●徐国虎 董 慧

基于语义的数字图书馆推理检索研究*

摘 要 传统数字图书馆检索技术处理语义能力弱。为此,应运用本体的思想和方法组织数字图书馆资源,在资源内容描述语义形式化中引入规则推理技术,把检索从基于关键词匹配层提高到基于语义的知识推理层面。图 1。参考文献 9。

关键词 数字图书馆 语义检索 推理 本体
分类号 G250.76

ABSTRACT Because traditional digital library techniques are weak in processing semantics, we propose to use the conceptions and methods of ontology to organize digital library resources, introduce rule inference techniques and raise searches from the keyword match level to semantically-based knowledge inference level. 1 fig. 9 refs.

KEY WORDS Digital library. Semantic search. Inference. Ontology.
CLASS NUMBER G250.76

目前数字图书馆的信息资源检索技术主要有 Web 数据库检索技术、搜索引擎检索技术和元数据检索技术。它们主要借助于目录、索引、关键词和各种元数据等方法来实现,或者要求了解检索对象数据结构或元数据格式,基于语法结构进行检索,或者不能处理复杂语义关系,常常检索出大量相关度很差的文献。也就是说,它们存在着 3 个深层次的问题:忠实表达的问题,表达差异的问题,词汇孤岛的问题。

这 3 个问题的本质在于资源概念的语义复杂性以及传统的信息检索技术缺乏知识语义理解能力和处理能力。而实际上,人们在检索时经常面对复杂的难以事先了解的资源类型和内容结构,但又希望能针对复杂概念进行准确的语义检索,这就要求检索系统能解析数字图书馆各种内容结构的信息资源及其相互之间的复杂语义关系,要求把数字图书馆信息检索技术从基于关键词匹配层面提高到基于语义的知识推理层面。

1 基于本体的数字图书馆资源描述模式

要让检索系统能理解数字图书馆信息资源中丰富的逻辑语义并进行推理检索,我们必须采用一定的知识体系(比如本体描述框架)来表达信息资源对象及其相互间的逻辑语义关系,运用一定的词汇体系(包括同义词、上位词、下位词等)来描述这些对象类及其关系;并建立数字图书馆元数据元素与信息资源语义描述的映射关系;分析数字图书馆领域资源复杂的语义关系来确立推理的规则,对利用元数据(或目录)和语义描述框架描述的信息资源进行搜索和推理[1]。如果上述这些机制都能以计算机可理解、可处理的方式建立起来,就能实现数字图书馆环境下基于语义的检索和推理。

1.1 数字图书馆资源描述的形式化

运用本体技术来组织和描述数字图书馆所有信息资源,其描述模式包括以下 4 个部分。

(1)资源对象类层级体系(Resource hierarchical system,RHS)。在数字图书馆诸多的信息资源之间存在 is-a、kind-of、part-of 等关系,通过这些层级关系从而构成整个领域的对象类体系。父子类之间往往存在继承关系。类别间还可能存在组合关系,例如某类是其他若干类的交集或并集,或是另一类的等同集、反集或补集。比如一个关于 XML 技术资源的数字图书馆中,XML Schema 类资源、XSL 类资源、XSLT 类资源、XPath 类资源、XLink 类资源、XQuery 类资源与 XML 资源就存在 Part-of 关系,而 SGML 类资源又与 XML 类资源存在父子继承关系。资源对象类的层级体系决定了信息资源实例的语义层次关系。

(2)对象类的属性层级体系(Property hierarchical system,PHS)。这些属性由数字图书馆中各种对象,比如信息资源对象、资源创作者对象、资源存放对象、资源格式对象、资源所有权人对象、资源出版对象等等之间的相互关系决定。PHS 主要是参考元数据中关于知

* 本文系国家自然科学基金资助项目(批准号 70373047)和教育部基地重大课题(批准号 05JJD870004)研究成果之一。

识产权的描述元素和外部属性的描述元素以及它们之间的逻辑关系。其中的属性可按照层级关系继承，有取值对象类和实际值范围、取值基数的限制，并有关于交换性、对称性、唯一值性、可传递性、有序性等的规定。某些属性还可以是其他属性的子属性。

(3)资源主题语义描述体系(Semantic Relation System, SRS)。它主要对资源的内容进行描述，并揭示不同资源描述主题词之间的语义关系。它包括数字图书馆资源所涉及领域的所有主题术语，在一定基础上，可以直接在都柏林元数据、MARC 元数据关于资源内容描述元素与 SRS 中术语建立映射，借鉴过来。

(4)推理规则体系(Inference Rule System, IRS)。对资源的语义检索，必然涉及推理规则。推理规则体系主要是建立在数字图书馆资源描述的层级体系、属性体系和语义关系体系的基础之上，主要包括属性继承规则，父子资源对象类(属性)对象类传递规则、资源对象类组合规则、逻辑关系推理规则等。这些规则往往用一阶或高阶谓词逻辑等形式表示，并在检索过程中被集成到本体推理引擎(如 RACER, PELLET, JESS)中用于语义推理[2]。

具体而言，数字图书馆中的一个资源 R，我们可以定义为一个四元组：R(ID, Schema, Presentations, Semantics)[3]。

其中 ID 表示资源 R 的唯一标识符，Schema 表示资源 R 的模式，Presentations 用于描述资源的载体形式，Semantics $\subseteq$ *SRS* 用于描述资源的具体语义。

Schema 为一个三元组 Schema(Incs, Attributes, Associations)。其中 Incs $\subseteq$ *RHS*，用于揭示两个资源实例之间的层级关系，Attributes $\subseteq$ *PHS* 是资源实例所具有的属性集。

Associations $\subseteq$ (Associations_name, Entity_name1, Entity_name2, Cardinality 1, Cardinality2) 表示资源与其他实体(比如出版社，作者，版权人等等)之间的关系。Cardinality 1, Cardinality2 分别表示在 Associations 中对资源和其他实体的约束。

1.2　信息资源之间的关系

数字图书馆资源众多，表现形式多样，相关主题的资源和资源之间存在着复杂的语义逻辑关系。要从语义的角度提高数字图书馆资源检索的查全率与查准率，就必须探讨资源之间的关系。

假设数字图书馆中存在着两个资源 R_1 (ID_1, $Schema_1$, $Presentations_1$, $Semantics_1$) 和 R_2 (ID_2, $Schema_2$, $Presentations_2$, $Semantics_2$)，则这两个资源之间可能存在如下的关系[4]。

(1)同一关系，Identical(R_1, R_2)。如果 $Schema_1 = Schema_2$，并且 $Semantics_1 = Semantics_2$；同一关系表示两个资源在外在载体形式与内在语义内容方面一致，是同一资源，同一关系具有对称性和传递性。

(2)同涵关系，Synonymous(R_1, R_2)。如果 $Schema_1 \neq Schema_2$，并且 $Semantics_1 = Semantics_2$；同涵关系表示两个资源虽然外在模式不同，但描述的都是同一主题的内容，同涵关系具有对称性和传递性。

(3)包含关系，InheritFrom(R_1, R_2)。如果($Schema_1$, $Schema_2$) $\in Incs_2$；包含关系表示两个资源描述的主题内容在语义上存在层级关系，也就是说资源 R_1 描述的内容是 R_2 语义主题的一部分，包含关系存在传递特性。

(4)参照关系，Reference(R_1, R_2)。如果 $Schema_1 \neq Schema_2$，并且 $Semantics_1 \neq Semantics_2$，但 $Semantics_1 \cap Semantics_2 \neq \emptyset$；参照关系表示两个资源虽然描述的内容不是同一领域的，也不存在对象类实例父子关系，但两者描述的内容互有参考价值。参照关系具有非严格的传递性，其传递性的存在有一定的语义范围限制。

1.3　信息资源检索推理规则

探讨了资源之间存在的这 4 种关系，我们就可以根据资源之间的语义关系，来确立资源检索时的推理规则，从而扩展检索范围，提高检索精度。资源检索推理时，涉及的规则除了父子对象类的传递规则、属性继承规则之外，考虑资源描述具体语义的关联性以及事务的外在逻辑，推理规则将更多更复杂[5]。一般情况下，如果只考虑父子对象类以及资源语义的简单关系的话，推理主要包括以下几条基本规则。

(1)资源对象父子类规则：If X subClassOf Y and Y subClassOf Z, Then X subClassOf Z。其形式化描述(采用 JENA2 推理引擎所支持的规则格式)为[6]：[classTransitiveRule:(? x rdfs:subClassOf ? y),(? y rdfs:subClassOf ? z) -> (? x rdfs:subClassOf ? z)]。

(2)资源对象父子类实例规则：If X subClassOf Y and individual R ISInstanceOf X, Then R ISInstanceOf Y。其形式化描述为[instance TRule:(? x rdfs:subClassOf ? y)(? r rdf:type ? x) -> (? r rdf:type ? y)]。

(3)资源同一关系规则：If R_1 IdenticalOf R_2 and R_2 IdenticalOf R_3 Then R_1 IdenticalOf R_3。其形式化描述为[IdenticalTRule:(? x rdfs:IdenticalOf ? y)(? y rdf:IdenticalOf ? z) -> (? x rdf:IdenticalOf ? z)]。

(4)资源同涵关系规则：If R_1 SynonymousOf R_2

and R_2 SynonymousOf R_3 Then R_1 SynonymousOf R_3。其形式化描述为[SynonymousTRule:(? x rdfs:SynonymousOf ? y)(? y rdf:SynonymousOf ? z) -> (? x rdf:IdenticalOf ? z]。

(5)资源包含关系规则:If R_1 InheritFrom R_2 and R_2 InheritFrom R_3 Then R_1 InheritFrom R_3。其形式化描述为[InheritFromTRule:(? x rdfs:InheritFrom ? y)(? y rdf:InheritFrom ? z) -> (? x rdf:InheritFrom ? z)]。

以上探讨的规则主要基于资源对象类的父子层次以及资源的3种语义关系的传递性,规则简单,涉及的资源语义也简单易懂。而在推理中,考虑到数字图书馆资源描述的具体领域以及资源描述的外部属性和版权属性,推理规则将更加复杂,检索将更深入语义逻辑层面,检索精度将相应提高。下面将以一个检索"熟悉XML的专家"的例子来探讨。

2 基于语义的数字图书馆推理检索体系

2.1 推理检索框架

实际应用中,我们需要一系列相互关联的工具、系统和资源来利用体现在信息资源描述本体和其他元数据中的语义信息进行智能检索和推理,它们所构成的检索框架可由图1所示[7~9]。在这个框架中,本体构建工具利用有关本体描述语言来标注数字图书馆的资源,并将标注所形成的元数据存入元数据库,检索系统调用推理引擎利用本体术语、元数据和推理规则进行语义推理,实现语义检索和其他智能处理;领域资源分析工具主要在元数据涉及多个领域资源描述本体时,辅助推理引擎分析推断概念归属的领域本体和处理规则,支持多领域本体引用和复用。

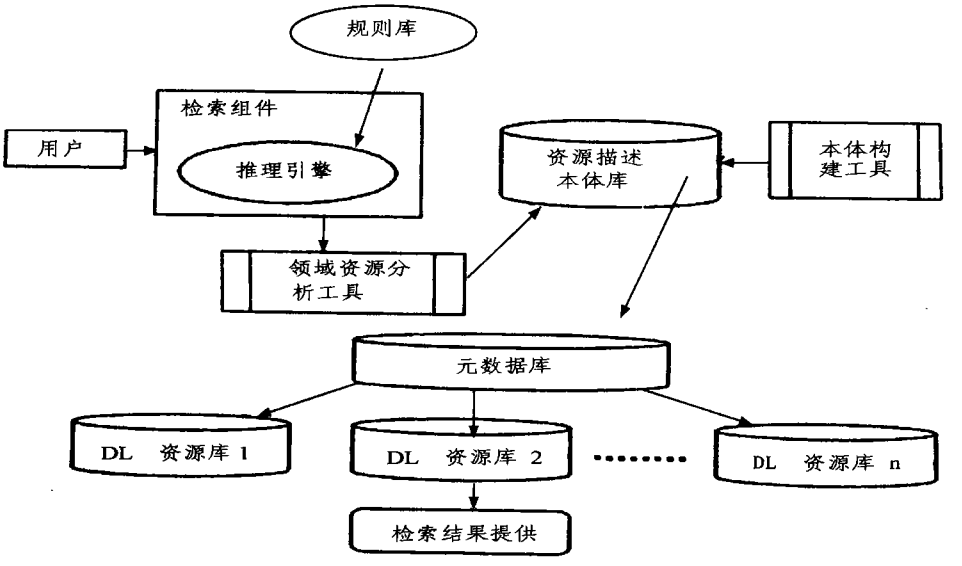

图1 基于语义的数字图书馆推理检索框架

2.2 检索示例

我们欲检索"熟悉XML的专家",这个看似简单的检索语句实际上涉及复杂的逻辑概念、语义和语法关系,如:

(1)XML指扩展标记语言,属于数据标记语言,可看成SGML的简化版本,本身又包括XML Schema, XSL, XSLT, XPath, XLink, XQuery等标准和技术。

(2)"专家"指具备特定专业知识的人,例如写过相关文献、讲授过相关课程、或参与相关研究项目的人员,甚至可是参加过有关课程或会议的其他人员(权威性较低而已)。

(3)"文献"可能是标准、著作、文章、报告、网页、辅导材料等。"课程"可能包括正式课程、讲座、报告会、培训辅导等。"研究项目"可能是立项课题或者自拟的课题、系统设计、实验、理论分析、文献综述活动等。"人员"则可能是具各种职衔甚至没有特定职衔的人。

(4)这些信息可能存在于关于描述XML技术文献、人员机构、会议活动、研究项目、经费计划、专门公告、电子邮件等资源里。这些资源可能具有不同内容结构、不同标记元素和元数据元素、不同自然语言,例如作者可能表示为Author、Contributor、Creator、Editor、

作者、编者、编撰者及其他名称。

假设在数字图书馆中有一个资源描述的内容是：王志东曾主持一个国家自然科学基金项目“基于 XSLT 的本体推理规则转换研究”，该项目重要参与者有杨兵，并且杨兵主持编著过一本《SGML 入门与提高》的书籍，该书关于 XSL 的一章主要由李东编写。在传统基于关键字匹配的检索方式中，该资源将被漏检，我们就发现不了王志东、杨兵和李东都是熟悉 XML 技术的专家。而基于语义关系，检索系统应能自动理解资源描述的内容结构和内容元素，能理解或推理这些内容元素之间的逻辑语义关系，从而将以上三人都检索出来。

语义推理利用文献的语义标注和本体语义关系及推理规则进行推理，从而实现智能检索和知识组织。对于该检索示例，在相应的文献资源描述本体、科研概念描述本体、XML 技术体系支持下，利用下列对象类层次定义和父子类属性继承特点以及逻辑推理规则就可实现智能检索，就可以发现上述资源中出现的王志东、杨兵和李东三人都是熟悉 XML 技术的专家。

section Subclass-Of chapter Subclass-Of book Subclass-Of publication Subclass-Of document

//文献中节、章、书、出版物、文献之间的逻辑层次关系

document Has-Subject(x,y)

Written-By(x,y)

author EQUIVALENT - TO contributor EQUIVALENT-TO“编者”QUIVALENT-TO——

//文献创作者不同称谓之间的等价传递关系

project Subclass-of research Subclass-Of work

//科研活动中项目、研究、工作之间的逻辑层次关系

Xpath/XSLT/XSL Included-By XML Included-By SGML

//通用标记语言中，Xpath/XSLT/XSL、XML 与 SGML 之间的语义包含关系

work Has-Subject(x,y)

Participated-By(x,y)

(Subclass-Of(x,y) and Has-Property(y,z)) => Has-Property(x,z)

//子类属性继承规则

(has-subject(x,y) and Included-By(y,z)) => Expand-subject(x,z)

//资源(部分)主题泛化规则

(has-subject(x,y) and Included-By(z,y)) => Specialization-subject(x,z)

//资源(部分)主题具体化规则

(Written-by(y,x) and has-subject(y,z)) => has-knowledge(x,z)

//资源创作者知识技能拥有规则

(has-knowledge(x,z) and Included-By(y,z)) => has-knowledge(x,y)

//知识技能具体化规则

(participated-by(y,x) and has-subject(y,z)) => has-knowledge(x,z)

//科研参与者知识技能拥有规则

(collaborated-with(x,y) and has-knowledge(y,z)) => has-knowledge(x,z)

//合作者知识技能拥有规则

在此基础上，未来成熟的基于语义的数字图书馆检索技术应当可以自动抽取资源文件中的相关具体内容，或者根据特定内容将相关资源或其中内容组织到一定知识体系中，或者合并资源内容来建立新的资源文件，并按用户要求的格式予以提供，从而真正实现对知识的检索和操作。

参考文献

1 张晓林. Semantic Web 与基于语义的网络信息检索. 情报学报，2002(4)

2 Elena Paslaru Bontas. Reasoning Paradigms for SWRL-enabled Ontologies. Springer，2004

3 McCray，Gallagher. Extending the role of metadata in a digital library system. In Proceedings of the IEEE research and technology advances in digital libraries，1999

4 Su-Hsien Huang. Enhancing semantic digital library query using a content and service inference model. Elsevier，2005

5 F. Bry，T. Furche. Data Retrieval and Evolution on the SemanticWeb：A Deductive Approach. In Workshop on Principles and Practice of Semantic Web Reasoning. Springer，2004

6 Jena2-A Semantic Web Framework. http://www. hpl. hp. com/semweb/jena2. htm[2005-07-25 查询]

7 S. Decker. An Information Food China for Advanced Applications on the WWW. http://citeseer. nj. nec. com/decker00information. html[2005-07-26 查询]

8 Onto Broker Home Page. http://ontobroker. aifb. Uni-karlsruhe. de[2005-07-26 查询]

9 M. Erdmann，R. Studer. Ontologies as Conceptual Models for XML Documents. http://wern. ucalgary. ca/KSI/KAW/KAW99/papers/ErdmannlPerdmann. pdf[2005-07-28 查询]

徐国虎　武汉大学信息管理学院博士生。通信地址：武汉大学。邮编 430072。

董　慧　武汉大学信息管理学院教授，博士生导师。通信地址同上。

（来稿时间：2005－10－26）