

基于知识元的数字图书馆学科标引研究

刘 薇¹, 刘柏嵩^{2*}, 王洋洋¹

(1. 宁波大学 信息科学与工程学院, 浙江 宁波 315211; 2. 宁波大学 网络中心, 浙江 宁波 315211)

摘要: 对数字图书馆图书按中图法分类和学科分类进行了比较, 定义了新的知识元结构, 利用知识元中出现的概念, 通过中文 Wikipedia 概念库得到的丰富的语义信息, 对抽取的图书概念与学科分类类目之间计算语义相似度、语义关联度, 进行语义链接, 最后展示了初步实现的数字图书馆学科标引系统。

关键词: 知识元; 语义链接; 语义知识; 学科标引

中图分类号: TP391.1

文献标志码: A

文章编号: 1001-5132 (2013) 01-0100-04

计算机网络技术的发展, 不仅在改变用户的科研学习环境, 也在改变用户的科研学习惯和工作生活方式, 实现从信息服务转向知识服务的需求高涨, 数字图书馆作为用户获取知识的重要途径之一, 迫切地需要提升服务水平。

目前我国影响最大、使用面最广的是《中国图书馆分类法》(简称《中图法》)。根据 2010 年由北京图书馆出版社出版第 5 版《中国图书馆分类法》, 将人类的知识分成 22 个基本大类^[1]。《中图法》按学科知识的系统性分门别类地组织与揭示图书资料, 但是没有能够清晰揭示图书的学科属性, 给科研学习, 高校教学带了诸多不便。

学科是相对独立的知识体系^[2]。学科分类标引, 是根据书名、大标题、小标题、内容提取的知识元, 形成的关键词概念集合, 进行分类标引, 实现对图书的学科知识属性的梳理, 以便于人们对各学科内容的图书进行有效地表述、组织、传递和检索。

学科标引是知识标引的具体应用。知识标引将知识的控制单位从文献深化到文献中的数据、公式、事实、结论等最小的独立的知识元, 发现知识之间潜在的联系, 实现知识元的链接^[3]。链接最关键的一步是定义上下文概念和相关联实体的相似度量^[4]。笔者对数字图书定义新的知识元结构, 提

出通过中文 Wikipedia 概念库得到的丰富的语义信息, 对抽取的图书知识元中的概念与学科分类类目之间计算语义相似度、语义关联度, 进行语义链接, 初步实现数字图书馆学科标引系统。

1 知识元介绍

1.1 知识元的定义

知识元是指具有完备知识表达的不可再分割的知识单位。在知识结构中知识元是最小的元素, 可以由多个知识元直接组合构成新的知识单元。温有奎等^[5]定义知识元由名称、属性、操作、导航 4 要素组成。鉴于数字图书馆学科分类标引的特殊应用, 知识元的内容应包括以下几个方面: (1) 所属学科类别、关键词、分类号; (2) 知识元名称; (3) 知识元描述; (4) 关联条目。

1.2 知识元的结构

结合都柏林核心元数据(Dublin Core, DC)标准及数字图书馆学科分类的特点, 我们定义知识元结构为一个 7 元组 CKIE=(KEID, N, C, D, KC, KL, L), 其中: KEID 为知识元标识号, 采用 URI 表示; N 为知识元名称, 即知识元的标题, 是对本知识元的知识内容的一种高度概括表述。C 为知识元的关键字集, 即一组可用于检索本知识元的关键字集。

收稿日期: 2012-09-07.

宁波大学学报(理工版) 网址: <http://nbdxlg.paperopen.com/>

基金项目: 国家自然科学基金 (08CTQ014); 大学数字图书馆国际合作计划 (B2014)。

第一作者: 刘 薇 (1986-), 女, 浙江建德人, 在读硕士研究生, 主要研究方向: 知识工程、网络信息智能处理. E-mail: liuweil986@mail.nbu.edu.cn

*通信作者: 刘柏嵩 (1971-), 男, 安徽安庆人, 博士/研究员, 主要研究方向: 人工智能及计算机网络. E-mail: lbs@nbu.edu.cn

D 为简要说明,是对本知识元的内容的简要描述.通过该描述,一方面可使其在构建知识元结构模型时选择知识元的准确性提高;另一方面,在实现知识共享时,可以使知识使用者直接获得知识内容,而不需要再到庞大的载体中去寻找. KC 为知识类别,指对知识体系按照学科标准分成若干领域,并且与学科分类号分别对应;对知识体系按照一定标准划分后的各个子项,如采用学科为标准,知识类别可以分为生物、哲学、管理学、文学等领域,表现为层次关系. KL 为知识级别,按照认知方式分成概念、公理、规则和方法 4 类;知识级别由简单到复杂包括概念、公理、规则和方法 4 个层次. L 为知识地址,指在构建知识库时所赋予某一知识元的唯一位置标识,一般是一个链接到知识元所在的载体的超级链接.通过该信息,知识的使用者可以获得对该知识元更深入完整的信息.

1.3 知识元语义关系

知识元(概念)关键词之间存在着多样关系.分类关系主要考虑上下位关系(Is-A)和部分整体关系(Part-Of). Is-A 关系表示一个泛指概念或抽象概念的具体化, Part-Of 关系表示概念之间部分与整体的关系.非分类关系反映了概念间的某些语义关系,其类型多种多样,主要包括:空间关系、时间关系、序列关系、因果关系、源流关系、发展关系、推理关系、联想关系以及整体部分关系等^[6].概念可分为上位概念和下位概念.按同一标准划分并处于同一层面的概念称为同位概念.对任意的类 C_1 和 C_2 ,如果对于任意的概念 i ,若 i 是类 C_2 的实例可以推出 i 也是类 C_1 的实例,即类 C_2 的概念领域包含于类 C_1 的概念领域之中,则称 C_2 为 C_1 的子类;反过来, C_1 称为 C_2 的超类.

2 数字图书馆学科标引

2.1 数字图书馆学科分类结构

为了便于人们对各学科的图书内容进行有效地表述、组织、传递和检索,实现对图书的学科知识属性的梳理,按照学科分类体系,运用知识元标引技术,构建数字图书馆学科分类体系.层次结构能够很好地反映学科知识的属性,按学科门类展现,将每本图书分类到第 2 级或第 3 级子类目,清晰展示了数字图书的归属.分类层次结构框架如图 1

所示.

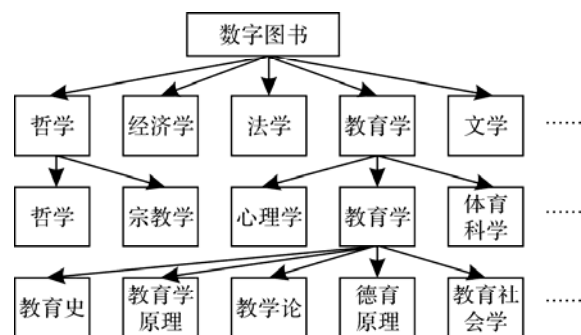


图1 数字图书馆学科分类层次框图(部分)

本案例假设图书知识元抽取已完成,以大标题处提取知识元概念集为主概念、从前言、章节标题、小标题、内容等处提取上下文概念为辅概念,从中选取相关知识元概念集同学科类目计算语义关联度、语义相似度,以此为依据进行学科分类标引.所有知识元的关键字(概念)集可以从中文 Wikipedia 挖掘出丰富的语义信息.

2.2 语义关联度

为了衡量链接关系的强弱,准确地将图书信息链接到相应的学科类目,采用文献[7]中描述的基于 Wikipedia 超链接结构的谷歌相似距离为范本,计算从图书提取的知识元(上下文概念)与学科类目之间的语义关联度.给定一个知识元 x 与一个学科类目 y ,他们之间的语义关联度为:

$$SmA(x, y) = 1 - \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}}$$

其中, $f(x)$, $f(y)$ 分别为链接到 x , y 的 Wikipedia 概念集合; $f(x, y)$ 为链接到 x , y 概念集合的交集; N 为所有 Wikipedia 概念的集合.每本图书与学科类目 y 的语义关联度 $SA(m)$ 为图书知识元概念集合 U 中的每个上下文概念语义关联度平均值:

$$SA(y) = \sum_{t \in U} SmA(t, y) / |U|$$

2.3 语义相似度

为了测量图书中提取的知识元概念 x 与学科类目 y 之间的语义相似度,首先确定他们超类之间的语义相似度,假设 x , y 的超类分别为 X 和 Y ,那么对于 X 中每一分类 S 与学科类目 Y 的语义相似度定义为 $sim(S, Y)$.为计算 $sim(S, Y)$,采用一种信息论方法,假设分类是一棵树, S 是一个类,

“ $s \in S$ ”的信息量为 $-\log(P(S))$, $P(S)$ 是随机选择的对象属于根为 S 的子树的概率. 假设 S_1, S_2 是2个分类, S_p 是离他们最近共同祖先, 则分类 S_1, S_2 之间的语义相似度为:

$$\text{Sim}(S_1, S_2) = \frac{2\log(P(S_p))}{\log(P(S_1)) + \log(P(S_2))}.$$

下面计算知识元类集 X 到学科类目集 Y :

$$\text{SimS}(X, Y) = \sum_{S \in X} \text{Sim}(S, Y) / |X|.$$

我们定义知识元概念 x 与学科类目 y 之间的语义相似度等于他们超类之间的语义相似度, 即 $\text{SimS}(x, y) = \text{SimS}(X, Y)$. 因此, 定义每本图书与学科类目 y 的语义相似度 $SS(y)$ 为图书知识元概念集合 U 中的每个上下文概念语义相似度平均值:

$$SS(y) = \sum_{t \in U} \text{SimS}(t, y) / |U|.$$

结合以上介绍的语义关联度和语义相似度, 对每本图书计算到任一学科类目 y 的链接可能性 $F(y)$:

$$F(y) = \alpha SA(y) + \beta SS(y),$$

其中: α, β 为常数, 根据经验所得. 每本图书选取与其链接可能性最大的学科类目进行链接, 完成学科分类标引. 若与同一层次的学科类目的链接可能性都小于一个特定阈值, 则回溯到上一层, 与该层次的学科类目, 进行语义关联度、语义相似度计算, 得到其链接可能性, 如此反复, 直至将图书链接到某层次的学科类目.

3 实验结果与分析

笔者采用的设计工具为 Delphi 和 Myeclipse, Delphi 用于开发图书的学科分类标引程序, C/S 模式. Myeclipse 用于开发展现学科分类书目信息, B/S 模式.

系统具体开发环境为: 开发用 PC 机; 操作系统为 Window 2003 企业版; Java 环境为 J2ee1.5; Web 服务器为 Tomcat5.3; 开发工具为 Delphi, Myeclipse, MSSQL2008 关系数据库系统.



图2 图书标引系统

图书标引系统如图2所示, 将需要标引的图书的上下文提取的知识元(概念)输入, 设定好参数, 标引系统依据知识元概念计算语义关联度和语义相似度, 将图书链接到相应的类目, 清晰地揭示图书的学科属性. 图书标引系统得到数据, 用于学科分类标引展现系统.

学科分类标引展现系统, 采用图书学科分类标引系统得到数据, 将图书信息按树形结构展现: 按学科门类展现首级目录(学科门类与学科关系); 按学科展现二级、三级等目录, 同时在右栏展现图

题名	第一作者	出版时间	出版社
当代西方教学模式		1991	山西教育出版社
我和老师		08 1 1987 12:00AM	编者
焊剂的性能与使用	陈处昌	1989 12	机械工业出版社
教师道德学	王庆之	02 1 1990 12:00AM	山东人民出版社
课程变革概论	白月桥	1999	河北教育出版社
教学论	游正伦	06 1 1982 12:00AM	教育科学出版社
教育调查研究法入门		05 1 1984 12:00AM	[不详]
教育目标分类学 情感领域	张云高	05 1 1989 12:00AM	华东师范大学出版社
教育目标分类学 情感领域	张云高	04 1 1989 12:00AM	华东师范大学出版社
怎样培养学习能力	李锐	08 1 1983 12:00AM	内蒙古人民出版社
给教师的建议	杜殿坤	01 1 1980 12:00AM	教育科学出版社
教育评价技术	张笛梅	09 1 1988 12:00AM	湖南教育出版社
教学拾萃	黄世利	01 1 1986 12:00AM	海南人民出版社
评价人员手册	龚伟民	06 1 1987 12:00AM	上海翻译出版公司
如何写评价报告	李定一	01 1 1988 12:00AM	上海翻译出版公司
如何测量方案实施	李正中	01 1 1989 12:00AM	上海翻译出版公司
如何处理评价目标	洪邦裕	01 1 1988 12:00AM	上海翻译出版公司
如何设计方案评价	王钢	08 1 1988 12:00AM	上海翻译出版公司
教育评价	李守福	01 1 1988 12:00AM	吉林教育出版社
论知识教学和智力发展		03 1 1982 12:00AM	人民教育出版社
教学论 教学理论的历史发展	吴杰	08 1 1986 12:00AM	吉林教育出版社
论教学过程最优化	吴文佩	06 1 1982 12:00AM	教育科学出版社
教育过程最优化问答	利兰	02 1 1985 12:00AM	北京师范大学出版社
教学教育过程最优化问答	冯克难	02 1 1986 12:00AM	教育科学出版社

图3 学科标引展现

书目录信息(题名、作者、出版社及 URI 等)。实验结果如图 3 所示, 学科标引展现成如下 3 层结构:

(1) 清晰地展示了系统按照学科属性, 将数字图书分为哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、军事学、管理学、其他等 13 个大类。

(2) 在教育学这个大类, 分为心理学、教育学、体育科学 3 个二级子类目。

(3) 在二级子类目教育学中又包括教育史、教育学原理、教育论、育德原理等 17 个三级子类目。

在教学论这个三级子类目的图书中, 如果简单地从书名、摘要入手, 很难将某些图书与教学论这一类目联系起来。学科标引系统清晰地揭示图书的学科属性, 给科研、学习、教学提供了便捷。

4 总结与展望

数字图书馆服务面临着如何深入到细小的知识单元, 进行组织、整理、服务的巨大挑战。学科标引可使知识被有效地检索、利用, 为用户提供针对性的知识服务, 在一定程度上满足了人类对知

识组织、知识管理、知识服务的需求。在以后的研究中可进一步完善充实数据库, 从规范元数据、提高中图法与汉语词表准确性、程序改进等方面着手, 提高学科分类标引的准确率。

参考文献:

- [1] 中国图书馆分类法编委会. 中国图书馆分类法[M]. 5 版. 北京: 北京图书馆出版社, 2010.
- [2] GB/T13745-92, 学科分类与代码[S].
- [3] 温有奎, 徐国华, 赖伯年, 等. 知识元挖掘[M]. 西安: 西安电子科技大学出版社, 2005.
- [4] Shen Wei, Wang Jian-yong, Luo Ping, et al. Linking named entities with knowledge base via semantic knowledge[DB/OL]. [2012-4-16]. <http://www.2012.www-conference.org/proceedings/proceedings/p449>.
- [5] 温有奎, 徐国华. 知识元链接理论[J]. 情报学报, 2003, 22(6):665-670.
- [6] 刘柏嵩. 基于 Web 的通用本体学习研究[D]. 杭州: 浙江大学, 2007.
- [7] 魏凯斌, 冉延平, 余牛. 语义相似度的计算方法研究与分析[J]. 计算机技术与发展, 2010, 20(7):102-105.

On Discipline Indexing of Digital Library Based on Knowledge Element

LIU Wei¹, LIU Bai-song^{2*}, WANG Yang-yang¹

(1.Information Science and Engineering College, Ningbo University, Ningbo 315211, China;

2.Network Center, Ningbo University, Ningbo 315211, China)

Abstract: In this paper, we compare e-book category based on Classification of disciplines with the one based on Chinese Library Classification (ZTF), and propose a novel method to define new knowledge element structure by using the concepts involving knowledge element and the rich semantic information derived from Chinese Wikipedia concept database. We compute the semantic similarity and semantic associativity between the extracted book concepts and the discipline categories, followed by linking the book to the category with semantic information. In the end, the proposed system of discipline indexing of digital library is presented.

Key words: knowledge element; semantic linking; semantic knowledge; discipline indexing

(责任编辑 史小丽)