面向STKOS的概念映射及其系统实现

王 波
(杭州电子科技大学 浙江 杭州 310018)

摘 要:随着科技文献知识服务需求的不断提高以及信息的爆炸式增长,为了更加方便、快捷的对信息进行检索,本文提出了在科技知识组织体系(Science and Technology Knowledge Organizing System,简称STKOS)的基础上对概念进行映射分析的方法,并实现了其系统化。文中对多种标准的分类体系和各领域的文献词表进行异构分析,并对其进行了映射。实验结果表明本文的方法可以在多重标准下的分类体系和词表中方便快捷地对概念进行检索。

关键词:STKOS;分类体系;映射;词表;概念

0 引言

目前,网络信息资源以指数级速度增长,传统检索模式已不能满足人们的需求,因此,新的信息检索模式成为近年来研究的热点。从信息来看,所谓的情报就是知识或信息经传递并起作用的部分,即运用一定的形式,传递给特定用户并产生效用的知识或信息[1]。简单地说,情报就是从某种情况中所获得的有用的消息和报告。对于现今信息量大幅度增加,叙词表和分类法成为了支撑高效信息检索的基础知识体系。由于分类法和叙词表的种类众多及其各自的标准不同,因此对于这一知识体系,存在着一定的问题:叙词表和叙词表之间存在异构,分类法之间也存在异构,分类法和叙词表之间缺少有效的相互支撑。对于分类法和叙词表之间所存在的异构问题的研究,如OCLC研究机构通过采用同现映射和直接映射组合的方法,研究了《杜威十进制分类法》DDC和《美国国会图书馆标题表》[2](Library of Congress Subject Headings,简称LCSH)的兼容操作[3]以及Northwestern大学的研究项目LCSH和《医学主题词表》(Medical Subject Headings,简称MeSH)的映射[4]。本文基于科技知识组织体系对概念进行映射分析,能够在一个统一的标准中对概念进行检索,这样可以更好地提高对概念的检索效率。

1 面向分类法和叙词表的概念映射

1.1 分类法和叙词表的概述

本文所用到的分类法有国际专利分类表[5](International Patent Classification,简称IPC)、国际标准分类法(International Classification for Standards,简称ICS)、中国标准分类法(Chinese Classification for Standards,简称CCS)、国际十进制分类法[6](Universal Decimal Classification,简称UDC)和杜威十进制分类法[6-7](Dewey Decimal Classification,简称DDC)。下面是对它们和叙词表的简单介绍及特点分析。

IPC是使各国专利文献获得统一分类的一种工具。IPC是从专利技术领域进行分类的,且结构上是按等级进行分类的,通过字母和数字逐级编号得到分组类号,分组类号后的圆点数表示各分组的从属关系。ICS是由国际标准化组织编制的标准文献分类法。它按标准文献主题内容所属学科、专业进行分类的,采用的是层累制分类法,由三级类目构成,每个级别上都是有编号的,并通过小圆点进行隔离,且在二级和三级中存在着一些注释对类目进行详细说明。CCS简称中标分类,它的分类标准是以专业划分为主,适当结合科学进行分类。CCS所采用的分类法是二级分类,一级分类和二级分类分别以字母和数字作为编号。UDC是用单纯阿拉伯数字作为标记符号的,它采用的分类法是三级分类。它用个位数(0~9)标记一级类,十位数(00~99)标记二级类,百位数(000~999)标记三级类。DDC分为10大类。它采用阿拉伯数字作标记符号,并采用小数制(即十进制)的层累标记制,以三位数(000~999)形成前三级的等级结构。在三位数中,凡带"0"的号码均表示总论性类目:后二位为"0"的号码表示一级类(大类),末一位为"0"的号码表示二

作者简介:王 波(1986-),男,硕士生,主要研究方向:智能信息检索。E-mail: lovewbo1230@163.com

级类，凡末尾不带“0”的三位数号码均属三级类。

叙词表是一种概括某一学科领域，以规范化的、受控的、动态性的叙词（主题词）为基本成分，以参照系统显示词间关系，用于标引、存储和检索文献的词典。叙词表的标准中没有明确的等级分类，但叙词之间存在这多种关系：等同关系、等级关系和相关关系。在叙词表中，对于某些叙词都有注释对其进行详细说明。

1.2 对分类法和叙词表的特点分析

根据对 IPC、ICS、CCS、UDC、DDC 和叙词表的描述，它们之间存在异构性，同时，它们之间也有着一定的共性。首先，IPC、ICS、CCS、UDC 和 DDC 都是按等级进行分类的，且叙词表也有等级分类。其次，ICS、CCS 和叙词表都是以专业和学科为主进行分类的。

对以上几种标准进行分析，可以得出几种分类法和叙词表的异构情况。其中，IPC 和叙词表必须单独分类，其他几种分类法有较多的共同点。故对它们各自的类目进行分析，可总结出 7 种不同标准的结点：分别是部结点（IPC 的 8 个大部）、目录结点（IPC 的类、主组，ICS 的一级分类，CCS 的一级分类，DDC 的一级、二级分类，UDC 的一级、二级分类，CLC 的一级、二级分类）、分类结点（IPC 的小类，ICS 的二级分类）、专利叶结点（IPC 的分组及其后面带圆点的信息）、学科叶结点（CCS 的二级分类，DDC 的三级分类，UDC 的三级分类，CLC 的三级分类）、ICS 叶结点（ICS 的三级分类）和叙词结点（叙词）。对于结点间的关系，即所有概念间的关系，有等价、继承、相关、“是一种特征”和“是一部分”5 种。其中继承关系可以细分出“是一种”、“应用于”及“具备特征的”三类关系；“是一种特征”可细分出“是一种属性”和“是一种表现”两类关系。

1.3 面向分类法和叙词表的映射

根据对几种分类法和叙词表的分析，总结得出 7 种结点和 8 种关系，本文将得到的 7 种结点进行两两结合，分析它们是否存在某种关系。在这 7 种结点中，并不是所有的结点之间都存在关系，故将不存在关系的结点组合去除，可以得出以下结点组合存在某一种或多种关系，分别是：IPC 的部和类的关系为“是一种”；IPC 的类和小类的关系为“应用于”；IPC 的小类和主组的关系为“应用于”和“是一部分”；IPC 的主组与分组的关系为“是一种”和“是一种属性”；IPC 的分组间的关系为“具备特征的”；ICS 的一级与二级分类的关系为“是一种”和“应用于”；ICS 的二级与三级的关系为“是一种”和“应用于”；CCS 的一级与二级的关系为“是一种”和“应用于”；DDC 的一级与二级的关系为“是一种”；DDC 的二级与三级的关系为“是一种”和“是一部分”；UDC 的一级与二级、二级与三级以及三级与三级的关系都为“应用于”；CLC 的一级与二级、二级与三级的关系为“是一种”；主题词表中除了同级关系以外，还有“是一种”、“是一种表现”及相关 3 种关系。

2 面向科技知识组织体系的概念映射系统的实现

本文根据总结得出的各种分类法和叙词表之间的映射关系，通过开发平台 Eclipse 及 GMF 来对其进行系统上的实现。其系统平台的结构如图 1 所示。

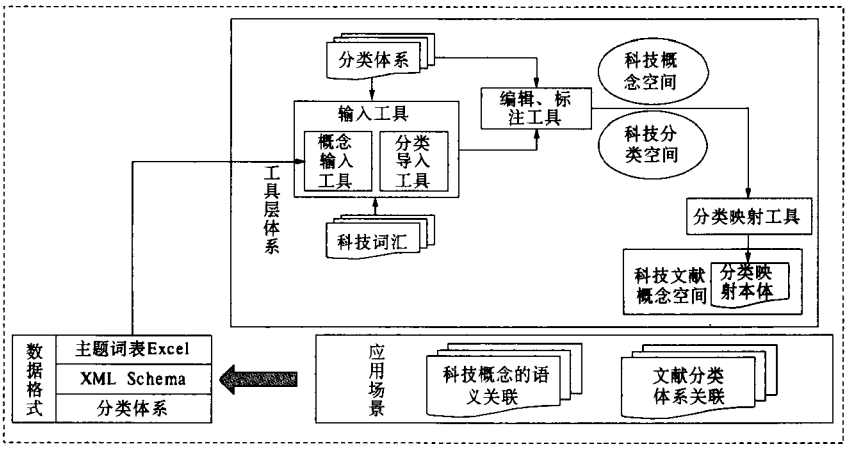

图 1 系统的结构图

Eclipse 是著名的跨平台的自由集成开发环境（IDE），主要用于 Java 语言开发。Eclipse 的基础是富客户机平台（Rich Client Platform，RCP）。GMF 提供了图形化编辑器的开发环境和运行时框架。GMF 框架主要包括两个部分：运行时框架，图形化开发环境。图形开发环境依赖底层的运行时框架。在 GMF 的图形化开发环境中，GMF 编辑器的主要功能是用来定义一组用来生成代码框架的模型，主

要包括:graphical model, tool model 和 map model。

其应用场景为科技概念的语义关联和文献分类体系关联。输入数据即各种分类法和叙词表的概念,其格式为 Excel。通过编码实现对数据的导入功能,期间进行二次标注。

2.1　系统框架

本文中,系统的实现分为 3 大模块:功能层、模型层和数据层。其框架图如图 2 所示。

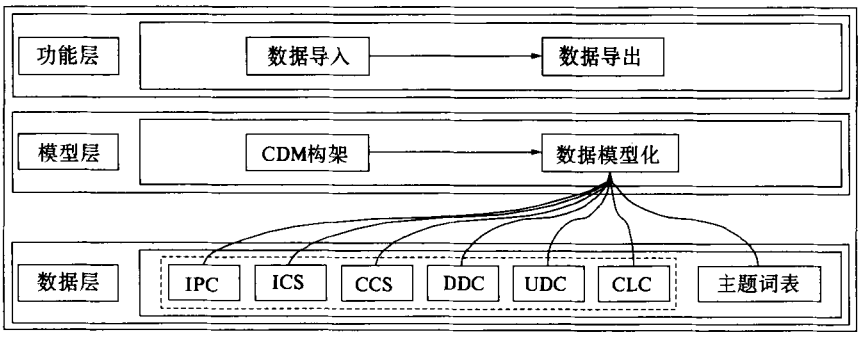

图 2　系统的框架图

其中功能层有两个方面:数据导入和数据导出。模型层的实现是先由上文分析得到的公共数据模型在 Eclipse 中用 Ecore 模型描述,驱动生成 CDM 构架,接着将所需要的数据导入,实现数据的模型化。数据层部分包含的是该研究的数据,其中包括分类体系中的 IPC、ICS、CCS、DDC、UDC 和 CLC 以及主题词表。

2.2　系统的技术体系

在本文的研究中,由于系统的实现是需要集多种功能于一体的,故在系统开发的过程中涉及多种技术。系统的技术体系如图 3 所示。该研究中,所用到的技术有 Eclipse 的插件 GMF 和 RCP 以及其他插件。通过将原始数据经过编码分析,使数据标准化,然后经由 GMF 来实现数据的可视化。

3　系统实现的认证结果

本文中,用于实例认证的数据主要是主题词表、IPC 和 UDC,主要选取了石油方面的一些概念。由于概念数据的信息量过大,故只截取其中的一部分来引用。主题词表的原始数据如表 1 所示。IPC 的原始数据片段[7]如图 4 所示。UDC 的原始数据片段[8]如图 5 所示。通过系统实现后得出的结果分别如图 6、图 7 和图 8 所示。

表 1　石油主题词表

中文	英文	属性	值(中文)
……	……	……	……
安全接头	SAFETY JOINT	C	钻柱
安全卡子	SAFETY CLAMP	S	安全设备
安全卡子	SAFETY CLAMP	C	钻井设备
安全系数	SAFETY FACTOR	S	安全
……	……	……	……
安山岩	ANDESITE	S	火山岩
安山岩	ANDESITE	C	中性岩
安装	MOUNTING	D	吊装
……	……	……	……

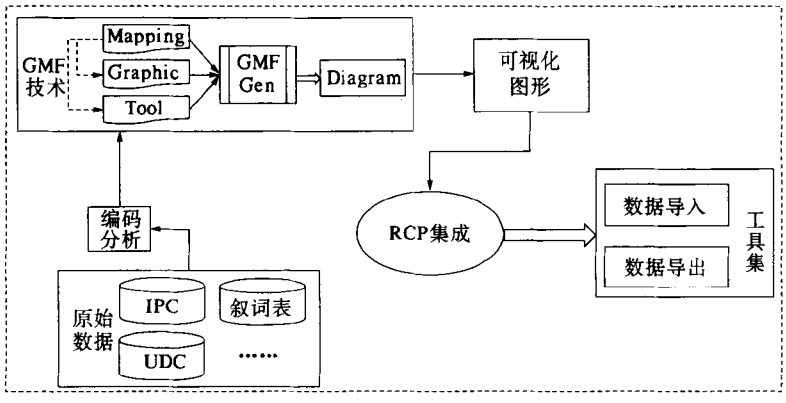

图 3　系统的技术体系图

C01 无机化学
C01B 非金属元素；其化合物
C01C 氨；氰；其化合物
C01D 碱金属，即锂、钠、钾、铷、铯或钫的化合物
C01F 全属铍、镁、铝、钙、锶、钡、镭、钍的化合物，或稀土金属的化合物
C01G 含有未列入C01D或C01F小类之金属的化合物
C02 水、废水、污水或污泥的处理 (沉淀池，过滤，如砂滤池或筛选装置入B01D)
C02F 水、废水、污水或污泥的处理
C03 玻璃；矿棉或渣棉
C03B 制造、成型及辅助工艺
C03C 玻璃、釉或搪瓷釉的化学成分；玻璃的表面处理；由玻璃、矿物或矿渣制成的纤维或细丝的表
C04 水泥；混凝土；人造石；陶瓷；耐火材料
C04B 石灰；氧化镁；矿渣；水泥；其组合物，例如砂浆、混凝土或类似的建筑材料；人造石；陶瓷

图 4　IPC 数据片段

5 数学、自然科学
502/504 环境科学、自然资源保存、环境威胁与保护
502 环境及其保护
502.3/.7 环境组成部分
504 环境威胁
51 数学
510 基础数学、数学原理
510.2 数理逻辑一般问题、基础数学
510.3 集合论
510.6 数理逻辑
511 数论
512 代数
514 几何
514.7 微分几何、几何代数与分析法
515.1 拓扑学
517 数学分析
517.9 微分方程、积分方程、其他函数方程、有限差分、变分法、

图 5　UDC 数据片段

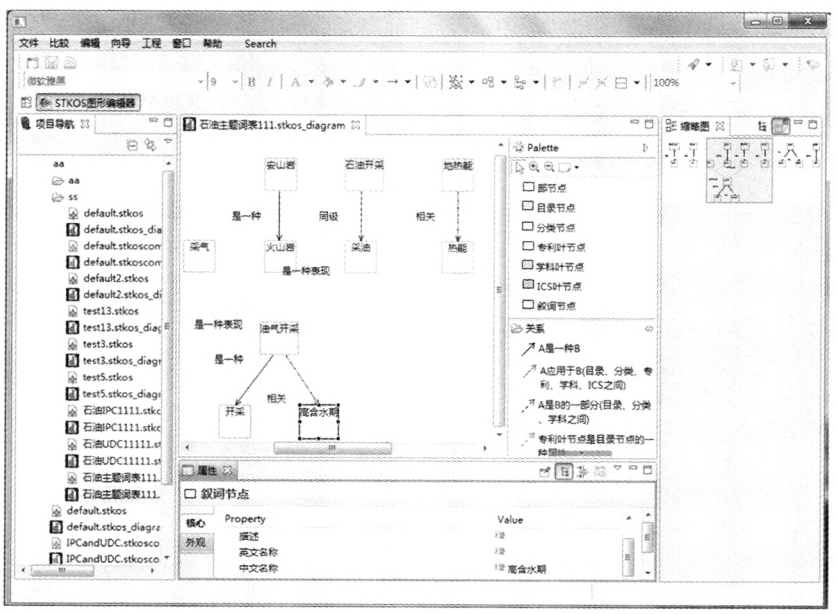

图 6　系统实现后的叙词表的结果

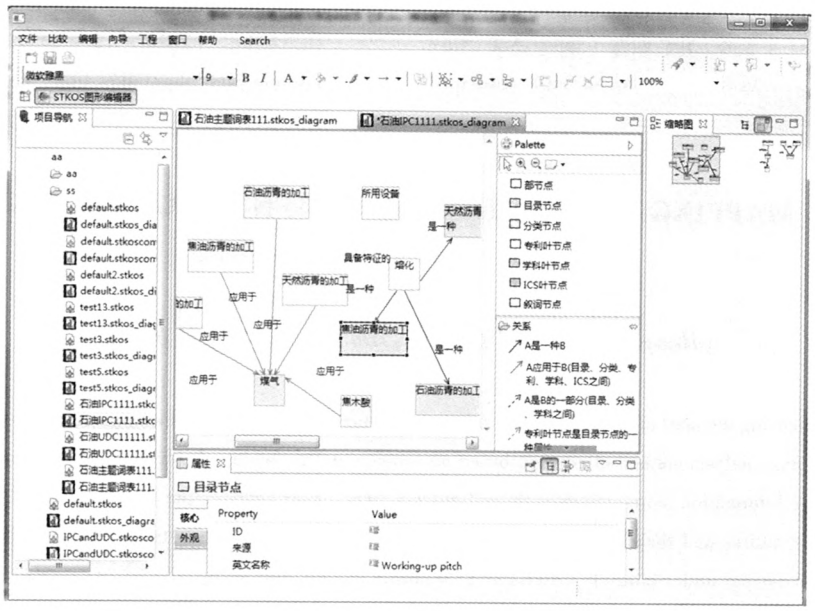

图7　系统实现后的IPC的结果

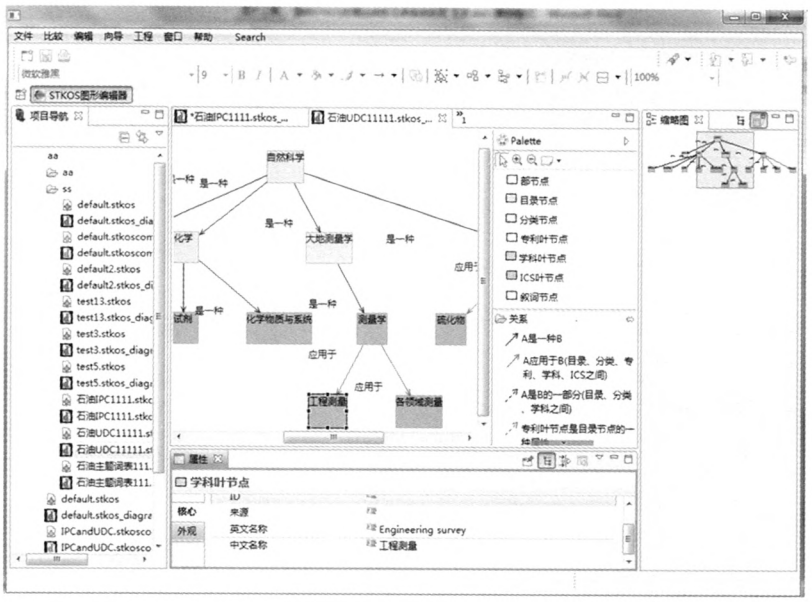

图8　系统实现后的UDC的结果

4　总结

本文通过对科技知识组织体系进行概念上的映射,构建了统一的概念空间(用公共数据模型描述)。然后基于统一的概念空间,并通过开发平台Eclipse和GMF开发了具有数据导入、数据导出及数据可视化功能的系统,实现了对概念的可视化图形编辑和导入。通过概念的映射和系统的实现,使得我们在这个信息呈爆炸性增长的时代,可以更快捷、方便的对概念进行检索,同时也可以更加直观地了解概念间的关系。

参考文献

[1] 高祀亮. 高校图书馆情报信息服务的新模式[J]. 现代情报, 2006, 3(3):138-139.

[2] 曹树金.《杜威十进制分类法》的电子化及未来研究重点[J]. 图书馆, 1995(5):19-21.

[3] 曹树金, 严丽君, 汪东波. DDC、LCC、UDC网络版评价[J]. 中国图书馆学报, 2002(6):61-65.

[4] Tony Olson. The Integration of Information languages and Interoperability[EB/OL]. [2011-07-20]. http://www.ala.org/ala/lita/litamembership/litaigs/2002authcontrol.pdf.

[5] 马磊, 宋建玮. IPC分类法在科技查新工作中的应用[J]. 图书馆学刊, 2012(3):32-33.

[6] 曹树金, 严丽君, 汪东波. DDC、LCC、UDC网络版评价

[J]. 中国图书馆学报, 2002(6):61 - 65.
[7] 浙江杭诚专利商标事务所. IPC 国际专利分类表[EB/OL]. [2012 - 12 - 25]. http://www.hczl.com/ipc/ipc.htm.
[8] UDC Consortium Office. Universal Decimal Classification [EB/OL]. [2012 - 12 - 25]. http://www.udcc.org/udcsummary/php/index.php? lang = chi.

CONCEPT MAPPING AND SYSTEM IMPLEMENTATION ORIENTED STKOS

WANG Bo

(*Hangzhou Dianzi University*, *Hangzhou* 310018, *China*)

Abstract: With the increasing demand of science and technology literatures knowledge and the explosive growth of information, this paper puts forward a mapping analysis method of concept based on science and technology knowledge organization system, and realizes its systematization to inquire information more conveniently and quickly. The paper analyzes the heterogeneity of a variety of standard classification systems and literatures and thesaurus in all areas and do mapping. The experimental results show that the method can search concepts of classification system under multiple criteria and thesaurus quickly and easily.

Key words: STKOS; classification system; map; literatures; concept

面向STKOS的概念映射及其系统实现

作者： 王波， WANG Bo
作者单位： 杭州电子科技大学 浙江杭州310018
刊名： 南阳理工学院学报
英文刊名： Journal of Nanyang Institute of Technology
年，卷(期)： 2012,4(6)

参考文献(8条)

1. 高祀亮 高校图书馆情报信息服务的新模式 2006(03)

2. 曹树金 《杜威十进制分类法》的电子化及未来研究重点 1995(05)

3. 曹树金;严丽君;汪东波 DDC、LCC、UDC网络版评价 2002(06)

4. Tony Olson The Integration of Information languages and Interoperability 2011

5. 马磊;宋建玮 IPC分类法在科技查新工作中的应用 2012(03)

6. 曹树金;严丽君;汪东波 DDC、LCC、UDC网络版评价 2002(06)

7. 浙江杭诚专利商标事务所 IPC国际专利分类表 2012

8. UDC Consortium Office Universal Decimal Classification 2012