基于本体的资源描述和检索方法研究

肖　君[1,2]　彭　鑫[3]　赵文耘[3]
[1](华东师范大学教育信息技术系,上海　200062)
[2](上海远程教育集团,上海　200086)
[3](复旦大学计算机科学与工程系软件工程实验室,上海　200433)
E-mail:xiaoj@shtvu.edu.cn

摘　要　计算机和网络技术的飞速发展使得网络日益成为资源获取的重要来源。网络资源由于其数量具大、多元化等特点,使得资源的组织管理更加复杂。资源的描述和检索是其中的主要问题。基于目录分类、关键字等的资源描述方法缺乏语义信息,因此无法较好地满足用户的检索请求。文章针对这一问题,引入领域本体作为资源描述的知识基础,以刻面作为资源描述框架,从而更加智能地满足用户的资源检索请求。

关键词　本体　描述　检索　匹配　刻面

文章编号　1002-8331-(2005)36-0009-03　　**文献标识码**　A　　**中图分类号**　TP393

Research on Ontology-based Representation and Retrieval of Resources

Xiao Jun[1,2]　Peng Xin[3]　Zhao Wenyun[3]
[1](Huadong Normal University, Shanghai 200062)
[2](Shanghai Distance Education Group, Shanghai 200086)
[2](Department of Computer Science and Engineering, Fudan University, Shanghai 200433)

Abstract: The rapid development of computer and network makes Internet resource more and more important in our lives.Organization and management of Internet resource is much more complex than that of conventional resource due to the features of large quantity and variety.Representation and retrieval of resources is the main problem.Conventional representing and retrieving methods lack semantic information, so can not meet the resource requirements of users. Aiming at this, we introduce domain ontology and facets as the knowledge base and framework of resource representation respectively.Then the resource requirements of users can be met more intelligently.

Keywords: ontology, representation, retrieval, matching, facet

1　引言

计算机和网络技术的飞速发展使得网络日益成为资源获取的重要来源。网络上涌现出了一大批信息资源站点,例如教育教学资源站点(如上海教育资源库[1])、软件工具资源站点(如华军软件园)、软件构件库(如上海构件库[2])、电子图书馆(如上海数字图书馆)等。这些资源库虽然都偏重某一类资源,但都具有资源数量巨大、资源类型多元化、变化性大等特点。由此带来的一个关键问题是如何对资源进行有效的组织管理以满足用户的使用需求,其中主要是资源的描述和检索问题。

目前常用的资源描述方案有属性—值、编目结构、关键字、元数据等。而在软件构件库领域(如上海构件库[2]),基于刻面的构件描述和检索是目前应用最为广泛的一种方法,例如青鸟构件库[3]采用的就是以刻面分类为主多种分类模式相结合的构件描述方案。

这些描述和检索方法的主要问题是缺少语义信息。可以设想一下,在人类社会交往中,人们可以利用主观知识对信息进行加工,这样提到‘诗仙’人们就可以想到‘李白’,提到‘唐宋八大家’就会想到‘韩愈’、‘柳宗元’等。然而对于软件系统,无论是关键字还是属性值,都是作为机器符号存在,完全与它们所关联的语义无关。这样,系统在处理用户查询请求时只能进行机械的符号匹配,不能基于相关知识进行关联和扩展,因此无法灵活地满足用户的查询请求。

在基于刻面以及传统的基于关键字的方法中,检索操作都是基于词汇进行的,存在较大程度上的语义缺失问题。虽然术语辞典可以提供语义信息,但大多只能描述术语间的一般特殊关系和同义关系。针对这些问题,我们在国家863软件重大专项《Linux多媒体网络教学资源管理和应用平台软件研究》支持下依托上海教育资源库,进行了基于本体的教育资源描述和检索方面的研究。

2　领域本体和刻面描述

2.1　领域本体

近年来,本体论作为共享知识的表达基础已经被广泛应用于信息科学中,例如软件复用[4]、信息检索[5]、需求获取[6]等。本体论是一个哲学概念,用于描述事物的本质,知识工程学者借用

基金项目:国家863高技术研究发展计划资助项目(编号:2004AA1Z2330);国家自然科学基金资助项目(编号:60473061),上海市科委科研攻关项目(编号:04DZ15022)

作者简介:肖君(1974-),男,安徽人,博士生,主要研究方向:知识管理技术。彭鑫(1979-),男,湖北人,博士生,主要研究方向:软件工程、企业应用集成(EAI)。赵文耘(1964-),男,江苏人,教授,博导,主要研究方向:软件工程及电子商务、企业应用集成(EAI)。

这个概念，是为了解决知识共享中的问题[6]。语义的深度和广度总是与相应的领域空间相关，领域越具体相应的语义越具有深度，针对性也越强。从资源描述的角度来说，其所包含的精确语义都与特定领域相关，例如上文提到的‘唐宋八大家’这一语义信息可以在‘中学语文’领域中定义，而在上层的‘教育教学’领域中就可能由于领域过于宽泛而无法体现。

领域本体为领域内的概念以及概念间广泛存在的各种关系提供了共享的描述，因此可以作为领域内资源描述的知识基础。基于领域本体一方面可以引导用户更加准确、完整地描述自己的检索要求，另一方面可以为资源描述提供丰富的语义注解，从而更好地弥合用户资源需求与资源描述之间的‘鸿沟’。

2.2 基于刻面的资源描述

刻面是事物的刻画面，基于刻面的描述是从多个刻画面对资源进行描述的方法。刻面按照组成关系构成一刻树，每一个叶子刻面下的取值空间称为术语空间。基于刻面的描述方案主要由三部分组成[7]：刻面分类方案、资源的刻面描述集合以及刻面描述术语之间的关系，即术语辞典。刻面可以按照观点或特定领域的维度进行组织[8]，分层的刻面方案使各刻面的含义更加容易理解，并且使资源描述的维度更为丰富。图1是我们在中学语文领域资源描述中所使用的刻面树，其中‘资源信息’下的刻面描述资源使用方式、对象方面的信息，而‘原文信息’下的刻面用来描述与资源内容相关的信息。由于我们主要关注资源的领域语义，因此资源本身的作者、入库时间等信息都不在刻面描述中体现。按一般特殊关系构成的术语空间为用户查询要求以及资源的描述提供了多种抽象层次上的选择。例如图2中叶子刻面‘作者’下定义了‘诗人’、‘政治家’等术语类以及‘王安石’等术语实例。

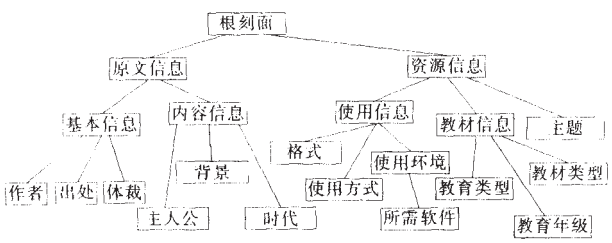

图1 中学语文领域资源描述刻面定义

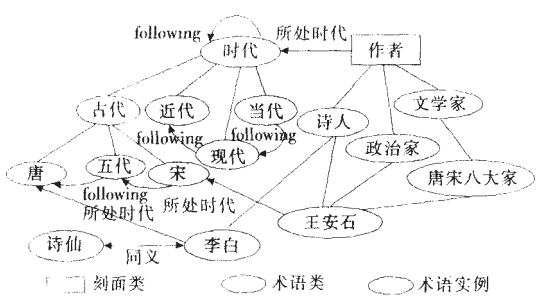

图2 本体中的术语定义

确定了刻面及术语方案后，每一个资源都将按照这一框架进行描述，即确定每一叶子刻面下的术语选择。根据资源的具体语义，每个刻面下可能存在零个或多个术语选择。基于刻面的方法相比传统的基于关键字的方法有了很大进步，但也存在着一定的局限性，主要体现在领域知识表达不完整以及匹配方式不灵活。这些不足可以由领域本体中的知识进行弥补。

2.3 基于本体的刻面和术语定义

由于本体在语义层面上详尽描述了领域内的概念以及概念间的关系，因此可以成为用户资源需求以及资源描述的公共知识基础。引入本体后，资源描述仍然以刻面方案为主，不同的是领域本体将作为构件描述和检索的知识基础存在。领域知识主要体现在叶子刻面下的术语空间上，而刻面树主要体现一系列正交的描述方面的组成关系。我们将叶子刻面和术语定义为本体中的概念，而叶子刻面、术语以及其它概念之间的关系也将在本体中定义。OWL[9]是W3C推荐的Web本体语言，资源描述本体将主要使用OWL来描述。

OWL中的概念主要由类(Class)、实例(Instance)和关系(Property)组成。叶子刻面在本体中定义为类，而术语中的抽象术语和具体术语分别定义为类和实例。同一叶子刻面下定义为类的术语构成一系列继承关系(在OWL中表示为rdfs: subClass of)，相应的叶子刻面类是它们共同的父类。定义为实例的术语是上层术语类的实例。

例如图2中的刻面‘作者’是一个类，术语‘诗人’是它的子类，而术语‘李白’是类‘诗人’的实例。由于本体中允许多继承，因此同一刻面下的术语空间也不再是树状结构，例如图2中‘王安石’既是‘诗人’的实例又是‘政治家’的实例。除了继承关系，本体中还可以定义其它关系，例如定义类‘作者’和‘时代’之间的‘所处时代’关系，或者定义‘时代’与‘时代’之间的following关系(先后关系)。除了自定义关系，本体中还可以使用一些基本关系，例如实例‘诗仙’和‘李白’是同义关系(在OWL中表示为owl: sameAs)。这些关系能够为用户检索请求和构件描述提供附加语义。例如，通过‘following’关系，用户就可以表述‘唐之后’的这样一个时代限定，而通过‘所处时代’关系，用户就可以表达‘唐代诗人的作品’这一资源需求。

3 基于本体的资源检索

3.1 检索过程

在一般的刻面检索方法中，用户按照刻面方案指定查询树，即确定每一刻面下的术语要求，然后系统将查询树与资源描述进行匹配并返回符合要求的资源列表。在这一过程中，用户只能按照术语空间中定义的术语描述资源需求，而资源匹配过程则按照刻面进行机械的术语匹配。因此匹配有赖于用户要求与资源描述之间的‘刚性’一致，资源重用的机会大大降低。引入领域本体作为知识基础后，不仅用户更加灵活地表达查询要求，查询过程也可以基于本体进行联想和扩展，从而更好地满足用户的资源需求。

基于本体的资源检索过程主要包括以下步骤：

(1)用户检索请求描述及处理：用户按照刻面方案，直接或间接指定各叶子刻面下的查询术语，系统根据领域本体对查询请求进行预处理从而生成查询树。

(2)联想及扩展：以领域本体为基础，对查询树进行联想及扩展，从而使与用户要求近似或相关的资源纳入查询目标中。

(3)资源匹配：根据查询树以及资源刻面描述进行资源匹配，匹配过程中将按照匹配算法计算匹配度，然后按匹配度大小依次返回检索结果。

基于本体的推理过程主要体现在前两个步骤中。资源匹配由于是最耗时的一个环节，因此我们通过对领域本体进行预处理省去了匹配时的推理过程。在我们的资源检索方法中，领域本体是在领域专家的参与下通过领域分析获得的。我们使用了protégé[10]作为本体建模工具，得到的OWL文件中包含了相关

领域的直接知识。在预处理中，我们使用 Jena[11]的推理引擎利用 OWL 文件中的知识进行推理，得到全部显式或隐式的知识后以 RDF 三元组的形式存储在 MySQL 数据库中，这样在匹配环节仅通过数据库查询操作就可以完成资源匹配。

3.2 查询树生成

在关键字或刻面检索方法中，用户只能直接指定目标，例如如果指定作者，则必须选择“李白”、“王安石”等确定信息。这主要是由于传统的检索方法是按字形进行术语或关键字匹配，不包含任何语义信息。而在基于本体的资源检索中，由于相关术语都是本体中包含语义信息的概念，因此除了直接指定术语用户还可以通过概念间的关系描述查询要求，主要有以下三种方式：

(1)指定抽象术语：指定该刻面下的一个术语类而不是实例，例如在“作者”刻面下指定“诗人”，系统通过本体中的实例关系可以列举出属于“诗人”实例的那些术语。

(2)通过同刻面类间关系间接指定：通过定义在同一个类上的关系（即关系的主体和客体都是同一个刻面或术语类）实例来间接指定，例如通过定义在“时代”类上的“following”关系可以指定“宋以后的时代”，通过定义在“作者”刻面类上的“同时代”关系可以指定“与李白同时代的作者”。

(3)通过与其它类间的关系间接指定：与方式(2)类似，不过使用的是当前刻面下的类与其它类之间的关系。例如“作者”刻面下可以在指定“诗人”这一抽象术语后，继续通过“时代=宋”、“性别=男”，“风格=豪放”这样一些关系实例来进一步限定原文作者。

用户直接指定或者通过以上三种方式间接指定各刻面下的术语后，系统将对查询要求进行解析，得到完全实例化的查询树。除了对关系的处理外，系统还要对同义词进行替换，例如将“诗仙”替换为同义的术语“李白”。这样得到的查询树中各刻面下术语全部用与资源描述树一致的实例术语表达，可以以刻面为单位进行资源匹配。

3.3 基于查询树的扩展联想

文献[5]认为传统的信息检索机制存在三个深层次问题，即“忠实表达”问题、“表达差异”问题和“词汇孤岛”问题。“词汇孤岛”问题可以通过领域本体中丰富的概念间关系来解决，而另两个问题则与用户对检索请求的表达相关。如果直接按照查询树进行资源检索，那么可能由于用户表达的缺陷造成结果不理想。另一方面，资源检索过程中的相关联想也很重要，这一点在教育资源领域尤其突出。例如，教师用户在检索“阿 Q 正传”相关课件时，一般也会对介绍作者鲁迅本人以及辛亥革命这一历史背景的资源感兴趣。因此我们在进行检索前还应该对查询树进行联想和扩展，以期将与用户直接目标近似或相关的资源纳入检索范围。因此，我们通过对原始查询树进行扩展联想来扩大资源检索范围。查询树扩展联想包括以下两种基本方式：

(1)同刻面术语扩展：根据查询树上同刻面下实例术语的语义相似性或密切关联性，由已有术语扩展到相似术语。例如，“格式”刻面下，由于“txt”、“bmp”等格式均能插入到 ppt 中，即具有“格式兼容”关系，因此由用户指定的“ppt”术语可以扩展出“txt”、“bmp”等术语，从而增大了检索范围。

(2)跨刻面术语联想：根据查询树上某些刻面上的实例术语联想到其它刻面上的术语取值，从而体现资源使用时的相关性。例如，一个《阿 Q 正传》的课件根据原作作者“鲁迅”可以联想到主人公刻面取值为“鲁迅”，根据背景“辛亥革命”联想到原文主题刻面取值为“辛亥革命”。这样，与《阿 Q 正传》课件相关的介绍鲁迅和辛亥革命的其它资源就能通过联想纳入检索范围，满足了相关资源在使用上的相关性。

这两种方式分别代表了不同的扩展方向：前者着眼于检索条件的近似性，从而“柔性”满足用户的直接检索目标；后者着眼于资源使时的相关性。在用户查询树基础上，通过这两种基本方式的组合可以获得各种扩展查询树，从而提供更加符合用户需求的检索结果。这两种扩展联想方式的实现有赖于领域本体中定义的各种关联关系。例如，定义在“格式”刻面上的“格式兼容”关系实例包括“ppt-bmp”、“ppt-avi”、“doc-jpeg”等。如果通过这两种基本方式的综合获得各种合理的联想扩展方案仍然是一个难题，需要进一步加以研究。

3.4 资源匹配算法

确定查询树后，资源检索就演变成一个查询树与资源描述树之间的资源匹配过程。查询树中同一刻面下各术语间是并的关系，而多个刻面之间是交的关系[3]。刻面匹配以叶子刻面为单位进行，对于查询树 Q 以及资源描述树 C，匹配度 M 的计算公式为：

$$M=\sum_{i=1}^{n}\alpha_i m(Qt_i, Ct_i)$$

其中 n 表示查询树中的叶子刻面数，α_i 表示第 i 个查询刻面的相对权重，Qt_i 和 Ct_i 分别表示 Q 和 C 在第 i 个叶子刻面上的术语集合，$m(Qt_i, Ct_i)$ 代表它们的匹配度：

$$Sq=\{ins|ins\in termIns\wedge\exists t\in Qt_i: ins=t\vee ins<t\}$$

$$Sc=\{ins|ins\in termIns\wedge\exists t\in Ct_i: ins=t\vee ins<t\}$$

其中 termIns 表示本体中的实例术语集合，< 表示 instance of 关系。

于是，$m(Qt_i, Ct_i)$ 可以定义为：

$$m(Qt_i, Ct_i)=\frac{|Sq\cap Sc|}{|Sq|}$$

叶子刻面的相对权重体现了用户的资源需求中各方面因素的相对重要性。一般来说，与内容相关的刻面权重较高。而 Q 与 C 在每个叶子刻面上的匹配度 m 则取决于本体中该刻面下二者公共实例术语数量的比例。这种计算方法较适用于本体中术语分类方案较为匀称的情况，即术语空间中的概念层次数在各处比较平均。这对于一个构造良好的本体模型而言是可以满足的。

4 总结和展望

传统的资源描述和检索方法具有较大的语义缺失性，因此无法很好地满足用户的资源检索要求。我们在“Linux 多媒体网络教学资源管理和应用平台软件研究”这一课题中，引入领域本体作为资源描述和检索的语义基础，在资源描述和用户资源需求表述间建立起一座基于语义的桥梁，从而能够更好地满足用户的资源需求。目前，该项目已经在上海地区和西部三省多个示范点推广和应用，实现了上海和西部的资源联动，对国家教育信息化工作的推动具有一定的示范性。

(收稿日期：2005 年 9 月)

(下转22 页)

表 1　9/7 小波滤波器与 MPEG 滤波器下采样的 PSNR 比较

foreman	9/7 wavelet				MPEG filter			
QP	Y PSNR/dB	U PSNR/dB	V PSNR/dB	Bitrate/(kbits/s)	Y PSNR/dB	U PSNR/dB	V PSNR/dB	Bitrate/(kbits/s)
36	33.436	39.775	42.162	230.026 7	33.47	39.757	42.129	229.92
38	32.544	39.277	41.628	188.28	32.53	39.297	41.44	187.146 7
40	31.72	38.994	41.218	160.84	31.638	39.077	41.168	159.84
42	31.336	38.839	40.711	150.16	31.246	39.011	40.812	148.706 7

表 2　9/7 小波滤波器与 MPEG 滤波器下采样的 PSNR 比较

football	9/7 wavelet				MPEG filter			
QP	Y PSNR/dB	U PSNR/dB	V PSNR/dB	Bitrate/(kbits/s)	Y PSNR/dB	U PSNR/dB	V PSNR/dB	Bitrate/(kbits/s)
36	29.548	36.349	38.363	819.613 4	29.633	36.558	38.504	843.466 7
38	28.469	35.645	37.794	659.626 7	28.5	35.858	38.029	676.746 7
40	27.599	35.306	37.548	555.373 4	27.676	35.424	37.666	577.106 7
42	26.857	35.002	37.317	488.906 7	26.853	34.99	37.349	497.933 4

4　实验结果

提出的想法在文[3]得以实现，并且对 9/7Daubechieshe 小波下采样与 MPEG 下采样进行了比较，输入的序列格式是 CIF 格式，包括 foreman 和 football，帧率是 30fps。

表 1 是运动不太剧烈的 foreman 序列，提供了 9/7 小波下采样和 MPEG 下采样的比较。从表中可以看出，在高端和底端 9/7 小波下采样与 MPEG 的下采样 bit-rate 相差不大，但是 PSNR 有所提高，这是因为提出的方案使基本层的运动向量能够比较准确地预测出增强层的运动向量，以至于 PSNR 有所提高。

表 2 是运动比较剧烈的 football。可以看出，9/7 小波下采样的 bit-rate 与 MPEG 相比，有了明显的下降，但是 PSNR 差不多，这是因为从基本层预测出的运动向量，能够使增强层运动估计补偿比较准确，比特率有了较大幅度的下降。

以上的实验结果表明，本文提出的可伸缩视频编码方案有机地将空域和频域结合起来，充分挖掘了层间的宏块分割信息，运动向量关系，可以进一步的提高编码效率，同时，基本层视频序列的主观质量也有了较大的提高。

5　结论

在可伸缩视频编码中，本文提出了一种新的下采样方法，有效地促进了基本层运动信息和增强层运动信息的相关性，并给出层间的运动向量关系和推导过程，提高了编码效率。可以看出，小波下采样的序列比 MPEG 下采样的序列视觉效果清晰。在以后的工作中，下采样因使用自适应小波滤波器，层间的运动向量关系和基本层的上采样滤波器应进一步加以研究，以提高编码效率。（收稿日期：2005 年 9 月）

参考文献

1.I Daubechies, W Sweldens.Factoring wavelet and subband transforms into lifting steps.Preprint, 1996

2.Peisong Chen, John W Woods.Bidirectional MC-EZBC With Lifting Implementation[J].IEEE Trans circuits and systems for video technology, 2004; 14(10): 1183~1194

3.Scalable Video Model 3.0[S].ISO/IEC JTC 1/SC 29/WG11 MPEG 2004/N6716, 2004-10

4.X Li.Scalable video compression via overcomplete motion compensated wavelet coding[J].Signal Process: Image Communication, 2004; 19: 637~651

5.Y Andreopoulos, A Munteanu, J Barbarien et al.In-band motion compensated temporal filtering[J].Signal Process: Image Commun, 2004; 19: 653~673

6.ITU-T.Video Coding for Low Bitrate Communication[S].ITU-T Recommendation H.263, 1995-11, 1998-01

7.ISO/IEC JTC1.Coding of audio-visual objects-Part 2: Visual[S].ISO/IEC 14496-2(MPEG-4 Visual), Version 1, Amendment 1(Version 2), 2000-02

8.X Li, L Kerofsky, S Lei.All-phase motion compensated prediction for high performance video coding[C].In: Proc Int Conf Image Processing, 2001; 3: 538~541

（上接 11 页）

参考文献

1.上海教育资源库.http://www.sherc.net

2.上海构件库网站.http://www.sstc.org.cn

3.常继传，李克勤，郭立峰等.青鸟系统中可复用软件构件的表示与查询[J].电子学报，2000; 28(8): 20~23)

4.Sidney C Bailin.Software Reuse as Ontology Negotiation[C].In: Proceedings of the 8th International Conference on Software Reuse(ICSR 2004), 2004: 242~253

5.董慧，杜文华.基于本体和多代理的数字图书馆信息检索模型[J].中国图书馆学报，2004; 30(150): 63~65

6.金芝.基于本体的需求自动获取[J].计算机学报，2000; 23(5): 486~492

7.王渊峰，张涌，任洪敏等.基于刻面描述的构件检索[J].软件学报，2002; 13(8): 1546~1551

8.Prieto-Diaz R.A faceted approach to building ontologies[C].In: Proceedings of IEEE International Conference on Information Reuse and Integration(IRI 2003), 2003: 458~465

9.Sean Bechhofer et al.Owl Web Ontology Language Reference.http://www.w3.org/TR/owl-ref/, 2004-02-10

10.Protégé home.http://protege.stanford.edu

11.Jena home.http://jena.sourceforge.net