《智能信息处理》课程考试

一种有效的图书本体学习算法

吴敌

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 18 日

一种有效的图书本体学习算法

吴敌

（大连海事大学 信息科学技术学院 辽宁大连 116026）

摘 要：为了解决本体的学习和集成在本体构建中的瓶颈问题，提出一种本体学习算法；该算法利用树形结构，以分层的方式进行概念及其关系的添加与剪枝，设计相应的本体学习的 2 个主要过程；根据图书的检索特征，将图书本体分为图书外部本体和内容本体，可以分别学习图书的 2 类本体结构。结果表明：该本体学习算法可以有效地实现图书的准确检索。

关键词：本体；本体学习；本体构建

中图法分类号 TP302.1

Effective Algorithm for Book Ontology Learning

Wudi

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026，China)

Abstract:To solve the bottleneck problem of ontology learning and integration in ontology construction，an ontology learning algorithm was suggested．The algorithm applied the tree structure，added and pruned the concepts and their relations in a hierarchical manner and designed two procedures for the corresponding ontology learning．According to retrieval features of books，the book ontology was divided into external and content ontology to respectively learn two kinds of ontology construction of books．The results show that the algorithm can effectively realize the accurate retrieval of books．

Keywords: ontology; ontology learning; ontology construction

1 引 言

基于本体的图书检索利用本体对概念化对象明确的形式化规范说明[1]和具有的良好概念层次结构及对逻辑推理的支持，并且通过概念之间的关系来表达概念语义，能较好地为语义检索和概念检索提供知识基础。基于本体的智能信息检索优于关键词搜索，原因是本体包含机器可判断的概念，从而使系统对领域内的概念、概念之间的关系及领域内的基本公理知识有一个统一认识，系统通过分析用户提出的查询中所包含词(组)的语义，理解用户的查询，并准确地映射到信息资源，提高信息检索系统的查全率和查准率; 然而，领域本体的获取通常比较困难，特别是图书资源涉及到的领域较多，如果构造应用于图书检索的本体涉及到多个领域，构造图书本体的难度就比较大。本文中设计并分析一种有效的图书本体学习算法，根据图书的检索特征，将图书本体分为图书外部本体和内容本体，可以分别学习图书的这 2 类本体结构; 利用学习到的本体实现图书检索系统，并通过检索结果分析该本体学习算法的高效性和检索的准确率。

2 相关工作

本体一般被作为一个有向无环图来构造[2]。Gruber[3]最早给出了有效的本体设计原则，用于指导和评价本体设计的客观标准。从初始知识源中学习一个领域本体是本体构建的有效方法之一，目前已有多种支持这种本体构造的方法，这些方法在对本体语言的支持能力、表示能力以及可扩展性、灵活性、易用性等方面所实现的功能都不尽相同。

本体学习的主要目的是方便本体的构建，尤其是涉及多个领域知识时，利用机器学习实现知识获取的方法完成本体构建[4]，较为经典的是“平衡合作建模”范式，即结合人工与机器学习算法实现协同交互进行本体的构建[5]。文献[6]中给出了一种自动学习本体的方法，该方法利用通用型主体模型

生成的主题作为概念并构造这些概念之间的相容关系来学习本体，而无需种子本体。Jacinto 等[7]利用关联规则挖掘算法，从关系数据库中发现基本术语，用来描述某领域中的隐藏断言，从而实现本体的构建。

本文中针对已有的本体学习方法的不足，提出一种有效的本体学习算法。该算法利用树形结构，以分层的方式进行概念及其关系的添加与剪枝，设计相应的本体学习的 2 个主要过程，并将学习到的本体应用于智能图书检索过程中进行实验验证，根据实验结果分析本文中所提出的本体学习算法的有效性。

3 用于图书检索的本体结构

当使用本体论表示图书资料时，主要从 2 个角度描述一本图书或者一项资源: 一是图书外在资源，如题名、作者、出版者、出版年等; 二是内容资源，如图书的主题、目录或者图书内容等。2 个方面都可以使用本体来建立具有层次等级关系的知识模型，在层次等级结构的知识模型中可以定义图书信息概念、信息客体的内容和相关的属性与关系。图书资料本体构建的总体结构如图 1 所示。图书外部本体是指图书的外在属性，主要包括书名、分类号、国际标准书号(ISBN) 或国际标准连续出版物(ISSN) 、主要责任者、出版社、定价、版次、其他属性等。

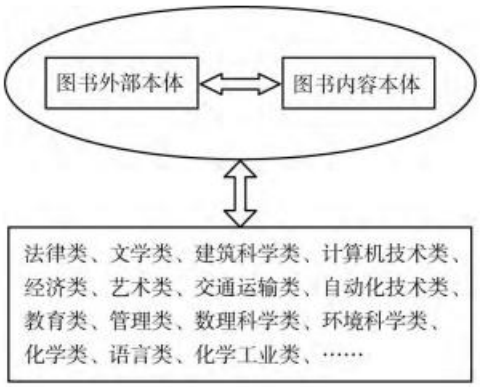

图 1 图书资料本体构建的总体结构

图书外部本体的结构如图 2 所示。图书资料用户在检索时，主要是通过图书外在属性进行输入搜索，因此图书的外部本体对于用户来说使用频率很高，而这些外在属性在很多时候需要语义规范，例如书名、出版社等。同一本图书由用户输入的名称可能不同，如“计算机基础”“计算机文化基础”“大学计算机基础”等名称从语义上讲都是相似度极高的图书，“人大出版社”“人民大学出版社”等都是可能由不同用户输入相同的出版社名称。建立图书外部本体尽管比其他语义本体较为简练，但对于用户的方便性和系统的应用与可扩展性等方面都是非常重要的。在图 2 中，分类号节点分为分类号 1、2、3 等，原因是很多专业图书实际上是跨专业学科领域的，也就是属于交叉学科。

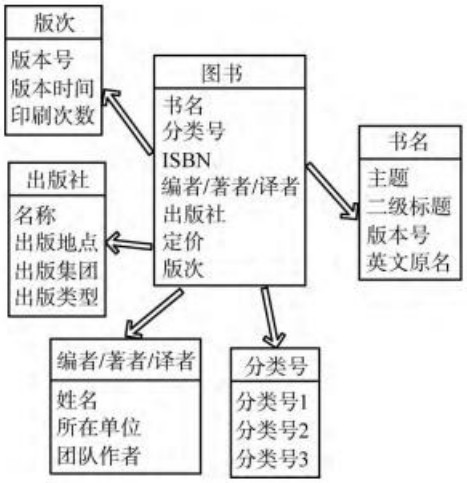

图 2 图书外部本体结构

图书外部本体的部分 OWL 表示如下:

〈owl: Class rdf: ID =“Book”〉
〈rdfs: subClassOf rdf: resource =“#User”/〉
〈/owl: Class〉
……
〈owl: DatatypeProperty rdf: ID =“Category”〉
〈rdfs: domain rdf: resource =“#User”/〉
〈rdfs: range〉
〈owl: Data Range〉
〈owl: oneOf〉
〈rdf: List〉
〈rdf: first rdf: datatype =“＆xsd; string”〉经济类
〈/rdf: first〉
〈rdf: rest〉
〈rdf: List〉
〈rdf: first rdf: datatype =“＆xsd; string”〉计算机类〈/rdf: first〉
〈rdf: rest rdf: resource =“＆rdf; nil”/〉
……
〈/ rdf: List〉
〈/ rdf: rest〉

〈/ rdf: List〉
〈/owl: oneOf〉
〈/owl: Data Range〉
〈/ rdfs: range〉
〈rdfs: comment
rdf: datatype =“&xsd; string”〉 〈/ rdfs: comment〉
〈/owl: DatatypeProperty〉
……
〈owl: DatatypeProperty rdf: ID =“Title”〉
〈rdfs: domain rdf: resource =“#Book”/〉
〈rdfs: range rdf: resource =“&xsd; string”/〉
〈rdfs: comment
rdf: datatype =“&xsd; string”〉 〈/ rdfs: comment〉
〈/owl: DatatypeProperty〉
……
〈owl: ObjectProperty rdf: ID =“Add_Resources”〉
〈rdfs: domain rdf: resource =“#Ordinary_Book”/〉
〈rdfs: range rdf: resource =“#Resource_Library”/〉
〈/owl: ObjectProperty〉
……
〈/ rdf: RDF〉 。

由于学科专业门类较多，涉及到的不同专业图书种类也很多，因此图书内容本体的构建较为复杂。

4 基于层次结构的本体学习算法

本体所涉及的概念具有层次结构，各层概念之间可以用一组关系进行连接。这些关系包含的类别有很多，如属性关系 is－a、部分与全局关系 part－of 等。本体也区别于一般的关系数据库。关系数据库是由数据库管理系统管理的关系数据的集合，因此，本体也不能简单地由关系数据库存储和管理。本文中从构建本体的角度实现本体，其描述如下:

1) 某领域相关的概念描述，记为 C_i，i = 1，2，…n，其集合可以表示为此概念集。

2) 所有概念不同属性的描述，记为 ak i，其中 k =1，2，…，m。

3) 属性约束描述，记为 sk i。

于是，本体中的概念表示为 C_i(ak i，sk i) 。

一个本体所涵盖的所有概念及其关系的层次结构可以记为一个五元组，T^C = (N，G，R，C_0，E_i)，其中，N 表示概念分层树结构的节点集，记为{ n_0，n_1，…，n_r} ; G 表示每一个节点对应的术语集{ g_0，g_1，…，g_r}，其中 g_j(j = 0，1，…，r)表示一个关键词，在概念分层树中是唯一存在的;R 表示概念分层树中节点之间的连线集合，也就是连接的 2 个概念的关系集合，记为“＜”，其中＜=〈g_a，g_b〉表示 g_a 是 g_b 的子节点; g_0 表示本体所描述领域的最抽象概念;E_i 表示本体所描述领域的最特化的实例。

概念分层树的学习即抽取过程，需要设计相应的学习算法。学习本体树结构过程由于受限于知识源的不完整性、不一致性和模糊性等问题，因此需要不断地对概念分层树节点添加、剪枝和求精处理。

4.1 节点添加算法

从知识源获取的术语集 K 只是领域中所有术语集 AK 的子集，即 K 包含于 AK，K = { k_0，k_1，…，k_m} ，AK = { k'_0，k'_1，…，k'_n}，m≤n。

利用符号“＜ ”表示概念之间的从属关系，即如果“a ＜ b”，则表示概念 a 是概念 b 的细节级概念，也就是 b 比 a 更抽象。设 k_i ＜ k_j，若∃k'_t ∈AK，k_i ＜ k'_t ，k'_t ＜ k_j，则将 k'_t 插入到概念分层树中，且

$$K \leftarrow K \cup \{ k'_t \} \ , \qquad (1)$$

$$R \leftarrow (R - \{<k_i, k_j>\}) \cup \{<k_i, k'_t>, <k'_t, k_j>\} \ . \qquad (2)$$

节点添加过程 AddItem 的实现算法如下。

1) 输入初始知识源 KS_0 ;

2) 从知识源中抽取初始术语集，构造初始本体 K_0 ;

3) i = 1;

do while(true) {

4) 更新知识源，记为 KS_i ;

5) while(KS_i 中存在新术语) {

6) if(k_i<k_j，and k'_t∈KS_i，k_i<k'_t ，k'_t<k_j)

7) 按照式(1)和(2)将 k'_t 添加到概念分层树中;

}

8) 根据概念分层树获取本体 K_i ;

9) i ++ ;

10) }until(无新的知识源，或学习完成) 。

4.2 剪枝算法

从知识源 KS 中自动获取的初始本体的一些概念不是领域中的概念，或者与领域关系不大，需要把一些无关紧要的概念及其关系删除掉，称为剪枝。记 $k_x <^* k_j$，k_x 表示 k_j 的所有后继节点，则

$$\begin{cases} K \leftarrow K - \{k_x\}, \\ R \leftarrow R - \{<^*>\}。 \end{cases} \quad (3)$$

剪枝算法的基本思想是，每进行一次知识源 KS_i 更新以后，检测概念分层树 K_i，如果存在无关概念，则删除该概念对应的节点及其子树，再对新更新的知识源 $KS_i + 1$ 进行下一步处理。剪枝过程 PruneTree 的实现算法如下。

1) i = 1;
2) while(true) {
3) 检测概念分层树 K_i;
4) if(K_i 中存在无关概念) {
5) 删除该概念对应的节点及子树，得树 K'_i;
6) $K_i \leftarrow K'_i$;
}
7) i ++;
8) } 。

4.3 求精过程

求精过程通常需要领域专家和知识工程师的参与。该过程一般涉及到相似概念的合并与相似度计算、本体之间的映射等，这些方面可以利用已有的方法来进行。

5 结束语

本文中提出了一种有效的本体学习算法，利用树形结构以分层的方式进行概念及其关系的添加与剪枝，利用概念分层原理，不断地利用概念的粒度大小，对概念树节点添加、剪枝和求精处理，设计了相应的本体学习的 2 个主要过程，即 AddItem 过程和 PruneTree 过程，在一定程度上较好地解决了获取本体的难度大、周期长等缺点。该方法可以从 2 个角度获取图书的本体结构: 一是图书的外部本体，如题名、作者、出版者、出版年等; 二是图书的内容本体，如图书的主题、目录或者图书内容等。结果表明，该系统的检索效率较高，实用性强。

参 考 文 献

[1] STUDER R，BENJAMINS V R，FENSEL D． Knowledge engineering，principles and methods［J］． Data and Knowledge Engineering，1998，25(1 /2) : 161 － 197．

[2] KHAN L，LUOF． Ontology construction for information selection［C］/ /Proceedings of 14th IEEE International Conference on Tools with Artificial Intelligence: ICTAI 2002: 4—6 November 2002，Washington，D C． Washington D C: IEEE Computer Society， 2002: 122 － 127．

[3] GRUBER T R． Toward principles for the design of ontologies used for knowledge sharing［J］． International Journal of Human－Computer Studies，1995，43(4 /5) : 907 － 928

[4] MAEDCHE A，STAAB S． Ontology learning for the semantic web［J］． IEEE Intelligent Systems，2001，16(2) : 72 － 79．

[5] MORIK K，WROBEL S，KIETA J U，et al． Knowledge acquisition and machine learning: theory，methods，and applications［M］． London: Academic Press，1993．

[6] LIN Z，LU R，XIONG Y，et al． Learning ontology automatically using topic model［C］/ /2012 International Conference on Biomedical Engineering and Biotechnology (ICBEB) ． Washington D C: IEEE Computer Society，2012: 360 － 363．

[7] SENTHILNAYAKI B，VENKATALAKSHMI K，KANNAN A． An ontology based framework for intelligent Web based e-learning［J］． International Journal of Intelligent Information Technologies，2015，11(2) : 1 － 17．