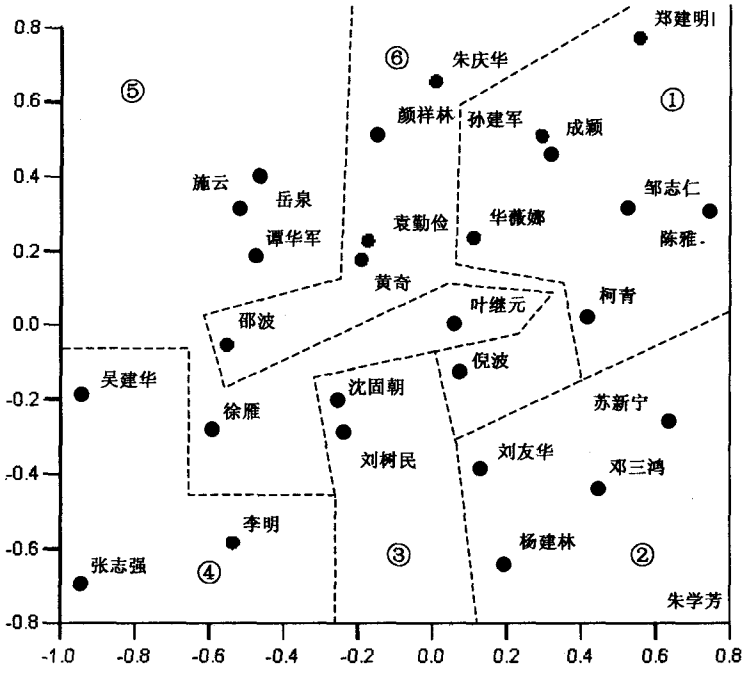

图6 “南京大学信息管理系”28位学者间关系的多维尺度分析结果

点。此外,基于主题和来源种刊之间的关系,可以为论文选择合适的投稿期刊。

(4)基于本体的学科或学者的研究热点及趋势分析。根据 CSSCI_Onto 中的学科或学者与各年度主题的关系,选择学科或学者的热门主题,再基于主题间的关联度对热门主题进行聚类,从而可以获得各年度学科或学者的研究热点,绘制各个热点的年度发展趋势图,从中可以发现学科或学者的研究趋势。

7 结束语

本文在分析 CSSCI 数据和服务现状的基础上,提出通过本体面向对象的知识结构来组织 CSSCI 中的学术资源,以达到提高 CSSCI 信息检索和引文分析服务质量的目的。本文提出的基于本体构建 CSSCI 学术资源网络模型的解决方案是可行的,通过具体的实践——建立基于 CSSCI_Onto 的知识检索服务平台和实现基于本体的引文分析,证明了构建的本体具有实际应用价值。

在本体的具体构建过程中,也遇到了一些问题。本文构建的 CSSCI_Onto 概念模型基于原有的关系型数据库的元数据结构,在领域专家的协助下完成,并通过实际应用对该模型进行了循环改进,但是概念模型的构建需要考虑多方面因素,而且经专家评价和实验验证的机会都不多;本文基于 CSSCI(2000~2006 年)数据来建立学术资源知识网络,数据量相当大,虽然采用机器手段完成了资源的语义标引,但是整个过程是分步完成的,还没有建立一套完整的自动化步骤;对语义标注过程采用的 TF-IDF 算法,虽然采用归一化方法实现了全局范围内关联度的比较,具有一定合理性,但是由此获得的关联度会随主体实例总量的变化而发生改变,不利于实例库的扩展;CSSCI_Onto 在信息检索和引文分析中的应用还只是初步的,需要更多的实践来进一步验证和改进。这些问题都有待于笔者进一步研究和探讨。

参 考 文 献

[1] Tim Berners-Lee. Semantic Web Road map[OL]. [2008-10-30]. http://www.w3.org/DesignIssues/Semantic.html.

[2] Tenier S, Toussaint Y, Napoli A, et al. Instantiation of relations for semantic annotation[C]// Proceedings of 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 2006: 463-472.

[3] 廖述梅. 基于本体的语义标注原型评述[J]. 计算机工程与科学,2006,28(9):123-128.

[4] Welcome to protégé[OL]. [2008-10-30]. http://Protege.stanford.edu.

[5] Gruber T R. Towards principles for the design of ontologies used for knowledge sharing[J]. International Journal of Human Computer Studies ,1995 ,43 :907-928.

[6] Maedche A, Staab S. Ontology learning for the semantic Web[J]. IEEE Intelligent Systems, 2001, 16(2):72-79.

[7] Uschold M, King M. Towards a methodology for building ontologies[C]// Workshop on Basic Ontological Issues in Knowledge Sharing, Held in Conjunction with IJCA I-95, Montreal, Canada, 1995.

[8] Astrova. Reverse engineering of relational database to ontologies [C] // Proceedings of the ESWC 2004. Heidelberg: Springer-Verlag, 2004:327-341.

[9] Noy N, Rector A. Defining N-ary Relations on the Semantic Web[OL]. [2008-10-30]. http://www.w3.org/TR/2006/NOTE- swbp-n-aryRelations-20060412/.

[10] Weng S S, Tsai H J, Liu S C, et al. Ontology construction for information classification[J]. Expert Systems with Applications, 2006, 31(1): 1-12.

[11] Swartout B, Ramesh P, Knight K, et al. Toward distributed use of large-scale ontologies [C/OL] // Symposium on Ontological Engineering of AAAI. Stanford (California), Mars, 1996. http://ksi.cpsc.ucalgary.ca/KAW/KAW96/swartout/Banff_96_final_2.html.

[12] 苏新宁. 中国人文社会科学学术影响力报告(2000-2004)[M]. 北京:中国社会科学出版社,2007.

[13] 邓三鸿. 知识地图的构造和利用[D]. 南京:南京大学,2003.

(责任编辑 许增棋)