# 基于图书馆语料库的英汉双语图书 Ontology 的构建

薛 峰[1],黄新艳[2]

(1.山东教育电视台,山东 济南 250014;2.山东经济学院 计算机科学与技术学院,山东 济南 250014)

**摘 要**:将单语 Ontology 转化为英汉双语 Ontology,主要涉及英汉双语图书语料的收集、抽取翻译等价对、用 OWL 语言描述单语图书 Ontology、双语图书 Ontology 的转换以及英汉图书 Ontology 的不断完善与发展等。构建的英汉双语图书 Ontology 能够有效地遵循用户的查询意图,极大地提高中英文文献的查全率和查准率,有利于读者对图书馆双语资料的使用。

**关键词**:图书馆语料库;英汉双语;图书 Ontology

**中图分类号**:G250.74 **文献标识码**:A **文章编号**:1008-7699(2011)02-0091-05

Ontology 原本是一个哲学概念,用于描述事物的本质,是对客观存在的系统的解释和说明,通常被译为本体论。在人工智能领域,Neches 等人将 Ontology 解释为定义了包含相关领域词汇的基本术语和关系,以及组合这些术语和关系定义词汇外延的规则。[1]其目标是捕获相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇(术语)和词汇(术语)间相互关系的明确定义。[2]近年来,Ontology 受到了广泛关注,在许多方面发挥着重要的作用。Ontology 应用于数字图书馆是研究的一个热点问题,具有重要的应用价值。美国的 IBM — DBZ 数字图书馆、英国的不列颠数字图书馆(The British Library)、已经数字化了 100 000 幅图像的法国国家图书馆(BNF)等都是 Ontology 应用于数字图书馆的研究成果。双语 Ontology 也具有重要的作用,而在将单语 Ontology 转化为英汉双语 Ontology 的过程中,主要涉及的问题是:英汉双语图书语料的收集;抽取翻译等价对;用 OWL 语言描述单语图书 Ontology;双语图书 Ontology 的转换;英汉图书 Ontology 的不断完善与发展等。

## 一、英汉双语图书语料的收集

收集英汉双语图书语料是构建英汉双语 Ontology 的第一步。构建本体一般是基于某个领域的本体。我们收集的语料库主要是应用在数字图书馆中。因此,我们选择计算机专业方面的英汉双语图书语料作为研究的目标。在研究中,我们选取了数字图书馆数据库中计算机类图书的 2 万多条双语信息,并且浏览了图书馆馆藏的核心书籍的摘要。

计算机专业类图书主要包括计算机的理论与方法、一般性问题、计算机软件、一般计算器和计算机、电子数字计算机、电子模拟计算机、微型计算机、多媒体技术、计算机的应用等。一个计算机专业的核心课程如下设置:专业基础课:程序设计(C,C++)、数据结构、模拟电路、数字电路;专业课:计算机组成原理、操作系统、汇编语言、数据库原理、编译原理;专业限选课:计算机网络与通讯、软件工程、图形学、人工智能、系统结构、图形学等。以上课程基本覆盖了计算机类的核心概念,通过对以上资料的收集和阅读,可以勾勒出计算机图书知识的总体框架。

**收稿日期**:2011-03-17

**作者简介**:薛 峰(1970-),女,山东济南人,山东教育电视台副研究馆员.

此外，在转换过程中对于语料库中没有的英汉对应词，我们采用参考《牛津中阶英汉双解词典》(第3版)与《计算机英汉辞海》，进行人工翻译的方式来解决。例如，如果“Computer Network and Communication”在语料库中找不到对应词的话，我们就会首先参考《计算机英汉辞海》将“Computer Network”翻译成“计算机网络”，“Communication”翻译成“通讯”，参考《牛津中阶英汉双解词典》将“and”翻译成“与”。这样，“Computer Network and Communication”就与“计算机网络与通讯”成了对应词。

## 二、翻译等价对的抽取

基于共现信息的翻译等价对获取的基本思想是在句子对齐结果(双语句对，也称句珠)的基础上，[3,4]统计双语词汇的同现概率，共现概率越大，则它们的关联强度就越大，就越可能成为翻译等价对。计算同现概率的模型有很多种，其中Dice系数，互信息，联列表和对数相似性为最常用的4种。[5,6]

这里假设：英语侯选单元为eu，汉语侯选单元为cu；a为同时出现英语单元eu和汉语单元cu的句对数；b为仅出现英语单元eu的句对数；c为仅出现汉语单元cu的句对数；d为不出现英语单元eu和汉语单元cu的句对数。如果语料库PC中共有n个句对组成，那么有n=a+b+c+d。可以利用下列模型来计算英语单元eu和汉语单元cu之间的关联程度：

(1)Dice系数的值介于[0,1]之间。数值越大，表示二者共现频率越大，越有可能成为对译词汇。

$$\mathrm{Dice}(\mathrm{eu},\mathrm{cu})=\frac{a}{(a+b)(a+c)}$$

(2)点式互信息是一种基于信息论中的互信息概念，来计算词间关联程度的方法。

$$MI(\mathrm{eu},\mathrm{cu})=\log_2\left[\frac{na}{(a+b)(a+c)}\right]$$

(3)联列表方法

$$\Phi^2=\frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Gale和Church等人设计的一个服从$X^2$分布的随机变量。

(4)对数可能性分值(Log Likelihood Ratio)

$$\mathrm{LLR}(\mathrm{eu},\mathrm{cu})=2[\log L(p1,a,a+b)+\log L(p2,c,c+d)-\log L(p,a,a+b)-\log L(p,c,c+d)]$$

其中，$\log L(p,k,n)=k\log(p)+(n-k)\log(1-p)$，$p1=a/(a+b)$，$p2=c/(c+d)$，

$p=(a+c)/(a+b+c+d)$，$\log(0)=0$。[7]376

利用上述四种计算同现概率的模型的任何一种，都可以计算出汉语文本以及英语文本中所有词对之间的同现概率，我们以同现概率最高的为最优翻译等价对。以“communication”为例，在英汉双语词典里关于“communication”的释义有：通讯，通信，交流，沟通。如果在我们的语料库中，与“communication”同现概率最高的为“通讯”，那么我们抽取的最优翻译等价对则为“communication”与“通讯”。[7]379

## 三、用OWL语言描述的单语图书Ontology

OWL，Web本体语言，是一种定义和实例化“Web本体”的语言，可以用来描述Web文档和应用中内在的类和关系。一个“Web本体”可能包含子类、属性和他们的实例描述。一般来说，OWL为我们提供了三种表达能力递增的子语言，以用于各个特定的实现者。OWL Lite用于提供给那些只需要一个分类层次和简单约束的用户。OWL DL支持那些需要最强表达能力的推理系统的用户。OWL Full支持那些需要尽管没有可计算性保证，但有最强的表达能力和完全自由的RDF语法的用户。使用OWL的本体开发者要考虑哪

种语言最符合他们的需求。我们采用的是 OWL Lite。

为了测试结果,用 Protégé 工具创建了一个 library. owl 的本体,主要用来描述馆藏书籍、论文及电子书籍等相关类及实例(如图 1)。该示例主要是做测试用的,定义了 Authors(作者)、Book(书籍)、Organization(组织机构)、Paper(文章)、Periodical(出版时间)五个大类,大类下面通过类继承层次生成若干个子类。

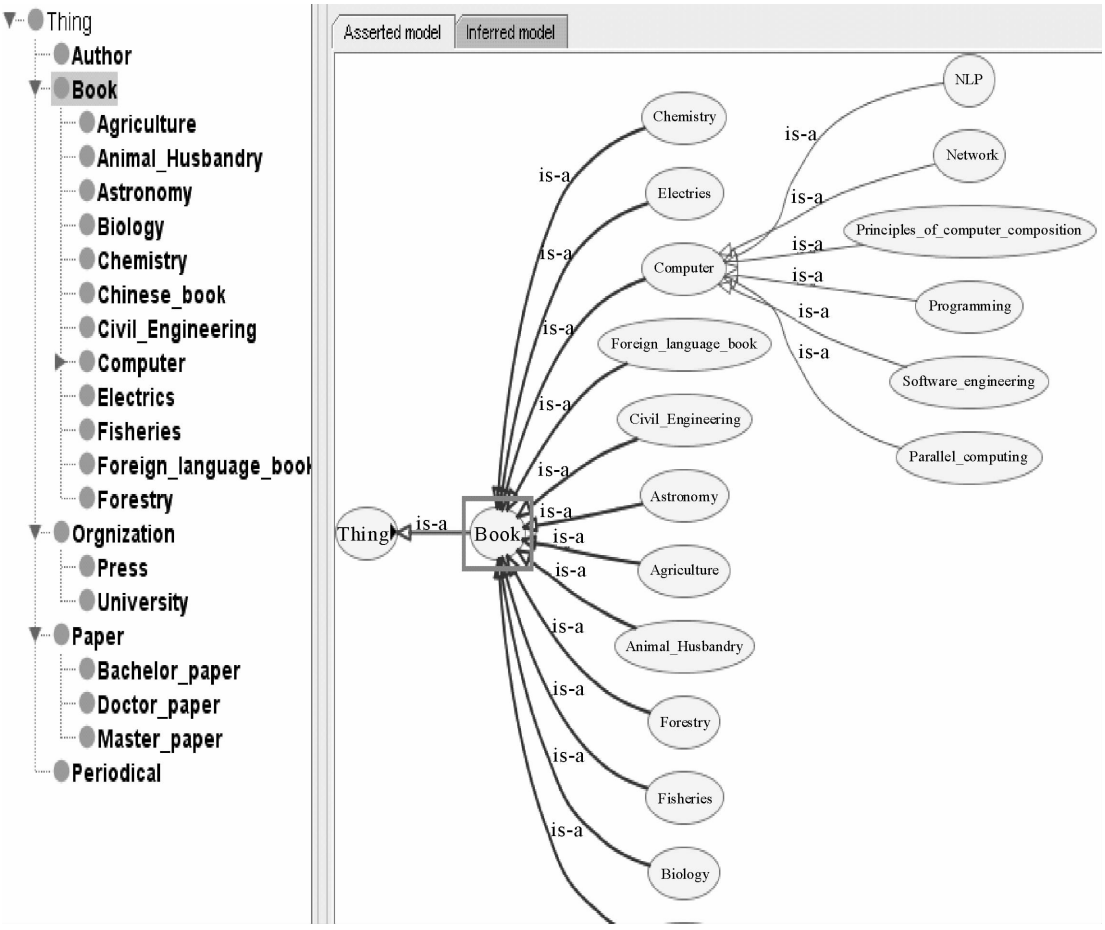


**图 1　用 protégé 构建的 library. owl 截图**

## 四、双语图书 Ontology 的转换

Jena 是来自于惠普实验室语义 Web 研究项目的开放资源,是语义 Web 和本体领域比较流行的 java 开发工具,目前的版本为 Jena 2. 5. 4。在研究中,我们用 Jean 来实现单语图书 Ontology(测试中我们用英语的)到英汉双语图书 Ontology 的转换。算法步骤如下:

1)首先将 OWL 描述的单语图书 Ontology 持久化成数据库中,生成相应的数据表格形式。测试中利用 MySQL 将 library. owl 持久化成如图 2 的结构。

这里将 OWL 文件转化到 MySql 关系数据库的表格形式。利用关系数据库的特点,可以很简单地将图中的英文类名字段转化为英汉双语类名字段。

2)利用前面所述的语料库将资源由英文表示翻译成相应的中文表示。图3显示的是转换之前的字段,

| Table Name | Engine | Rows | Data length | Index length |
|---|---|---|---|---|
| jena_g1t0_reif | InnoDB | 0 | 16 kB | 32 kB |
| jena_g1t1_stmt | InnoDB | 250 | 64 kB | 64 kB |
| jena_graph | InnoDB | 1 | 16 kB | 0 B |
| jena_long_lit | InnoDB | 0 | 16 kB | 16 kB |
| jena_long_uri | InnoDB | 0 | 16 kB | 16 kB |
| jena_prefix | InnoDB | 0 | 16 kB | 16 kB |
| jena_sys_stmt | InnoDB | 60 | 16 kB | 32 kB |

图 2 OWL 文件转化到 MySql 的表格形式界面图

图 4 显示的是转换之后相对应的字段。概念之间的关系并没有转换成中文,因为其可能会影响本体的推理能力。关于本体的推理,是我们将来要研究的问题。

从图 3 到图 4 的内容变化,我们可以发现本体中的类名例如“Doctor_paper”已经转化为“Doctor_paper | 博士论文”了(其中,“ | ”为我们规定的英汉连接符)。这样就实现了将本体中的英文类名转化为英汉双语类名的目的。

3)将翻译之后的本体从数据库写回至 OWL/RDF 文件,得到中英双语的图书 Ontology。

Resultset 1

| Subj |
|---|
| Uv::http://www.owl-ontologies.com/library.owl#Doctor_paper: |
| Uv::http://www.owl-ontologies.com/library.owl#Paper: |
| Uv::http://www.owl-ontologies.com/library.owl#Doctor_paper: |
| Uv::http://www.owl-ontologies.com/library.owl#Chinese_book: |
| Uv::http://www.owl-ontologies.com/library.owl#Book: |
| Uv::http://www.owl-ontologies.com/library.owl#Chinese_book: |
| Uv::http://www.owl-ontologies.com/library.owl#Foreign_language_book: |

图 3 翻译前的英文本体类界面图

Resultset 1

| Subj |
|---|
| Uv::http://www.owl-ontologies.com/library.owl#Doctor_paper博士论文: |
| Uv::http://www.owl-ontologies.com/library.owl#Paper论文: |
| Uv::http://www.owl-ontologies.com/library.owl#Doctor_paper博士论文: |
| Uv::http://www.owl-ontologies.com/library.owl#Chinese_book中文书籍: |
| Uv::http://www.owl-ontologies.com/library.owl#Book书籍: |
| Uv::http://www.owl-ontologies.com/library.owl#Chinese_book中文书籍: |
| Uv::http://www.owl-ontologies.com/library.owl#Foreign_language_book外文书籍: |

图 4 翻译后的中英文本体类界面图

## 五、英汉图书 Ontology 的不断完善与发展

转换后的英汉图书本体我们需要不断地完善与更新,不断地将一些新的英汉翻译等价对补充添加到该英汉图书本体中。[8]

为此,我们做了些小实验(例如添加一个新的书类“OperatingSystem | 操作系统”到该图书本体中)对构建的英汉图书 Ontology 进行更新和完善,其代码如下:

```
Resource rs=model.createResource("http://www.owl-ontologies.com/library.owl#
OperatingSystem | 操作系统");
Property prop=model.createProperty("http://www.w3.org/2001/XMLSchema#string");
RDFNode node=model.getRDFNode(
Node.create("http://www.owl-ontologies.com/library.owl#Computer"));
Statement stmt =model.createStatement(rs, prop, node);
model.add(stmt);
```

```
try{
FileOutputStream fo=new FileOutputStream(filePath);
model.write(fo,"RDF/XML-ABBREV");
fo.close();
catch(Exception e){
e.printStackTrace();
    }
```

通过该类程序就可以实现将一些新的英汉类,以及英汉实例等不断地补充添加到我们的英汉本体中,使该双语本体更加完善与实用,从而使其能够投入到数字图书馆的更广泛的应用领域中。

本文利用英汉双语图书语料探索了英汉双语图书 Ontology 的构建,主要涉及英汉双语图书语料的收集、抽取翻译等价对、用 OWL 语言描述单语图书 Ontology、双语图书 Ontology 的转换以及英汉图书 Ontology 的不断完善与发展等。按照本研究的思路构建的英汉双语图书 Ontology 能够有效地遵循用户的查询意图,提高了中英文文献的查全率和查准率,获得预期的检索信息,有利于读者对图书馆双语资料的使用。

**参考文献:**

[1]刘耀,穗志方.领域 Ontology 概念描述体系构建方法探析[J].大学图书馆学报,2006(5):31.

[2]潘红艳,林鸿飞,赵晶.基于 Ontology 的个性化推送系统[J].计算机工程与应用,2005(20):178.

[3]宋炜,张铭.语义网简明教程[M].北京:高等教育出版社,2004:102.

[4]GALE W. Identifying words correspondences in parallel texts[C]// In Proceedings of DARPA Speech and Natural Language Workshop. Asilomar: IEEE Computer Society, 1991:153.

[5]BECHHOFER S, HARMELENA F. OWL web ontology language reference[EB/OL]. 2009[2011-01-15]. http://www.w3.org/ TR/owl2-overview.

[6]Jena2 database interface[EB/OL]. 2004[2011-01-15]. http://jena.sourceforge.net/DB/index.html.

[7]李德俊.基于英汉平行语料库的词典编写系统 CpsDict 的研制[J].现代外语,2006(4).

[8]ZHOU Wei, WEI Zhiqiang, KANG Mijun, et al. A credit-based incentive mechanism for recommendation acquisition in multihop mobile ad hoc networks[C]// In 2009 Third International Conference on Emerging Security Information, Systems and Technologies. Glyfada:IEEE Computer Society, 2009:308.

## On the Construction of English-Chinese Bilingual Ontology Based on Library Corpus

XUE Feng[1], HUANG Xinyan[2]

(1. *SDETV*, *Ji'nan* 250014, *China*; 2. *School of Computer Science & Technology*, *Shandong Economic University*, *Ji'nan* 250014, *China*)

**Abstract**: The construction of English-Chinese bilingual ontology mainly involves: collection of English-Chinese bilingual corpus, extraction of the equivalent translated pairs, description of monolingual ontology using OWL, transformation of bilingual ontology and perfection of English-Chinese bilingual ontology. The English-Chinese bilingual ontology, following readers' intention effectively, can benefit them a lot in using the materials in library and greatly improve the accuracy and completeness of finding related data.

**Key words**: bilingual corpus; English-Chinese; bilingual ontology

(责任编辑:于凤银)