

基于刻面与本体的资源描述与检索系统的设计与实现

张海龙¹ 彭鑫¹ 赵文耘¹ 肖君²

¹(复旦大学计算机科学与工程系 上海 200433)

²(上海远程教育集团 上海 200086)

摘 要 给出一种基于本体与刻面的资源描述与检索系统的设计与实现。充分发挥基于刻面的描述与检索方案准确、高效的优点,在此基础上引入本体,利用本体知识中蕴含的丰富关系,弥补基于刻面的资源描述与检索方案中术语、刻面之间关系匮乏的缺点,大大提高检索的查准率与查全率。介绍了从本体的设计、编码、入库到基于本体的检索条件预处理,最后检索引擎执行检索的设计与实现。

关键词 刻面 本体 资源检索 推理

DESIGN AND IMPLEMENTATION OF A RESOURCE REPRESENTATION AND RETRIEVAL SYSTEM BASED ON ONTOLOGY AND FACET

Zhang Hailong¹ Peng Xin¹ Zhao Wenyun¹ Xiao Jun²

¹(Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China)

²(Shanghai Distance Education Group, Shanghai 200086, China)

Abstract The design and realization of a resource description and retrieval system based on ontology and facet is presented. This system exploits the advantages of accuracy and efficiency of resource description and retrieval system, and on this ground, ontology is introduced to cover the shortage of relations deficiency between terms and facets. Thus, the retrieval accuracy is improved. The design, coding and loading of ontology, the retrieval qualifications preprocessing, and the retrieval procedure are described.

Keywords Facet Ontology Resource retrieval Inference

0 引言

在资源的描述与检索领域,众多行之有效的方法被提出^[1,2]。其中基于刻面的描述与检索技术被广泛应用^[3,4]。该方案从资源的不同方面对其进行描述,有利于提高检索的准确率;同时,通过对查询结果匹配度阈值的控制增加查全率。这种检索技术的不足之处在于只能体现出刻面和刻面、刻面和术语以及术语与术语之间的组成关系,忽略了它们之间更为丰富的关系,而这些关系有助于系统对查询条件的理解,从而进一步提高查准率和查全率。

本体通过描述事物以及事物之间的关系来描述现实世界的语义。本体可以作为知识共享和知识发现的工具^[5]:一方面它可以作为不同系统之间的沟通媒介,另一方面通过已有的本体描述可以推知更多的知识。本体的这些优势可以弥补基于刻面的描述与检索技术的不足。通过建立刻面、术语的本体,可以极大丰富刻面和术语、术语和术语之间的关系,有利于发现它们之间的内在关系;同时这个本体可以作为检索条件到检索引擎之间的桥梁,通过它,更好地挖掘用户的检索意图,使得检索引擎检索出更接近用户意图的结果。基于这种思想,我们把本体与刻面描述检索方案相结合,构造基于本体和刻面的资源描述与

检索方案。

Linux多媒体网络教学资源管理和应用平台软件研究课题系上海远程教育集团承建的 863 软件重大专项课题^[6],在该课题的研究过程中,设计并实现了教育领域内基于本体的刻面描述与检索系统。

1 基于刻面的描述和检索技术与本体

1.1 基于刻面的资源描述与检索

在基于刻面的描述与检索系统对资源的描述是用术语按刻面进行的。一个刻面代表资源的一个特定方面,它们从不同侧面描述资源,不同刻面之间是一种正交关系。术语分属于不同刻面,用于在该刻面下描述资源^[4]。刻面与术语共同构成一棵刻面树,资源的描述和检索都是按照这棵树进行组织的。检索时把检索条件按照刻面树组织起来形成查询树,用这棵查询树和资源描述树进行匹配,计算出匹配度决定该资源是否为所需资源。基于刻面的描述与检索优点在于查准率、查全率较高。

收稿日期:2005-09-09。基金项目:国家 863 计划(2004AA1Z2330, 2004AA113050),国家自然科学基金(60473061),上海市科委科研攻关项目(04DZ15022)。张海龙,硕士,主研领域:软件工程。

1.2 领域本体

本体最早出现在哲学领域,是一个存在的系统说明,是概念及概念之间关系的模型,它通过概念之间的关系来描述概念的语义。W3C在本体语言标准化方面做了大量工作,推出了语义网本体描述语言:OWL。它又分为三种表述能力渐强的子语言:OWL Lite、OWL DL、OWL Full。Lite是三种子语言中表述能力最弱的一种,只需要一个分类层次和简单的属性约束的用户;DL支持那些需要在推理系统上进行最大程度表达的用户,这里的推理系统能够保证计算完全性和可决定性;Full支持那些需要在没有计算保证的语法自由的RDF上进行最大程度表达的用户。

1.3 基于本体与刻面的资源描述与检索方法

本系统在资源描述时完全采用立面描述方案:为每个入库资源指定在一些立面下的术语,对它进行描述。在进行检索时,首先利用本体知识对检索条件进行预处理,对检索条件进行加强与扩展,接下来引擎用处理后的条件进行检索,将得到更符合用户需求的返回结果。

2 系统的设计

2.1 本体库的生成

2.1.1 知识分析与构建本体模型

这里的提法“知识分析”与我们通常所说的领域分析有着紧密的联系,同时也有明显的区别,二者的不同主要体现在本体建模要求语义上严格正规。文献[8]中给出了一种基于方面的本体构建方法,基本步骤可以分为:抽取领域词汇、抽取关键词汇构造立面,进而把所有词汇分类、加入词汇与词汇之间的关系。Ruben方法的优点在于构建条理清晰,是一种递增的构建方法;但它也存在明显不足:由于没有方向性的指导,领域词汇的获取趋于混乱,很难给出较完整的词汇库。我们对文献[8]给出的建模方法做了一定改进,形成自己的本体建模方法:

(1) 在领域分析的基础上,抽取关键词汇,构造叶子立面。经过这一步,本领域中重要方面都被确定,我们的基于本体与立面的中学语文资源描述与查询系统中立面树的叶子立面包括:标题、作者、出处、体裁、主人公、背景、作品主题、时代、格式、使用方式、所需软件、教育类型、教材类型、教育年级、资源主题。

(2) 对抽取出的叶子立面进行分类,形成上层立面,最终构造出一棵立面树。我们的系统最终构造出的立面树结构如图1所示。

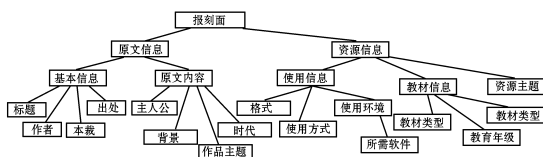


图1 立面树结构图

(3) 在(1)中得到的每个叶子立面下加入术语,用于在该立面下描述资源。这将在每个立面下形成庞大的术语库。在资源描述中所能用到的术语都应在该立面的术语库中出现。

(4) 加入立面、术语之间的关系。本体是以关系为中心的,没有关系或者关系很少很弱的本体不能起到应有作用。经过(1)~(3),得到的本体只具有简单的从属关系,需要加入它们之间存在的其它关系。对于本系统来说,本体库存在的一个重

要意义在于支持检索条件的预处理。在本系统的本体库中有几类特殊关系:DependOnObjectProperty、Adap teW ithObjectProperty、RelateW ithD iffFacetsProperty。DependOnObjectProperty关系定义相互依存关系,它意味着当主体或客体其一出现在查询条件中时,另一个也应该出现在查询条件中。Adap teW ithObjectProperty定义主体与客体的兼容性,它是一种同刻面下术语构成的关系,它意味着当主体出现在查询条件当中时,客体可以作为它的替换出现。RelateW ithD iffFacetsProperty,这是一种异刻面下术语所构成的关系。这三类关系之所以特殊,是因为它们直接关系到查询条件的扩展,参见后文。

2.1.2 本体编码

需要把3.1.1建立的本体模型正规表述出来。本系统采用OWL进行本体模型编码。我们采用英国剑桥大学开发的本体编辑DE:Protégé该工具是RDF编辑工具,配以OWL编辑插件可以编辑OWL文档。本体编码经过以下三个步骤:

(1) 立面编码 非叶子立面起到的作用是组织立面成一种层次结构,并不对资源描述与检索起作用;而叶子立面无论是对资源的描述还是检索都有重要作用,它们是本体中概念树的根。基于上面的考虑,不把整棵立面树编入本体,只把叶子立面以类的形式编入。在Protégé中,类的编辑在Class选项卡中进行。

(2) 术语编码 以一个叶子立面类为根,该叶子立面下所有术语为结点,通过SubClassOf与type关系,形成一棵概念树。在这棵概念树上的每个非叶子术语结点都被以类的形式编入本体;而那些叶子结点作为其父结点的实例被编入。

(3) 关系编码 关系编码分为两个步骤,首先定义关系;然后进行关系实例定义。关系定义主要确定关系类型、关系域、关系范围。一个关系实例形式为:

Subject Property Object

关系域用来限定Subject的类型,关系范围用来限制Object类型。例如:关系personTimes有如下定义:

```
<owl:ObjectProperty personTimes>
  <owl:domain rdf:about="character">
  <owl:range rdf:about="times"/>
</owl:ObjectProperty>
```

这样就意味着Subject必须为类character的实例,而Object必须为类times的实例。经过关系定义,本体中包含了我们需要的关系,但还没有各种关系的实例,还需要指出具体术语之间构成哪种关系,如:李白 personTimes唐。在Protégé中关系实例的指定也是在Individual选项卡下作的。

2.1.3 本体入库

本体入库是指根据OWL文件中本体信息推理生成知识闭包,存入关系式数据库的过程。这样做是为了提高检索效率。本体信息中蕴含着丰富的未直接表述的知识,但知识发现是一种代价很高的活动,这是由于推理过程十分复杂。如果知识发现在检索进行时进行,那么将导致检索效率非常低。我们的处理方法是事先发掘本体中知识,把这些知识存入RDB中,检索时避免推理活动,转而进行SQL查询。采用Jena来完成本体的知识发掘和入库。Jena是本体操纵开发包,为本体建模、操纵、推理等活动提供比较完善的支持。Jena配有基于规则的本体推理机,规则通过配置文件进行配置。我们在Jena原有配置文件基础上进行修改,去除了诸如基数限制等对本系统意义不大的规则。由于OWL文档是符合RDF文档规范的,所以Jena推理机将推理结果组织成一个个的RDF三元组,即Statement,基本

形式为:

< subject, property, object >

这就使得我们把这些 statement 存储在后台数据库中成为可能。基于 Jena 开发包,我们开发了 OWL (Lite) 到 DB (MySQL) 工具: OWL2DB。

对于本系统,我们采取的处理流程由图 2 所示。

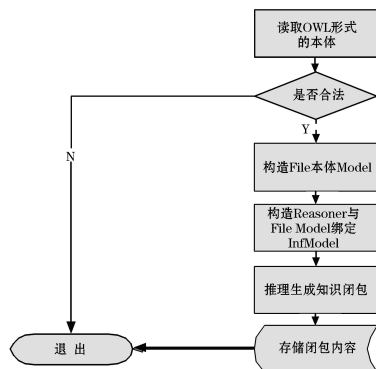


图 2 本体入库流程

3 检索条件的预处理

3.1 检索条件加强的实现

我们在本体中定义一种抽象关系: DependOnObjectProperty, 以及它的一些子关系, 如 canOpenWith 等。并指定术语形成该关系的实例。在得到用户检索条件后, 解析其中术语, 将具有依赖关系的术语所依赖的未选定术语加入到查询条件当中。这种检索条件的加强使得最终的结果更加精确、更加全面。

3.2 检索条件扩展的实现

我们令术语 ExtT 为由术语 T 经扩展所得到的术语, 有两种扩展结果: 其一, T 与 ExtT 同属于某一个刻面, 称为同刻面下扩展; 其二, T 与 ExtT 分属于不同刻面, 称为异刻面下扩展。对于同刻面下扩展, 由于扩展的影响范围被限制在一个刻面下, 并且不同刻面之间是种正交关系, 所以扩展出的术语不会与原有检索条件在其它刻面下的术语构成冲突; 而对于异刻面下扩展, 情况正好相反。正是由于这样的不同, 促使我们对两种扩展方式进行区分。通过扩展得到的新的查询条件需要加入到一棵新的查询树中, 这样不会对根据用户所指定条件进行的查询造成影响。

3.2.1 同刻面下扩展的实现 用户在刻面 F 下指定检索术语 T, 我们可以在刻面 F 下找出与 T 具有“密切关系”的术语, 用这些术语构造新的查询树。我们在本体中加入描述这种“密切关系”的元素, 比如, 同一刻面下术语之间的兼容关系。如在“Software 刻面下, doc 和 txt 之间构成这种兼容关系。在新的查询树中, F 下的检索术语是 ExtT, 而在其它刻面下, 检索术语应该和用户在刻面 F 下指定的术语保持一致。

3.2.2 异刻面下扩展 如果 ExtT 不在 F 下, 而在 ExtF 下, 那么那些在刻面 ExtF 下描述为 ExtT 的资源很可能也是用户所关心的。与同刻面下扩展中处理方法类似, 我们定义异刻面下扩展关系以及它的一些子关系。同时为控制异刻面下的术语扩展, 需要在系统中指定哪些刻面的术语之间可以构成异刻面下扩展关系。对于通过异刻面下扩展得到的新术语也应该放入一棵新的查询树中, 与同刻面下扩展时不同的是, 除了这些扩

展得到的术语外, 新查询树中不加入任何其它术语。

综上所述, 本系统中, 检索条件的预处理流程如下:

```

ArrayList sameFacetsExt, diffFacetsExt;
ArrayList terms = 检索条件中所有术语
for (I = 0; I < terms.length; I++) {
    dependTerms = 与 term[I] 构成依赖关系的术语集合
    diffFacetsExt.add(dependTerms)
    sameFacetsTerms = 与 term[I] 具有兼容关系的术语集合
    sameFacetsExt.add(sameFacetsTerms)
    diffFacets = 与 term[I] 所在刻面具异刻面扩展关系的刻面集合
    diffFacetsTerms = null;
    for (j = 0; j < diffFacets.length; j++) {
        diffFacetsTerms.add(diffFacets[j] 下与 term[I] 构成扩展关系的术语集合)
    }
}
整理 sameFacetsExt 与 diffFacetsExt
  
```

4 检索

基于刻面的资源描述、检索系统在查询条件描述、资源描述上都是采用树形结构(刻面树)进行的, 检索过程本质上是树匹配的过程。通过计算查询树与资源描述树之间的匹配度来选取资源。关于匹配度的计算方法, 请参考我们的另一篇文章“基于本体的构件描述和检索”。检索部分最大问题在于效率, 由于树匹配代价都是非常高的, 不能直接进行资源描述树与查询树的匹配, 我们采取的策略是直接 SQL 语句完成大部分计算, 余下小部分计算在程序中完成。我们的资源描述是以:

资源 D 刻面 D 术语

的形式存放在数据库中的。且用于描述资源的术语在刻面树中都是叶子术语, 对应的查询树中的非叶子术语也被它下面的叶子术语代替。在此基础上, 我们可以构造简洁 SQL 来完成匹配度的计算, 使得查询效率大大提高。我们所设计的 SQL 算法如图 3 示。利用该算法可以高效计算出资源在一个刻面下的匹配度。最后在程序中需要把一个资源在各个刻面下的匹配度加权相加, 这部分的代价不会太大。用这种方式很好地解决了查询效率问题。

对于每个刻面:

- 1) 在该刻面下, 对于检索条件所指定的叶子术语, 查找匹配资源, 一旦与该叶子术语匹配, 该资源在该刻面下匹配度为 1, 不匹配为 0;
- 2) 在该刻面下, 对于检索条件所指定的非叶子术语, 按照匹配度计算公式计算各个资源在该非叶子术语上的匹配度;
- 3) 对于各个资源在该刻面下各个检索术语上的匹配度, 取其中最大值作为在本叶子刻面下资源匹配度

图 3 单刻面匹配度算法

5 总结与展望

本文给出一种基于本体、刻面的资源描述与检索的实现。基于刻面的资源描述与检索方案是一种效率、准确率都很高的检索方式, 但它的检索完全依赖用户所指定的查询条件, 没有从语义上优化查询条件从而丰富结果的能力。本体可以真实地体现实体之间的关系, 以它作为知识库可以辅助刻面查询方式给出更准确、丰富、贴切的查询结果。当前系统实现基本体现本体带来的这些优势, 在如下方面还可以进一步扩展: (下转第 50 页)

是检查规则的约束部分。由于约束部分中的所有点都能在条件中找到。所以,对于(1.3)的每一种组合形式,(1.4)中的点都有相应的位置对应关系。我们将(1.4)中的点也从左到右排列:

$$R_1, R_2, \dots, R_{j-1}, R_j, \dots, R_s \quad (2.8)$$

对(2.8)的每一点都在(2.2)中从左到右查找,一旦发现,就将该点替换成被发现的序号,经过变换后(2.8)就变成了下面的一串数字符号:

$$v_1, v_2, \dots, v_{j-1}, v_j, \dots, v_s \quad (2.9)$$

它表示的是约束部分的第 j 个点对应条件部分的第 v_j 个点。因此,对应于条件特征串,有一组约束特征串:

$$v_{11}, v_{12}, \dots, v_{1s}$$

$$v_{21}, v_{22}, \dots, v_{2s}$$

$$\dots\dots\dots$$

$$v_{q1}, v_{q2}, \dots, v_{qs}$$

它标记了约束的点在条件语句中的位置。

2.3 结论特征串

和约束部分一样,结论部分(1.5)中的每一个点都可以在条件(1.3)中找到。和约束串的算法类似,结论特征串也有和条件特征串对应的一组:

$$w_{11}, w_{12}, \dots, w_{1m}$$

$$w_{21}, w_{22}, \dots, w_{2m}$$

$$\dots\dots\dots$$

$$w_{q1}, w_{q2}, \dots, w_{qm}$$

它标记了结论的点在条件语句中的位置。

3 根据规则进行推理

对于一条规则,我们用 $P = \{P_1, P_2, \dots, P_p\}$ 表示其点集合,也就是说该规则是关于 p 个点之间的几何关系。对于根据规则搜索到的语句组的点集合,我们用 $Q = \{Q_1, Q_2, \dots, Q_q\}$ 表示。我们的推理过程如下:

(1) 从数据库读取推理规则,将指针指向第一条记录。

(2) 如果指针为空,推理结束,则命题无法证明;否则读取指针指向的规则,指针向前移动一个记录。分解规则的条件语句类型。

(3) 根据条件语句类型,在几何信息库搜索相应的语句进行组合。如果任何一种类型没有被搜索到,转到(2);否则,计算搜索到的语句组合的点集合 Q 的大小 q ,执行(4)。

(4) 比较 q 和规则的点集合大小 p 。如 $q > p$, 转到(2);如果 $q = p$, 则执行(5);如果 $q < p$, 则需要先进行对象的替换,然后执行(5)。

(5) 计算搜索到的语句的特征串 t 。根据 t 在规则的条件特征串中搜索,如果没有被发现,转到(2);否则,从规则的约束特征串中取相应的特征串 v , 执行(6)。

(6) 将搜索到的语句组中的点从左到右依次放在链表 $PList1$ 中,根据 v 在 $PList1$ 中将约束部分的点找到,验证命题的约束。如果不符合,转到(2);否则,执行(7)。

(7) 从规则的结论特征串中取相应的特征串 w , 根据 w 在 $PList1$ 中将结论的点找到依次放在链表 $PList2$ 中。

(8) 分解规则的结论语句,从 $PList2$ 中取相应点,逐一生成新语句。并且验证新语句是否是要证明的命题,如果是,则命题

得证;否则,如果是新语句,则将该语句加入到几何信息库中,转到(2),则转到(2)。

4 结束语

本软件数据库用 Microsoft Access 2000, 开发工具用 Microsoft Visual C++ 6.0。作者对许多几何命题进行过推理,都能在较快的时间得到结论,添加的规则越多,其推理的能力也越强。在有些比较复杂的规则中,条件语句组合后,其特征串比较复杂,但由于规则的特征串计算是在用户添加规则时完成的,而在推理时只有一次特征串的计算,所以能保证推理的速度基本上不受规则的影响。

参 考 文 献

- [1] 张景中,高小山,周咸青.基于前推法的几何信息搜索系统.计算机学报,1996,19(10).
- [2] 张景中,杨路,高小山,周咸青.几何定理可读证明的自动生成.计算机学报,1995,18(5).
- [3] Wu Wen-tsun. On the decision problem and mechanization of theorem in elementary geometry, Science Sinica, 1979, 21(2).
- [4] 吴文俊.几何定理机器证明的基本原理.科学出版社,1984.
- [5] Tarski A (陆钟万译).初等代数和几何的判定方法.北京:科学出版社,1959.
- [6] 张景中.平面几何新路.四川教育出版社,1992.
- [7] Charles N Fischer, Richard J LeBlanc 编译器构造 C语言描述.郑启龙,姚震,译.机械工业出版社,2005.

(上接第3页)

(1) 丰富本体 给出更多的可以作为扩展依据的关系,为查询条件的扩展提供更多路径。

(2) 丰富查询条件扩展方式 目前同刻面下扩展与异刻面下扩展分别得到各自的新查询树,然而在一棵新查询树中同时包含同刻面扩展与异刻面扩展得到的术语是有意义的。这就要求能够避免两种扩展方式带来的矛盾结果,而取它们相互协调的部分组成新查询树。

(3) 实时推理 为了提高检索效率,本系统采取提前挖掘本体知识以备检索的方案。这种方案不支持本体知识的动态增长,对本体知识的演化造成障碍。进行本体知识的实时推理可以解决这些问题,但检索代价大大增加。解决这对矛盾需要找到一种提高推理效率的方式。

参 考 文 献

- [1] 张翔,张福炎.数字图书馆中数据对象的描述与检索.计算机工程与应用,2001,8.
- [2] 孙青,邓苏,黄宏斌,黄金才.基于LDAP的分布式异构信息检索模型研究.计算机应用研究,2003,11.
- [3] 王渊峰,张涌,任洪敏,等.基于刻面描述的构件检索.软件学报,2002,13(8):1546-1551.
- [4] 马亮,谢冰,杨美清.多构件系统刻面检索机制.电子学报,2002.
- [5] William S, Austin T. Ontologies IEEE Intelligent Systems, 1999 Jan/Feb: 18-19.
- [6] 上海教育资源库. <http://www.sherc.net>
- [7] Zuo Zhihong, Zhou Mingtian. Web ontology language OWL and its description logic foundation IEEE 2003.
- [8] Ruben Prieto Diaz. A Faceted Approach to Building Ontologies IEEE, 2003.