基于本体的信息检索研究

尹红健
(湖南化工职业技术学院，湖南 株洲 412004)

摘 要：介绍了本体 Ontology 的概念和理论知识，提出一种基于本体的 Web 信息检索模型。该模型利用本体技术对 Internet 上的各类信息进行领域分类，规范用户信息检索模式，以达到快速、准确找到用户所需信息的目的。

关键词：本体；信息检索；知识检索

中图分类号：TP316.2 文献标识码：B

Research of information retrieval system based on Ontology

YIN Hong Jian
(Hunan Vocational and Technical Institute of Chemical Engineering , Zhuzhou 412004, China)

Abstract: The thesis firstly introduces the concept and theory of Ontology, then presents web information retrieval model based on Ontology. In this model, information on the internet are classified and indexed by using of Ontology technology, and user queries are also normalized so as to achieve the purpose of finding the required information quickly and accurately.

Key words: Ontology; information retrieval; intellectual retrieval

随着计算机的普及与 Internet 的快速发展，我们已经进入了网络信息时代。信息的发布与共享不再受时空的限制，当网络规模越来越大，信息越来越多时，信息的查找和获取也变得越来越困难。面对庞大的信息资源，人们感到茫然，要在短时间内找到符合自己要求的信息越来越困难。

如何迅速、高效地检索和访问各领域的信息资源以促进信息的交流与共享已经成为一个急需解决的问题。人们迫切需要高效、准确的信息查找工具来快速定位自己感兴趣的信息和知识，现有的网络信息检索技术很难满足这种要求，基于本体 Ontology 的 Web 信息检索系统正逐渐成为当前研究的热点。

1 Ontology 的基本概念

1.1 Ontology 的定义

Ontology 最早是一个哲学上的概念，是研究“存在”的理论。从西方哲学史来看，Ontology 是指关于存在及其本质和规律的学说，是对客观存在的一个系统的解释或说明，关心的是客观现实的抽象。

Ontology 的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语和词语间相互关系的明确定义。Ontology 最为流行的定义是 Studer 在 1998 年提出的[1-2]：Ontology 是共享概念模型的明确的形式化的规范说明。它包含 4 层含义：概念模型、明确、形式化及共享。

1.2 Ontology 的组织方式

在计算机领域，作为一种语义和知识层面上的概念模型，Ontology 有其自身的结构，可以表示为[1-4]：本体(Ontology) = 概念(Concept) + 属性(Property) + 公理(Axiom) + 取值(Value) + 命名(Nominal)

Perez 等人用分类法组织了 Ontology，定义了 5 个基本的建模元语(Modeling Primitives)，其具体的描述表达意义如下：

(1)类(Classes)或概念(Concepts)：指任何事务，例如工作描述、功能、行为、策略和推理过程。从语义上讲，它表示的是对象的集合，其定义一般采用框架(Frame)结构，包括概念的名称、与其他概念之间的关系的集合以及用自然语言对概念的描述。

(2)关系(Relations)：概念之间在领域中的交互作用，可以在形式上定义为 n 维的笛卡尔积的子集 R: $C_1 \times C_2 \times \cdots \times C_n$，如子类关系(Subclass-Of)。在语义上关系对应于对象元组的集合，基本的关系有 Part-of、Kind-of、Instance-of 和 Attribute-of 4 种。

(3)函数(Fnuctions)：一组特殊的关系。在关系中的前 n-1 个元素可以唯一确定第 n 个元素。形式化的定义为 F: $C_1 \times C_2 \times \cdots \times C_{n-1} \rightarrow C_n$。如 Mother-of 就是一个函数，*Mother-of* (x,y)表示 y 是 x 的母亲。

(4)公理(Axloms)：代表永真断言，如概念乙属于概念甲的范围。

(5)实例(Instances)：代表元素。从语义上讲实例表示的就是对象，是某个类在现实世界中的具体反映。

2 Ontology 的理论研究

Ontology 在理论上主要研究如何合理地表示现实世界中的客观概念与抽象知识，包括概念和概念分类、确定概念之间的关系类型以及 Ontology 上的代数等。最值得一提的是 Guarino 等人对本体理论所作出的贡献[3-4]，他们对概念分类做了深入细致的研究，从一般意义上分析了概念的定义、概念的特性、概念之间的关系以及概念的分类，并提出了一套用于指导概念分类的可行理论。基于该理论，他们又提出了 Ontology 驱动的建模方法，在理论上为建模提供了一个通用的模式。

本体的本质是概念模型，表达的是概念及概念之间的关系。长期以来，本体应用的一个常见问题是分类结构不明确，没有一个统一的分类标准或分类理论。不同的应用从各自的角度出发，无限制地使用包含关系对概念进行各种分类，使得概念分类的一致性和合理性难于得到控制。按照 Guarino 的观点，概念之间的差别不仅体现在概念的定义上，同时也体现在概念的某些特性上。从这些特性出发，归纳出概念的元特性(最基本的特性)，从而用公式给出元特性严格的形式定义。在此基础上，又讨论了元特性之间的关系和约束，最终把研究结果作为概念分类的基本理论工具，并提出一套完整的概念分类体系结构[6-7]。

3 Ontology 的实际应用

20 世纪 90 年代，知识表示、信息组织、软件复用等方面的诸多问题对信息科学工作者们提出了种种新的挑战和课题。特别是由于因特网的迅猛发展，如何组织、管理和维护海量信息并为用户提供有效的检索服务成为一项重要而迫切的研究内容。为适应这些要求，Ontology 作为一种能在语义和知识层次上描述信息系统的概念模型建模工具，一经提出便引起了国外众多科研人员的关注，并在计算机的许多领域得到了广泛应用，如知识工程、数字图书馆、软件复用、信息检索、异构信息处理及语义 Web 等。

3.1 Ontology 在图书信息检索中的应用

目前，信息检索技术[5-7]可分为 3 类：全文检索(text retrieval)、数据检索(data retrieval)和知识检索(knowledge retrieval)。全文检索的特点是把用户的查询请求和全文中的每一个词进行比较，不考虑查询请求与文件语义上的匹配，这种方式虽然可以保证查全率，却大大地降低了查准率。数据检索的特点是查询要求和信息系统中的数据都遵循一定的格式，具有一定的结构，允许对特定的字段进行检索。数据检索需要有标识字段的方法。检索性能取决于所使用的标识字段方法和用户对这种方法的理解程度，因此具有很大的局限性。数据检索支持语义匹配的能力也较差。知识检索强调的是基于知识的语义上的匹配，因此在查准率和查全率上有更好的保证。目前知识检索已成为信息检索研究的重点，特别是面向 Web 信息的知识检索。本文研究了基于本体的图书资源查询。

本文建立了一个图书资源的本体图，描述了图书有关的概念和属性，其中定义 4 类资源对象，分别是图书(book)、作者(author)、出版社(press)和编审(editor)。在资源对象的基础上，还定义了 4 种对象属性：对象属性 creat 描述了作者与图书之间的写作关系，其定义域为作者类，值域为图书类；对象属性 has_auther 描述了论文所具有的作者，定义域是图书类，值域为作者；类对象属性 publish 描述图书与出版社之间的出版关系，其定义域为图书类，值域为出版社；类对象属性 has_editor 描述了图书编审，它们描述的是图书中包含的编审，其定义域为图书类，值域为编审类。此外，本体中还定义了各资源对象的数据属性，具体含义分别如表 1、表 2、表 3 所示。

表1 图书的数据属性

Class：Book

Book-category	图书类别
Book_year_moth	出版年月
Book_name	书名
Book_press	出版社
Book_chapterNUM	章节数
Book_Pagination	页数
Book_Price	价格

表2 出版社的数据属性

Class: Press

Press_name	出版社名字
Press_location	出版社地点
Press_class	出版社类型
Press_editor	主编

表3 作者的数据信息

Class: Author	
Auther_name	作者姓名
Auther_ jobs ite	作者工作单位
Auther_research area	作者研究领域

本体的结构根据使用需要设定类和属性，并加上必要的约束，在实用过程中逐渐完善、改进，这是一个长期的工作。根据前面研究的本体知识，本文提出了如图1所示的书信息资源的本体获取模型。

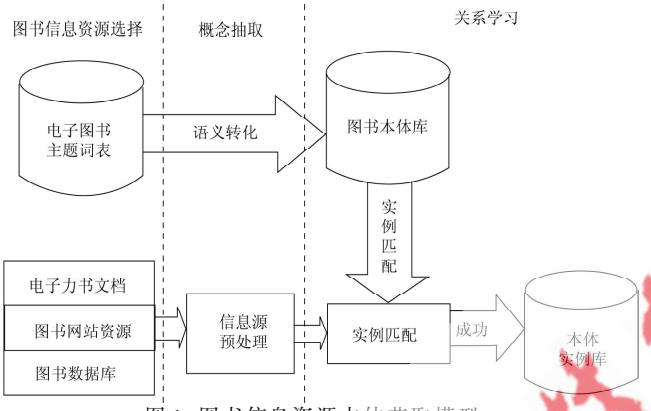

图1 图书信息资源本体获取模型

该模型有图书信息源选择、概念抽取和关系学习阶段，并从原始获取和后天学习两个层面完成图书信息资源本体的构建。

该图书信息检索从传统的关键字层面提高到知识或语义层面上。语义万维网具有良好的概念层次和对逻辑推理的支持，现已被广泛应用于知识表达、知识共享及重用，其中建立图书资源本体的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[8-9]上给出这些词汇术语和词汇之间相互关系的明确定义,从而提高了图书检索的效率和准确性，为用户节省更多的时间。

参考文献

[1] 刘升平,兰煜峰,译.OWL Web本体语言概述推荐标准(中文版) W3CHINA.ORG开发翻译计划(OTP)[EB/OL].[2004-07-3].http://zh.transwiki.org/cn/owloverview.htm.

[2] 刘昕鹏.Ontology理论研究和应用建模——Ontology研究综述、w3c Ontology研究组文档以及Jena编程应用总结[EB/OL].http://bbs.xml.org.cn/viewfile.asp?ID=265.

[3] 李善平,尹奇,胡玉杰,等.本体论研究综述[J].计算机研究与发展,2004,41(7):1041-1052.

[4] 邓志鸿,唐世渭,张铭,等.Ontology研究综述[J].北京大学学报(自然科学版),2002,6(5):34-36.

[5] VELARDI P, MISSIKOFF M, BASILI R. Identification of relevant terms to support the construction of domain ontologies[R]. Proc.of ACL-01 workshop on Human language Technologies, 2001.

[6] 高茂庭,王正欧. Ontology及其应用[J].计算机应用,2003(S2):35-37.

[7] 汪鹏.Ontology知识表示的艺术[J].计算机教育,2004,3(7):45-47.

[8] 杜小勇,李曼,王珊.本体学习研究综述[J].软件学报,2006,6(9):1837-1847.

[9] SMITH M K, WELTY C,MCGUINNESS D L. OWL Web ontology language guide recommendation[EB/OL]. http://www.w3.org/TR/2004/REC-owl-guide-20040210/.

(收稿日期：2009-05-18)