# 基于 Ontology 的智能信息检索研究

张 明[1], 王 煜[1], 杨敬伟[2], 袁 方[1], 赵 红[1], 石 强[1]
(1. 河北大学 数学与计算机学院, 河北 保定 071002; 2. 河北大学 科技处, 河北 保定 071002)

摘 要: 在介绍 Ontology 的基本概念及 Ontology 在信息检索中的应用前提下, 提出了一个基于 Ontology 的智能信息检索设计方案. 利用 Ontology 中规范的概念及概念之间明确的关系描述, 使信息检索过程更加智能化.

关键词: Ontology; 智能信息检索; XML
中图分类号: TP 391.3 文献标识码: A 文章编号: 1000-1565(2005)05-0561-06

随着计算机技术和 Internet 的迅猛发展, 全球信息化时代已经到来, 各类信息急剧增长. 如何在海量信息中寻找到需要的有用的信息, 是信息检索所要研究的问题. 目前常用的检索技术有全文检索(Text retrieval)和数据检索(Data retrieval)[1]. 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, 不考虑查询请求语义上的匹配, 虽然全文检索可以保证查全率, 但是查准率大大降低; 数据检索要求用户查询请求和信息系统中的数据要遵循一定的格式, 具有很大的局限性, 支持语义匹配能力差. 利用现有信息检索技术来进行信息检索, 经常返回大量无关的信息, 使用户大量的时间都花费在排除无关的信息上, 同时又可能丢失重要的信息. 寻求新的、智能化的检索方法也就成为研究热点. 近年来 Ontology 受到研究者的广泛重视, 探讨了 Ontology 在信息检索中的应用, 并提出了一个基于 Ontology 的智能信息检索系统设计方案.

## 1 Ontology

### 1.1 Ontology 的概念

Ontology(本体)的概念最初起源于哲学领域, 它在哲学中的定义为"对世界上客观存在物的系统地描述, 即存在论", 是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本质. 近一二十年来, Ontology 的概念和方法被计算机学科采用, 最早出现在人工智能领域[2], 现在 Ontology 在计算机的许多领域得到了广泛的应用, 如知识工程、软件复用、数字图书馆、Web 上异构信息处理、语义 Web、信息检索等.

越来越多的人研究 Ontology, 并对 Ontology 给出了不同的定义. 1997 年 W. N. Borst 经过对不同定义的深入研究, 认为 Ontology 是"共享概念模型的明确的形式化规范说明"("An ontology is a formal specification of a shared conceptualization")[3]. 这个定义包含 4 层含义[4], 概念模型(conceptualization): 通过抽象出客观世界中一些现象(Phenomenon)的相关概念而得到的模型, 其表示的含义独立于具体的环境状态; 明确(explicit): 所使用的概念及使用这些概念的约束都有明确的定义; 形式化(formal): Ontology 是计算机可读的; 共享(share): Ontology 中体现的是共同认可的知识, 反映的是相关领域中公认的概念集, 它所针对的是团体而不是个体.

尽管定义有很多不同的方式, 但是从内涵上来看, 不同研究者对于 Ontology 的认识是统一的, 都把本体

收稿日期: 2004-11-29
基金项目: 河北省教育厅科研计划资助项目(2004406)
作者简介: 张 明(1977-), 男, 河北唐山人, 河北大学助教, 在读硕士研究生.

当作是领域(可以是特定领域的,也可以是更广的范围)内部不同主体(人、机器、软件系统等)之间进行交流(对话、互操作、共享等)的一种语义基础.即由 Ontology 提供一种明确定义的共识,该共识更主要的是为机器服务,因为目前的计算机只能把文本看成字符串进行处理,并不能像人类一样理解自然语言中表达的语义.简单的说,Ontology 就是对客观存在的概念和概念之间关系的描述.

1.2 Ontology 的建模原语

Perez 等人用分类法组织了 Ontology,归纳出 5 个基本的建模元语(Modeling Primitives)[5]:类(classes)或概念(concepts)、关系(relations)、函数(functions)、公理(axioms)和实例(instances).

概念指任何事务,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;关系是在领域中概念之间的交互作用,形式上定义为 $n$ 维笛卡儿积的子集:$R: C_1 \times C_2 \times \ldots \times C_n$,如子类关系(subclass- of),在语义上关系对应于对象元组的集合;函数是一类特殊的关系,该关系的前 $n-1$ 个元素可以唯一决定第 $n$ 个元素,形式化的定义为 $F: C_1 \times C_2 \times \ldots \times C_{n-1} \rightarrow C_n$.如 Friend- of 就是一个函数,friend- of$(x, y)$ 表示 $y$ 是 $x$ 的朋友;公理代表永真断言;实例代表元素,从语义上讲实例表示的就是对象.

另外,从语义上讲,基本的关系共有 4 种:part- of,表达概念之间部分与整体的关系.kind- of,表达概念之间的继承关系,类似于面向对象中的父类与子类之间的关系.instance- of,表达概念的实例与概念之间的关系,类似于面向对象中的对象和类之间的关系.attribute- of,表达某个概念是另一个概念的属性,如"颜色"是汽车的一个属性.

在实际建立 Ontology 过程中,概念之间的关系不限于上面列出的 4 种基本关系,可以根据领域的具体情况定义相应的关系.

1.3 用于智能信息检索的 Ontology

根据相应的 Ontology 构建方法[6],为智能论文信息检索系统构建 Ontology.用于智能论文信息检索的 Ontology 由数据库领域词汇 Ontology 和论文 Ontology 组成,其中数据库领域词汇 Ontology 的片段如图 1 所示,论文 Ontology 的片段如图 2 所示.

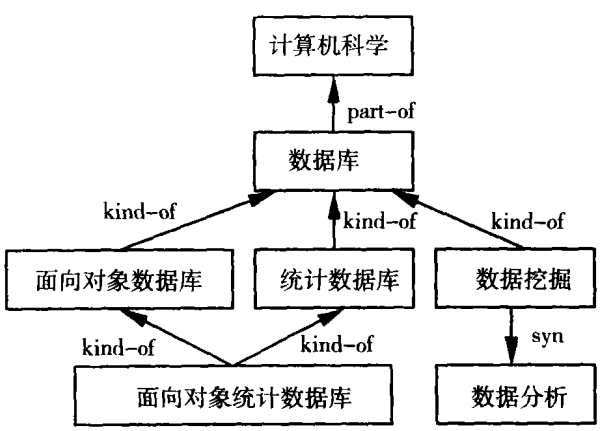


图 1 数据库领域词汇 Ontology 片段

**Fig. 1 Part of words of database field Ontology**

数据库领域词汇 Ontology 中包括数据库领域中的词汇(概念)和它们之间的关系,如统计数据库和数据库是 kind- of(继承)关系,数据挖掘和数据分析是 syn(同义词)关系.

论文 Ontology 中包括论文、标题、作者、单位、关键词的概念和标题、作者、单位、关键词各自与论文之间

的 part- of( 部分与整体) 关系, 作者和标题之间的 write( 写作) 关系, 作者和单位之间的 blong to( 隶属) 关系和标题、作者、单位、关键词的所有实例.

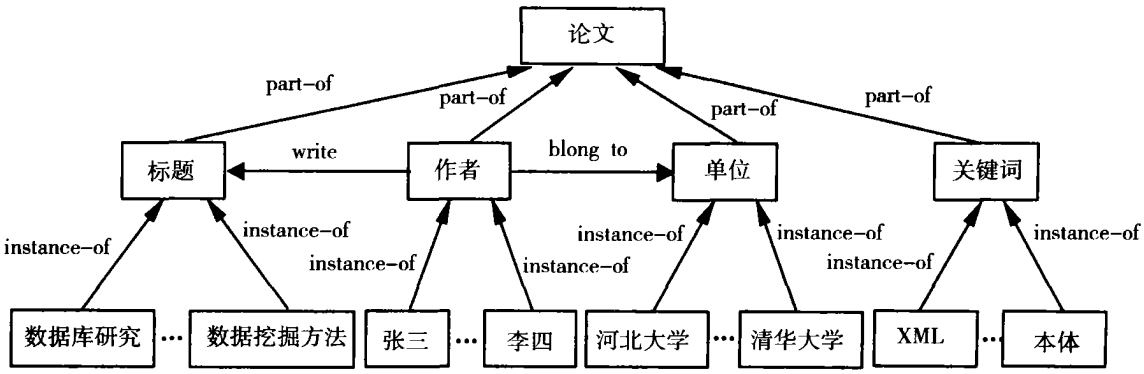


图 2　论文 Ontology 片段

**Fig. 2　Part of paper Ontology**

## 2　智能信息检索

### 2.1　理想的智能信息检索

理想的智能信息检索应该达到如下目标[7]: 提供友好的用户检索交互界面; 基于自然语言或实例的查询; 依据用户浏览和检索的习惯信息, 熟悉用户的兴趣爱好, 建立一定的用户描述, 主动向用户提供相关的信息; 针对用户查询请求自动向用户提供相关文档页面, 不需用户重复发现知识; 综合利用个性化检索和集中浏览的优势; 检索速度快, 能够快速地返回查询结果; 高查全率和高查准率. 即语义检索、个性化服务.

### 2.2　Ontology 在智能信息检索中的应用

由理想的智能信息检索达到的目标可知, 要实现智能信息检索的前提是对数据所包含内容的充分理解. 由于自然语言的灵活性和人们看待事务的角度不同会导致对同一概念的不同表达形式, 即使用不同的词汇表达相同或相近的概念, 例如计算机可以称为电脑、个人电脑、微机、微型计算机、PC 等. 这对于信息检索的查全率和查准率都会有很大的影响. 一般情况下, 用户所提出的查询请求是一个简单的词或者词组, 当系统接受到该查询请求后, 需要首先对其进行语义化处理. 通常处理考虑如下 3 种情况:

1) 同义词关系( synonym) : 词与词之间的意思相同或非常相近, 往往可以相互替换, 如计算机和电脑.

2) 上下位关系( hypernymy/ hyponymy) : 下位词是上位词的特例, 如动物和狮子、老虎、大象之间的关系. 在检索中有的时候通过该概念的上下位概念也能检索到潜在的有用信息.

3) 概念的歧义: 一词多义的现象. 例如: 笔记本即可以指笔记本电脑, 又可以指平常写字的笔记本. 为了排除这些歧义干扰, 应该将这些概念按主题分类.

其次, 如果用户提出的查询请求是词组的话, 还要考虑词组中各词之间的语义信息. 例如: 用户输入的是“张三　数据挖掘”, 则可以推测用户是想要查询“张三所写的关于数据挖掘方面的文章”, 使查询具有语义信息.

Ontology 具有良好的概念层次结构和对逻辑推理的支持[1, 8], 它提出了对特定领域知识的共同理解, 抽象出该领域内共同认可的词( 概念) , 并给出这次词( 概念) 及它们之间相互关系的明确定义. 基于 Ontology 的智能信息检索优于关键词搜索, 因为 Ontology 包含机器可以判断的概念的定义, 从而使系统对领域内的概念、概念之间的联系及领域内的基本公理知识有一个统一的认识, 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, 从而提高了信息检索系统的查全率和查准率.

2.3 基于 Ontology 的智能信息检索的基本设计思想

基于 Ontology 的智能信息检索的基本设计思想如下[9]:

1) 在领域专家的帮助下, 建立相关领域的 Ontology;

2) 收集信息源中的数据, 并参照已建立的 Ontology 把收集来的数据按规定格式存储在元数据库(RDB, KDB 等) 中;

3) 对用户检索界面获取的查询请求, 查询转换器按照 Ontology 把查询请求转换成规定的格式, 在 Ontology 的帮助下从元数据库中匹配出符合条件的数据集合;

4) 检索的结果经过定制处理返回给用户.

如果智能信息检索系统不需要太强的推理能力, Ontology 可用概念图的形式表示并存储, 数据可以保存在一般的关系数据库中, 采用图的匹配技术来完成信息检索. 如果需要比较强的推理能力, 一般需要用一种 Ontology 描述语言(OWL, Loom 等) 表示 Ontology, 数据保存在知识库中, 采用描述语言的逻辑推理能力来完成信息检索.

## 3 基于 Ontology 的智能论文信息检索系统

按照智能信息检索的基本设计思想, 构建了一个基于 Ontology 的智能论文信息检索系统, 该系统主要用于对数据库领域论文的检索.

3.1 论文数据存储

论文数据以 XML 文档形式存储. XML 文档是一种形式良好的半结构化文档, 由于文档本身有一定的结构性并且可以自行设计有意义的标记, 所以便于进行信息检索. DTD(Document Type Definition) 作为 XML 文档的句法规范模型, 定义了 XML 数据的结构, 说明了数据是如何组织的[10].

一篇论文主要由标题、作者、作者单位、论文摘要、关键词和论文正文组成, 并且标题、论文摘要、正文只能有一个, 而作者、作者单位、关键词可能有一个或多个. 因此论文结构的 DTD 如下:

```
<! DOCTYPE 论文[
<! ELEMENT 论文(标题, 作者+ , 单位+ , 摘要, 关键词+ , 正文)
<! ELEMENT 标题(# PCDATA) >
<! ELEMENT 作者(# PCDATA) >
<! ELEMENT 单位(# PCDATA) >
<! ELEMENT 摘要(# PCDATA) >
<! ELEMENT 关键词(# PCDATA) >
<! ELEMENT 正文(# PCDATA) >
]>
```

3.2 智能论文信息检索系统逻辑结构

智能论文信息检索系统逻辑结构如图 3 所示.

整个检索过程分以下几步:

1) 用户先在检索界面输入检索词或词组, 然后检索词(组) 被传递给语义化处理模块. 语义化处理模块参照数据库领域词汇 Ontology 进行检索词(组) 语义化处理, 包括找出各检索词的上位词、下位词、同义词.

2) 经语义化处理的检索词(组) 传递给词性确定模块, 词性确定模块参照论文 Ontology 确定各检索词的词性是标题、作者、单位还是关键词, 同时确定各检索词之间的关系, 把结果传递给查询转换模块.

3) 查询转换模块按照所得信息对本次查询进行查询转换, 使查询具有相应的语义信息, 然后在论文数据库中进行查询. 对不能确定语义信息的查询按照关键词匹配技术进行查询.

4) 查询所得的结果经定制处理模块处理, 按照检索词原词查询结果, 检索词同义词查询结果, 检索词上位词查询结果、检索词下位词查询结果进行排序, 然后由显示界面显示查询结果.

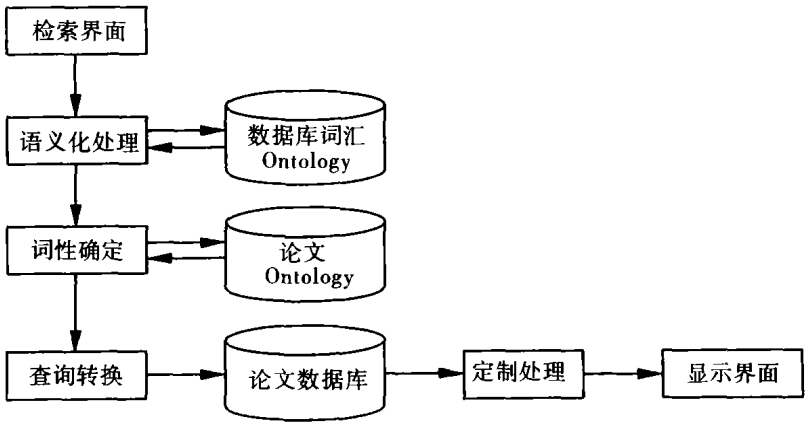


图 3　智能论文信息检索系统逻辑结构

**Fig. 3　Logical structure of intelligent paper information retrieval system**

在检索过程中, 如果用户输入数据库领域词汇 Ontology 中没有的概念, 系统将记录该概念. 如果该概念被检索的次数达到一个预先设定的值, 系统将提示管理员根据领域专家的意见将该概念扩充到数据库领域词汇 Ontology 中.

## 4　结束语

Ontology 在信息检索特别是智能信息检索中的地位越来越重要, 基于 Ontology 的智能信息检索技术也将会逐步完善. 本文在讨论 Ontology 和智能信息检索之后, 提出了一种基于 Ontology 的智能论文信息检索系统设计方案. 下一步的研究工作是具体实现笔者设计方案并完善 Ontology.

参　考　文　献:

[1] GUARINO N, MASOLO C, VETERE G. OntoSeek: content_based access to the web[J]. IEEE Intelligent System, 1999, 14(3): 70– 80.

[2] NECHES R, FIKES R E, GURBER T R, et al. Enabling technology for knowledge sharing[J]. AI Magazine, 1999, 12(3): 36– 56.

[3] BORST W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse[D]. Enschede: University of Twente, 1997.

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

[5] PEREZ A G, BENJAMINS V R. Overview of knowledge sharing and reuse components: Ontologies and Problem_Solving Methods[Z]. Stockholm, Sweden, Proceedings of the IJCAI_99 workshop on Ontologies and Problem_solving Methods(KRR5), Stockholm, Sweden, 1999.

[6] 杨秋芬, 陈跃新. Ontology 方法学综述[J]. 计算机应用研究, 2002(4): 5– 7.

[7] 许　珏. 体论与信息检索[J]. 中国信息导报, 2004(3): 57– 58.

[8] SHUN S B, MOTTA E, DOMINGUE J. Schol Onto: an ontology_based digital library server for research documents and discourse[J]. Intl J Digital Libraries, 2000, 3(3): 237– 248.

[9] 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, 38(5): 730– 738.

[10] 王海波, 姜吉发, 耿　晖. XML 搜索引擎研究[J]. 计算机应用研究, 2001, 18(4): 68– 71.

# Intelligent Information Retrieval Using Ontology

ZHANG Ming[1], WANG Yu[1], YANG Jing_wei[2], YUAN Fang[1], ZHAO Hong[1], SHI Qiang[1]
(1. College of Mathematics and Computer, Hebei University, Baoding 071002, China;
2. Secience and Technology Department, Hebei University, Baoding 071002, China)

**Abstract:** After introducing the basic concept of Ontology and the application of Ontology in intelligent information retrieval, we propose a schema of intelligent information retrieval using Ontology. Taking the full advantage of Ontology, the procedure of information retrieval is more intelligent.

**Key words:** Ontology; intelligent information retrieval; XML

(责任编辑: 孟素兰)

(上接第 553 页)

# Web Services Request and Response Pattern

MA Li[1], SHI Lei[2], SHI Qiang[2], YANG Jing_wei[3], ZHAO Hong[2]
(1. Hebei University Press, Baoding 071002, China;
2. College of Mathematics and Computer, Hebei University, Baoding 071002, China;
3. Science & Technology Department, Hebei University, Baoding 071002, China)

**Abstract:** Web Services has resolved a very difficult problem for interoperability between disparate systems. As a key technology of Web Services, it has defined Web Services by abstract language, and realized them by concrete data formats and protocols, but through WSDL definition , the Web Services abstract rank is very low and it's description to Web Services is not clear. So, the obstacle of interoperability between computers and applications in semantic still exist. In this paper, a pattern of Web Services request and response based on semantic is proposed, at the same time, it's working model is provided. This approach has resolved this problem through improving the exchange content between service provider and requester.

**Key words:** Web services; WSDL, XML; semantic

(责任编辑: 孟素兰)