

DDC 关联数据实现研究

贾君枝 赵 洁

摘 要 分类法作为规范的知识组织工具,实现其关联数据化,对书目数据的开放和网络知识组织体系的构建具有重要意义。《杜威十进分类法》是图书情报领域使用最为广泛的分类法,现已将其类目以关联数据的形式发布。本文基于 Dewey. info 平台,以 DDC 关联数据为例,通过对 DDC 关联数据发展脉络的梳理和技术构成的分析,从授权模型确定、实体对象的确、URI 命名、SKOS 模型化、RDF 表示五个方面解析 DDC 关联数据的具体实现。以此为我国《中图法》及其它分类法的关联数据化提供一定的参考。图 1。表 2。参考文献 13。

关键词 DDC 关联数据 Dewey. info 分类法

分类号 G254. 1

The Realization of DDC as Linked Data

Jia Junzhi & Zhao Jie

ABSTRACT The classification is a normative tool for knowledge organization. To publish classification as linked data is really important for the opening of bibliographic data and the construction of networked knowledge organization system. *Dewey Decimal Classification* is the most widely used classification in the field of library and information science. OCLC has published part of it as linked data. Based on the dewey. info platform, this paper takes DDC linked data as an example to study the concrete realization of DDC linked data by the following five basic steps, license model determination, entity object definitude, URI naming, SKOS modeling and RDF representation. This study is expected to provide a reference to the linked data publishing of *Chinese Library Classification* and other taxonomies in China. 1 fig. 2 tabs. 13 refs.

KEY WORDS DDC. Linked Data. Dewey. info. Classification.

网络知识组织体系(NKOS)是网络环境下知识组织体系的电子化描述,用于支持网络信息的表示与检索^[1]。关联数据是构建网络知识组织体系的重要工具,同时也是实现书目数据开放的有效途径。目前,书目记录、名称规范档、分类工具、叙词表及本体已相继实现了数据关联化。瑞典国家图书馆、英国哈德斯菲尔德大学图书馆、美国国会图书馆、OCLC、欧洲数字图书馆、德国国家图书馆等机构相继实现了部分资源的关联数据发布,并与其它资源建立了链接,一些机构还计划或已经将开放数据链接到关联数据云中。

分类法是成熟、规范、有效的知识组织工具,与其它知识组织工具有着不同的组织目的,适合学科导航、领域划分、数据挖掘等应用,有利于实现知识组织体系的互操作^[2-3]。在语义网发展的环境下,传统分类法由于类目众多、层次很深、形式单一、相对孤立等原因,无法进行有效的自动分类,也无法扩展分类法在图书情报领域之外的网络资源中的应用,阻碍其进一步的发展。随着越来越多的资源实现了关联数据化,分类法只有加入该行列,才能在未来利用庞大的关联数据网络,扩展分类法在其它领域的可用性,提高数据的重用性。因此,关联数据

化的分类法是网络知识组织体系的重要组成部分。《杜威十进分类法》作为世界图书馆使用最广、最具权威性的分类法,实现其关联数据化,不仅能够更好地链接各个图书馆内的资源,而且能够促进图书馆书目数据与网络资源的融合,对促进开放书目数据发展及资源发现具有重要意义。

1 DDC 关联数据的发展

OCLC(Online Computer Library Center,联机计算机图书馆中心)一直以来致力于将关联数据原则应用到《杜威十进分类法》中,并将 DDC 数据以“术语服务”的形式呈现^[4]。目前,在关联开放数据云(Linked Open Data Cloud)中,DDC 关联数据已占据一席之地。

1.1 DDC 关联数据的发展过程

Dewey.info 是 DDC 关联数据的发布平台,由 OCLC 杜威团队研究人员开发^[4]。该平台目前正处于试验阶段。2009 年,DDC 分类法首次转换为 DDC 关联数据,DDC 将 DDC 22 足本中最高三级类目以 11 种语言发布为关联数据。2010 年,Dewey.info 增加了英语、意大利语和越南语 DDC 节本 14 版中的指定类号和类目,这次扩展对已发布的每种语言的 DDC 关联数据增加了 3,500 条记录。之所以选择这三种语言,是因为它们都能以统一的 ESS XML 格式获取,而该格式已经完成了 RDF/SKOS 转换。2012 年 6 月,Dewey.info 将 DDC 23 足本中 000-999 的完整类目发布为关联数据,增加了超过 38,000 个类号,使可用的 DDC 关联数据增加了近十倍^[5]。2012 年 8 月,Dewey.info 增加了最新的 DDC 23 足本附表二地理区域、历史时期的指定类号和类目^[5],以 2--0 作为表二的类号,Table 2. Geographic Areas, Historical Periods, Biography 作为其类名,且目前只有英语一个语种。至此,DDC 完成了从 DDC 22 最高三级类目到 DDC 23 完整版的关联数据发布。接下来,DDC 将整合更多语言的 DDC 关联数据版本,并建立

与 LCSH 等映射词表的关联。同时,还将在表二中的地理区域部分和 GeoNames 叙词表之间建立链接,以管理 OCLC ESS(电子编辑支持系统)中所有已经编辑好的数据。

1.2 DDC 关联数据的应用及影响

DDC 发布为关联数据后,各大图书馆都与其建立了联系,如法国国家图书馆、大不列颠图书馆、德国国家图书馆已经在书目记录与规范文档中使用了杜威关联数据^[5]。联合国粮农组织(FAO)的农业多语种叙词表(AGROVOC)也关联到 Dewey.info^[5]。与此同时,研究人员也在探索 DDC 与虚拟国家规范文档(VIAF)、主题术语的分面应用(FAST)、ISO 639-3 语言编码、数学主题分类表(MSC)等规范词表的关联。

相较于 WebDewey 而言,Dewey.info 取消了 WebDewey 的用户授权,将语种从 WebDewey 的英语增加至 11 个语种,从而使 DDC 数据面向的用户扩展到了普通用户,为 DDC 数据的进一步开放奠定了基础。DDC 关联数据目前在以下三方面得到应用:

(1) 查询服务:这是 DDC 关联数据最简单的一种应用,通过查询杜威类号以获取与类号相关的类名、上下位、注释等信息。查询方法有两种:一种是通过 URI 直接查询,二是通过在 SPARQL endpoint 中输入相关 SPARQL 语句以返回相应的结果。

(2) 分面信息挖掘:通过在附表类目与主表类目间建立链接以合成组合类目。本文以现已发布为关联数据的地理区域为例进行说明。地理区域表中包含各种具体的地理位置,而主表中有些类是以地理位置进行区分的,如 641.59 Cooking characteristic of specific geographic environments, ethnic cooking,表示的是特定地理环境下的烹饪特点,在 641.59 与地理区域表间建立链接,可将 641.59 中的特定地理环境细化为具体的地理位置。采用这种分面的方法,便可对新信息进行挖掘。

(3) DDC 与 GeoNames 的链接:GeoNames 用于标识地理实体,不同词表、领域下的关于 GeoNames 同一地理实体的语词以 GeoNames 为

中介可建立关联。因此,建立 DDC 与 GeoNames 的链接,有助于利用 DDC 标识其它资源,从而扩展 DDC 在其它领域的应用。

2 DDC 关联数据的具体实现

基于 Tim Berners-Lee 于 2006 年提出的关联数据所遵循的四个原则^[6],Dewey. info 为每一个杜威分类法中的概念分配固定的 URI,并能根据实际查询情况动态转换 URI。在合适的授权模型下,提供针对机器和用户的两种表示方式,以使用户可以访问杜威数据。通过 SKOS 及其它词表模型化 DDC 数据,构建数据之间的 RDF 链接。提供 SPARQL 接口以查询杜威类目。DDC 关联数据的实现包含以下五个基本部分:授权模型确定,实体对象的明确,SKOS 模型化,URI 命名,RDF 表示。

2.1 授权模型确定

授权模型类似于 HTTP 的开放应用协议^[7],通过明确关联数据发布和使用过程中的各项法律问题,包括所有权、发布权、使用权、收益权等,针对不同对象限定不同的权限,构建合理利用关联数据的保护机制,以达到促进数据开放、保证开放数据安全、提高数据重用性的目的。授权模型提供的开放许可是语义网环境下关联数据能够真正开放并长久发展的必要条件。授权模型的授权许可嵌入数据后,用户在访问数据时无需征得数据发布者的许可,授权许可包含了用户被许可的所有操作,没有授权许可的操作将不予执行。目前已有的授权模型有:知识共享许可协议(Creative Commons License)、免费文献许可协议(GNU Free Documentation License)、共享文件许可协议(Common Documentation License)等,各个协议能够实现对不同类型开放数据的保护。

OCLC 发布 DDC 关联数据的目的是为了将 DDC 数据融入关联数据网络中,使其在公共领域发挥作用,因而选择较为通用的知识共享许可协议作为授权模型。Dewey. info 中的数据是

在署名—非商业性使用—禁止演绎(Creative Commons BY-NC-ND)授权模型下获取的^[8],表示只允许他人下载作品,而不能改变作品,也不能出于商业目的使用作品^[9]。该授权模型中的授权信息嵌套在 DDC 数据的 RDF/SKOS 表示中,遵循知识共享权利表示语言(Creative Commons Rights Expression Language)^[10]。在 RDF/SKOS 记录中,通过 cc 命名空间引用属性,并对数据使用进行授权。

2.2 实体对象的明确

DDC 关联数据是对 DDC 全部数据关联化,明确 DDC 数据的实体对象有助于数据的 SKOS 模型化和 URI 命名。Dewey. info 中的 DDC 关联数据有两个版本:DDC 22 的最高三级类目标和 DDC 23 完整类目。DDC 23 版本的组成部分与 DDC 22 版本大致相同,故以最新版的 DDC 23 为例,明确 DDC 数据的实体对象。

实体对象即 DDC 数据的数据类型,数据类型取决于数据所属的数据集类型。DDC 第 23 版由四卷内容组成:卷一包含的内容最多,包括第 23 版新特性、简介(introduction)、词汇表(glossary)、简介和词汇表的索引(index)、使用手册(manual)、附表(tables)。第 22 版和第 23 版的对比;卷二为 DDC 简表(summaries,DDC 最高三级类目)、主表(schedule:000-599);卷三为主表(schedule:600-999);卷四为相关索引(relative index)^[11]。由以上四卷内容可知,DDC 提供了以下不同完整程度和不同功能的数据集,分别为:分类法(scheme)、主表(schedule)、附表(table)、类(class)、使用手册(manual)、索引(index)、简表(summary)。这些数据既是 DDC 数据具体的存储位置,也是访问相应数据的入口。从不同的数据集进入,得到不同类型的数据,且数据有版本、格式、语言等方面的差异。需将数据集、数据及版本、格式、语言等以实体的形式表示。目前,DDC 关联数据仅对主表和附表中的表二实现了关联数据化,且附表中的表二也是以类的形式进行关联数据化。表 1 总结了 DDC 关联数据中的实体及实体属性。

表 1 DDC 关联数据实体

实体	实体属性
对象集	分类法、主表、附表、类、使用手册、索引、简表
对象	每一个类、附表中的九张表、索引内容
对象版本	DDC 22、DDC 23
版本时间	具体修改的年/月/日
对象语言	DDC 22 简表(14 种语言)、DDC 23 完整版(英语)
对象格式	面向用户的 html 格式、面向机器的 rdf 格式

各个实体之间存在一定的关联关系,Dewey.info 通过引用命名空间及其属性来表明实体间的关系。图 1 给出了实体之间的关联关系。对象是对象集中的元素,每一个版本、格式、语言的对象集均可作为一个 scheme。Scheme 中的对象有版本、语言、格式、版本时间的不同。版本、语言及版本时间通过 dct 命名空间中的属性进行标记。格式在 rdf 文件中并没有引用命名空间,而是通过在 URI 末端添加 html 或 rdf 直接获取相应的文件。具体如图 1 所示:

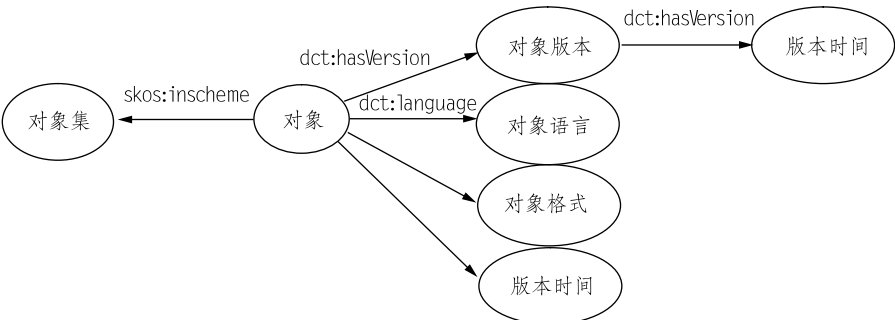


图 1 实体间的关联关系

2.3 URI 命名

关联数据是通过 URI 识别实体的,明确实体之后,要对实体进行 URI 命名。通过 URI 定位,能够识别到 DDC 中的所有实体。Dewey URI 具有以下两种模式:

```
http://dewey.info/{ object-collection }/{ object }/{ snapshot-collection }/{ snapshot }/about
http://dewey.info/{ object-collection }/{ object }/{ snapshot-collection }/{ snapshot }/{ resource-name }. { language }. { content-type }
```

这两种模式本质上是相同的。首先以 http://dewey.info 作为基本地址,该地址即 DDC 关联数据的发布平台。“object-collection”为实体中的对象集,其属性值可以有 scheme、table、class、manual、index、summary、id,必须指定对象集才能访问其中的对象。对象集后必须有相应的对象,URI 中用“object”表示,否则 dewey.info 无法找到相应的 URI。实体中的对象版本由

“snapshot-collection”来表示,其值有 e22、e23。版本时间由“snapshot”表示,格式为“年-月-日”,如 2009-08、2012-10-24 等。对象语言和对象格式分别对应“language”和“content-type”。其中语言取值有 14 种,如 en(英语),es(西班牙语),it(意大利语),fr(法语),zh(汉语)等。对象格式有 rdf、html。“resource-name”指定返回客户端的信息资源内容,通常指带有可变资源名称成分和特定语言、格式的具体文本。

“about”表示抽象概念的一般描述,而不是概念本身^[12]。抽象概念本身在网络上是不能表示的,服务器在收到查询抽象概念的请求时会指向与抽象概念有关的描述。不带有“about”字段的 URI 定位到 DDC 的具体类。查询时,会根据“303 重定向”到能够接受 HTML 或 XHTML 的客户端,即目前普通的网络浏览器,并自动转换为带有“about”字段的 URI,如 http://dewey.info/class/641/会自动转换为 http://dewey.info/class/

641/about。“about”字段也可以表明所描述概念的语言和格式,只需在“about”后面添加“language”和“content-type”标签,即 {about}. {language}. {content-type}。URI 的具体实例如下:

```
http://dewey.info/class/641/about.fr
http://dewey.info/class/641/about.fr.rdf
http://dewey.info/class/576.83/ancestors
```

(表示该类完整的向上的等级结构)

```
http://dewey.info/class/576.83/children
```

(表示该类向下的直接子类)

http://dewey.info/class/572/history.rdf (表示以 RDF 视图提供该概念随着时间发生的变化)

```
http://dewey.info/scheme/2008/07/up-
dates.de.atom
```

2.4 SKOS 模型化

SKOS 提供了一整套概念体系,包括概念标签、等级关系描述标签、注释标签、映射关系等,其作为一种公共格式能够使数据具有互操作性和可扩展性^[13]。由于关联数据最终要以 RDF 语言进行编码,而 SKOS 核心词汇表是 RDF 的应用,并且是目前知识组织系统模型化使用较多的词表。运用 SKOS 表示《杜威十进分类法》中的语义元素与类目体系,便可实现分类法体系结构的 RDF 化。

SKOS 模型化时,将 DDC 的每一个实体对象作为概念表 (ConceptScheme),概念表下包含若干个概念 (Concept),对应 DDC 类目数据,其中 DDC 分类体系中的十个基本大类被看作是顶层概念 (hasTopConcept)。分类号采用 skos:notation 标记,类名采用 skos:preLabel 标签表示。

分类法中类之间的关系主要分为等级关系、相关关系、等同关系。Dewey.info 平台中的关联数据仅对等级关系进行了描述,用 skos:broader 与 skos:narrower 两个标签分别表示上位关系与下位关系。

2.5 RDF 表示

Dewey.info 中的 DDC 数据有针对机器和用户的两种描述方式。用户版本采用 XHTML 和 RDFa 两种方式组合表示,机器版本采用 RDF 进行描述。机器版本有三种不同的 RDF 序列:RDF/XML、Turtle 和 JSON。经过 URI 命名和 SKOS 模型化之后,便可以运用 RDF 语言来描述实体及实体之间的关系了。

2.5.1 命名空间引用

命名空间及其属性的使用为数据之间的互操作及可扩展性提供了可能。RDF 语言通过引用命名空间中的属性来对数据进行语义描述,以实现数据的互联。Dewey.info 中,采用表 2 所示的命名空间及属性来描述实体类型及实体之间的关系。命名空间在使用之前,要首先对命名空间进行声明,dewey.info 中的声明语句如下:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dct="http://purl.org/dc/terms/"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xmlns:xhv="http://www.w3.org/1999/xhtml/vocab#"
xmlns:cc="http://creativecommons.org/ns#">
```

表 2 RDF/SKOS 表示中用到的命名空间及其属性

URI	缩写	属性
http://www.w3.org/1999/02/22-rdf-syntax-ns#	rdf	Description、about、type、resource、datatype
http://www.w3.org/1999/xhtml/vocab#	xhv	license
http://creativecommons.org/ns#	cc	attributionName、attributionURL、morePermissions
http://purl.org/dc/terms/	dct	isVersionOf、language、created、modified、hasVersion
http://www.w3.org/2004/02/skos/core#	skos	Concept、ConceptScheme、inScheme、hasTopConcept、topConceptOf、notation、preLabel、narrower、broader

2.5.2 授权模型引用

Dewey. info 中引用了知识共享许可协议。在命名空间声明部分,对该协议的命名空间进行了声明。在 RDF 表示过程中,主要使用了该命名空间中的三个属性(见表 2),分别是授权单位、授权 URL、更多授权许可。相应的 RDF 语句如下:

```
<cc: attributionName>OCLC Online Computer
Library Center, Inc. </cc: attributionName>
<cc: attributionURL rdf: resource = " http://
www. oclc. org/dewey/" />
<cc: morePermissions rdf: resource = " http://
www. oclc. org/dewey/about/licensing/" />
```

2.5.3 类及类间关系的 RDF 描述

类及类间关系的 RDF 描述是整个 RDF 描述的主体,涉及上面提到的多个命名空间的使用。在进行表示的过程中,要充分利用 URI 的层级识别和类间的等级关系,以 RDF 三元组的形式清晰地表示分类法的等级关系。RDF 三元组中,主语和宾语主要是 DDC 中的实体,用 URI 表示。谓语为表示主语和宾语关系的命名空间属性。每一个主语即实体的 RDF 描述要涵盖该主语的所有属性关系,如数据类型、授权信息、上下位等级关系、所属 scheme、类名、类号等。不同的实体要在不同的 RDF 文件中描述。以“6 Technology”为例,对应的描述语句如下:

```
<rdf: Description rdf: about = " http://dewey.
info/class/60/e23/2012-10-24/" />描述大类 60
<rdf: type rdf: resource = " http://www. w3.
org/2002/07/owl # Thing "/>类型为 owl 中的 Thing
<rdf: type rdf: resource = " http://www. w3.
org/2004/02/skos/core#Concept "/>类型为 skos
中的概念
.....
```

```
<dct:isVersionOf rdf: resource = " http://dewey.
info/class/60/e23/" />版本声明
<skos: notation rdf: datatype = " http://dewey.
info/schema-terms/Notation " > 60 </skos: notation
/>类号声明
```

```
.....
<skos: prefLabel xml: lang = " en " >Technology
</skos: prefLabel>类名为 Technology
<skos: narrower rdf: resource = " http://dewey.
info/class/600/e23/2012-10-24/" />下位类
<skos: narrower rdf: resource = " http://dewey.
info/class/601/e23/2012-10-24/" />
<skos: narrower rdf: resource = " http://dewey.
info/class/602/e23/2012-10-24/" />
<skos: narrower rdf: resource = " http://dewey.
info/class/603/e23/2012-10-24/" />
<skos: narrower rdf: resource = " http://dewey.
info/class/604/e23/2012-10-24/" />
<skos: narrower rdf: resource = " http://dewey.
info/class/605/e23/2012-10-24/" />
<skos: narrower rdf: resource = " http://dewey.
info/class/606/e23/2012-10-24/" />
<skos: narrower rdf: resource = " http://dewey.
info/class/607/e23/2012-10-24/" />
<skos: narrower rdf: resource = " http://dewey.
info/class/608/e23/2012-10-24/" />
<skos: narrower rdf: resource = " http://dewey.
info/class/609/e23/2012-10-24/" />
<skos: broader rdf: resource = " http://dewey.
info/class/6/e23/2012-10-24/" />上位类
.....
</rdf: Description>
```

以上代码中,谓语为 skos: narrower 和 skos: broader 的 RDF 语句均为等级关系描述语句,而多个下位关系的并列表示,间接表明了下位类之间的并列关系。需要注意的是,主要大类无上位类,最底层的类无下位类。

3 讨论

《杜威十进分类法》关联数据化的实现,是分类法实现关联数据化的起点。作为首个实现关联数据化的分类法,其它分类法在变身为关联数据的过程必然要借鉴 DDC 关联数据的实

现。《中图法》是我国使用最为广泛的分类法,在网络知识组织体系日渐发展的环境下,也面临着发布为关联数据的问题。本文从授权模型确定、实体对象的明确、URI命名、SKOS模型化、RDF表示五个实现步骤,对DDC关联数据做了

较细致的分析,以期为《中图法》的关联数据化提供参考。当前DDC关联数据并未完全实现,DDC中的注释说明、索引、附表等的关联数据化还有待进一步完善,研究者围绕此方面问题还需做深入的探讨。

参考文献

- [1] 王军,张丽. 网络知识组织系统的研究现状和发展趋势[J]. 中国图书馆学报,2008,34(1):65-69. (Wang Jun, Zhang Li. Network Knowledge Organization System: present status and development trends[J]. Journal of Library Science in China,2008,34(1):65-69.)
- [2] 贺德方. 国内外知识组织体系的研究进展及应对策略[J]. 情报学报,2010,29(6):963-972. (He Defang. Research advances in Knowledge Organization Systems at home and abroad and China's coping strategies[J]. Journal of the China Society For Scientific and Technical Information,2010,29(6):963-972.)
- [3] 张洁.《杜威十进分类法》的应用研究进展[J]. 图书情报工作,2013,57(18):110-115. (Zhang Jie. Application progress of Dewey Decimal Classification[J]. Library and Information Service,2013,57(18):110-115.)
- [4] OCLC. Dewey summaries as linked data[EB/OL].[2014-02-20].<http://www.oclc.org/dewey/webservices.en.html>.
- [5] Mitchell J S, Panzer M. Dewey linked data: making connections with old friends and new acquaintances[J]. JILIS (Italian Journal of Library and Information Science = rivista italiana di biblioteconomia, archivistica e scienza dell'informazione), 2013,4(1):177-199.
- [6] Berners-Lee T. LinkedData[EB/OL].[2014-03-17].<http://www.w3.org/DesignIssues/LinkedData.html>.
- [7] 张春景,刘炜,夏翠娟,等. 关联数据开放应用协议[J]. 中国图书馆学报,2012,38(1):43-48. (Zhang Chunjing, Liu Wei, Xia Cuijuan, et al. The open application licenses of linked data[J]. Journal of Library Science in China,2012,38(1):43-48.)
- [8] OCLC. Dewey Decimal Classification/Linked data[EB/OL].[2014-02-17]. <http://dewey.info>.
- [9] Creative commons. About the licenses[EB/OL].[2014-02-20].<http://creativecommons.org/licenses>.
- [10] Creative commons. CCREL[EB/OL].[2014-02-17]. <http://wiki.creativecommons.org/CcREL>.
- [11] OCLC. Introduction to the Dewey Decimal Classification[EB/OL].[2014-02-17]. <http://www.oclc.org/content/dam/oclc/dewey/versions/print/intro.pdf>.
- [12] OCLC. Using the API[EB/OL].[2014-02-20].<http://oclc.org/developer/documentation/dewey-web-services/using-api>.
- [13] Panzer M. DDC, SKOS, and Linked data on the Web[EB/OL].[2014-02-21].<http://www.oclc.org/news/events/presentations/2008/ISKO/20080805-deweyskos-panzer.ppt>.

贾君枝 山西大学经济与管理学院教授。

通讯地址:山西省太原市小店区坞城路92号。邮编:030006。

赵洁 山西大学经济与管理学院硕士研究生。通讯地址同上。

(收稿日期:2014-03-26)