数字图书馆的智能检索技术

廖明光*　王华敏**　廖明宏***

摘　要　信息检索是数字图书馆建设的关键技术之一，将人工智能新技术与信息检索技术相结合，设计一种基于本体论的智能检索工具，它可更准确、全面地满足用户检索信息的目的。

关键词　数字图书馆　信息检索　本体论

An Intelligent Search Technique of Digital Library

Liao Mingguang　Wang Huamin　Liao Minghong

Abstract　Information retrieval is a critical technique of digital library.Combining with the new technique of artifical intelligence,an intellignet search tool based on ontology is designed,which can much exactly and completely satisfied the goal of users.

Keyword　Digital Library　Information Retrieval　Ontology

信息检索技术是数字图书馆的一个关键技术。传统的检索方法主要借助于目录、索引和关键词等方法来实现的，其优点是简单、快捷；但缺点是无法挖掘信息之间的内在联系，检索的结果不能准确、全面地反映用户的需求。在数字图书馆建设中，将人工智能的先进技术引入信息检索中，可以改善传统的检索方法，提高信息查找的命中率。本文在介绍数字图书馆的基础上，重点讨论一种基于本体论（Ontology）的智能信息检索技术。

1 数字图书馆

目前，数字图书馆还没有统一的、严格的定义。从狭义上理解，它是一个数字化了的图书馆。广义而言，一个数字图书馆是计算机可处理信息的集合或此类信息的一个储存处[1]。数字图书馆的重要特征体现在它是建立在计算机网络基础上的，支持信息的远程访问和共享。从这个意义上说，它是一个虚拟图书馆。在 Kaye 的论文[2]中，虚拟图书馆被定义成利用电子网络远程获取信息与知识的一种方式。利用虚拟图书馆，人们不必到图书馆去，只要在办公室或自己家中打开与网络相联接的计算机终端即可获取信息。

数字图书馆的建设涉及到方方面面。有图书馆现有管理体制和馆员素质问题，更多的是一些技术问题。其中包括数字图书馆体系结构的建立；数字图书馆基础设施的建设：如网络的组建方案，网络的智能管理等等；数字化信息的生成、压缩及存储技术：数字化信息的索引与检索技术等。尤其是 Internet 技术在数字图书馆中的应用，人们可以使用 WWW 浏览器（如 Netscape，Internet Explore 等）上搭载的各种检索工具进行文本和超文本信息的检索。但 WWW 上的检索工具通常都是基于目录或关键词的，其查询的结果一方面包含过多的冗余信息，另一方面可能丢失一些重要的有用信息。问题的关键在于“数据”、“信息”和“知识”这三个基本概念的本质区别。

2 数据、信息和知识

事实上，在单纯的表示上，是很难区分数据、信息和知识的，任何东西在计算机中都表示成符号（sign），如 ASCII 字符、位等等，只有通过关系，这些符号才可能进一步区分为数据、信息和知识。借助语言学观点，符号可在语法、语义和语用三维空间加以解释：

（1）符号之间的关系：即“句子”的语法不涉及到符号与现实世界之间的关系，称为第一维空间，这时符号可看成数据。

（2）符号与含义之间的关系：即符号的语义构成符号的第二维空间，只有通过符号与符号含义之间的关系，数据才能变为信息。

（3）符号与符号的“用户”之间的关系，构成符号的第三维空间。在第三维空间引入了用户，他希望执行某些操作，可以称之为模式或符号知识，这个关系称之为语用。它定义了知识的一个重要特征，即只有知识才允许用户执行某些动作或决定。

通常，搜索信息本身并不是目的。相反，当人们在搜索信息时总带着某些目的，希望搜索到的信息能够帮助他达到这些目的。信息加上用户的目的，实际上才构成知识。因此，我们认为，在数据图书馆搜索的是知识，而不是毫不相

* 廖明光　华侨大学图书馆馆员（泉州 362011）
** 王华敏　哈尔滨工业大学外语系技术员（150001）
*** 廖明宏　哈尔滨工业大学计算机系教授（150001）

收稿日期：2001-01-12

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

干的信息，智能搜索工具就是为达到这一目的而设计的。

3 基于本体论的智能检索工具

一个本体论是对概念化的精确描述[3]。它刻划了概念之间的内在联系。传统的基于关键字匹配或基于学科分类的检索工具之所以不能令人满意，最主要的原因之一就是它们无法挖掘概念之间的内在联系，搜索出更深层的信息联系。采用本体论，可以达到这一目的。

可以用一组属性来表示每一条信息或知识项，这些属性能够表示信息的元模型、内容及其研究背景。因此，有三种不同的本体论，见图 1。

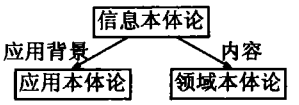

图 1 三种不同类型的本体论

其中，信息本体论用于描述信息元模型，它们是对不同信息源的结构、访问条件和形式化等特性的描述。领域本体论是对信息源的具体内容建模，比如，计算机科学、生物学、航天科学等领域。应用本体论是对信息的具体应用背景的刻划，比如企业本体论刻划了企业组织结构、生产过程等企业模型。

从形式上，一个本体论可以用语义网络来表示。在语义网络中，每结点表示一个概念，而结点之间的连接表示概念之间的关系。在实现上，可以用关系数据库来存放和管理一个本体论。

基于本体论的思想，一个智能检索工具的体系结构如图 2 所示：

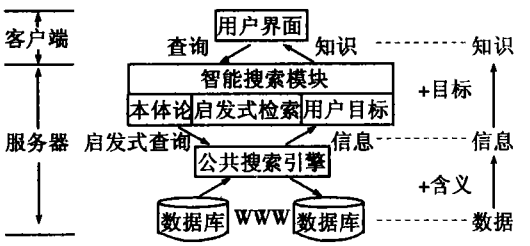

图 2 智能检索工具的体系结构

几点说明：

（1）智能检索工具是建立在客户端—服务器模型的基础上的。其中，在客户端为用户提供输入查询条件和显示查询结果的用户界面，采用 Java 语言实现，它与具体的机器无关，可被 Netscape 或 MS 的 Internet Explore 调用。在服务器一端，是智能检索工具的核心内容，它能够接收来自客户端的查询请求，进行智能搜索将查询结构返回给客户端用户。

（2）从数据流的角度看，智能检索工具提供一种变换，它将数据图书馆中的符号由数据变换成信息，最后到知识并提交给用户。这个变换过程由智能搜索模块和公共搜索引擎两部分组成。对用户提交的查询进行智能化处理，借助本体论，添加启发式信息，并将带有启发式查询的条件转换成公共搜索引擎所能识别的模式提交给公共搜索引擎。公共搜索引擎根据查询条件在数据图书馆中搜索相关内容，满足查询条件的数据以信息的形式提交给智能搜索模块。智能搜索模块根据用户的目标对提交的信息做进一步的过滤处理并保留那些与用户信息相关的信息，并以知识的形式提交给用户。从而达到数据图书馆中的符号由数据到信息，最后到知识的转换。

（3）每个用户在网上检索信息总是带有一定的目的性的。如果能使检索的信息使用户尽快达到目标，这样的信息对用户来说是最重要的。计划识别（Plan Recognition）和用户模型（User Model）可用来表示用户目标。计划识别就是有一组预定义的计划，系统观察用户的动作，并预计用户下一步采取的动作；用户模型用于描述用户与其一特殊应用有关的长期行为，如用户的目的、领域知识和用户的背景等等。通过用户模型，系统可以预见用户的信息需求。在智能搜索引擎中，用户模型用来表示用户目标。

（4）通用搜索引擎很多，如著名的 YAHOO、Excite 等等，这里采用 MetaCrawler[4]作为系统的底层搜索引擎。它是一个并行网页搜索服务引擎。与其它的搜索引擎不同，它不必维护自己的数据库。MetaCrawler 的每一个查询会被送到九个不同的服务：OpenText，Lycos，WebCrawler，InoSeek，Excite，Inktomi，Alta Vista，YAHOO 和 Galaxy。这九个引擎并行工作，并将查询结果返回给 MetaCrawler。选择 MetaCrawler 的好处在于它并行调用九个不同的搜索引擎，最大限度地收集所有信息。

4 结 论

数据图书馆的智能检索技术借助于本体论来挖掘信息之间的内在联系，使得信息检索的结果更能准确、全面地反映用户的目的，是数字图书馆建设中一项关键技术。

参 考 文 献

1 汪冰.数据图书馆：定义、影响和相关问题.北京：中国图书馆学报，1998，(6).

2 Gapen D.Kaye,The Virtual Library:Knowledge,Society,and the Librarian,The Virtual Library,Visions,and Realities.Ed. By Laverna M.Saunders,Westport,CT:Meckler,1992

3 T.Gruber,A Translation Approach to Portable Ontologies, Knowledge Acguisition,Vol.5,No.2,1993

4 NetaCrawler homepage,http:// www.metacrawler.com/

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net