基于扩展语义网的知识资源组织技术研究*)

冯 永 李 华 吴中福 钟 将 叶春晓
(重庆大学计算机学院 重庆 400030)

摘 要 采用知识点的方式组织知识资源,有利于知识的获取、分享、分配和存取。但是传统的树型结构对知识的整体关系描述能力不足,不利于分布式环境下对知识资源的查找和定位。语义网是一种可以详细描述本体间复杂关系并具有天然分布式特性的技术。然而一般的语义网本身不是按照知识点的方式进行组建。本文对语义网进行扩展,使其适用于描述基于知识点组织的知识资源。通过应用案例,按照知识点进行扩展的语义网可以有效地描述知识资源间的相互关系,便于知识的理解和利用,并且对知识的查找和定位也变得更加方便。

关键词 知识资源,扩展语义网,知识点,分布式环境

Knowledge Resources Organization Technology Research Based on Expended Semantic Web

FENG Yong LI Hua WU Zhong-Fu ZHONG Jiang YE Chun-Xiao
(College of Computer Science of Chongqing University, Chongqing 400030)

Abstract Knowledge points are used to organize knowledge resources, which is convenient to knowledge acquisition, sharing, distribution and visit. But the traditional tree structure is lack of ability to knowledge overall relationship's description, and it is not convenient to the search for knowledge resources and positioning in distributed environments. Semantic Web is a technology which can describe complex relationships between ontology in details and with the natural characteristics of the distributed. However, the general semantic Web is not constructed in accordance with knowledge points. The semantic Web has been expanded in the paper, so that it can be applied to describe knowledge points-based knowledge resources. Case studies demonstrated that the expended semantic Web can effectively describe the interrelationships between knowledge resources, it is easy to understand and use the knowledge, and location and positioning of knowledge have become more convenient.

Keywords Knowledge resources, Expended semantic Web, Knowledge points, Distributed environment

1 引言

在知识经济时代,知识已经超越了土地、工具、劳动力而成为决定生活水平的最重要的要素。随着以计算机、多媒体、通信、网络、人工智能等为代表的信息收集、处理、加工、传输等技术的飞速发展,信息技术正在向各个领域渗透,并不断地改变着人们的思维方式、知识获取方式和知识利用方式。知识管理是每个组织(指企业、科研机构、学校等)在此新形势下做出的战略性反应。知识管理就是一个组织对其所拥有的知识资源进行管理的过程,运用集体的智慧提高应变能力和创新能力,它的内容涉及如何在组织内部和运作过程中获取、分享、分配和存取知识[1]。

知识资源是某种专门知识的有机的、统一的集合,在信息技术领域,其表现形式可以是文本、图形、图像、音频以及视频等。知识资源在未进行有效的组织前,可能是零散的、无体系的,不利于人们对知识的获取、分享、分配及存取。因此,有必要采用一种有效的方式对知识资源进行组织。基于知识点的知识资源组织就是一种有效的知识资源组织方式。知识点是知识资源中的最小单位,具有不可分割性,拥有独立的完整概念,若干知识点可以构成一个知识资源应用案例[2]。

基于知识点的树型结构是一种便于人们直观理解的知识资源组织形式,它可以从形式上对知识资源进行处理,但是无法对知识资源的整体关系有效描述,特别是当知识资源存储在分布式环境中,这种矛盾更加突出,即传统的树型结构割裂了知识之间的关系,在分布式环境中不利于知识资源的查找和定位。

语义网(Semantic Web)是由 Web 的发明者 Bemers-Lee 等人[3]提出的下一代 WWW 技术,是一种可以详细描述本体间复杂关系,并具有天然分布式特性的技术[4,5]。本体定义了用于描述和表示领域知识的术语,它用于人、数据库和应用之间共享信息。它通常表达为一组对象(概念)、关系、函数、定理和示例[6~8]。

然而,一般的语义网本身不是按照知识点的方式进行组建的。因此,本文对语义网进行扩展,使其适用于描述基于知识点组织的知识资源,可以有效地描述知识资源间的相互关系,便于知识的理解和利用,并且易于知识的查找和定位。

2 基于知识点的知识资源组织

知识资源中知识的表征方式可以是一个个的知识对象。知识对象在信息领域是数字的或非数字的,在技术支撑的环境中能使用的、重复使用的或被引用的实体[9]。它们可以是一个 PPT 文档、一张图片、一段录音剪辑等等。

*)重庆市高等教育教学改革研究重大项目(0616001);重庆市科委自然科学基金计划资助项目(2007BB2192);重庆市科委自然科学基金计划资助项目(2007BB2372)。冯 永 博士,主研方向:知识管理、知识发现、知识学习;李 华 副教授;吴中福 教授;钟 将、叶春晓 副教授。

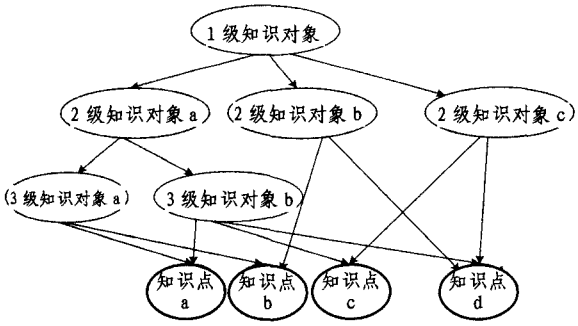

图1 某一知识资源的树型组织结构图

为了更好地实现知识资源的共享，对知识资源的组织引入了知识点这一概念。一个知识资源可以表示为：{1级知识对象，2级知识对象，3级知识对象，…，n级知识对象，知识点}。其中的1级知识对象包括若干2级知识对象，2级知识对象包括若干3级知识对象，等等。每一个1级知识对象包括的2级知识对象可能具有相关关系，每一2级知识对象包括的3级知识对象也可能具有相关关系，而知识点是知识资源的基本粒子，它们不隶属于任何知识对象。若干知识点可以按照应用要求组合为知识结构图，成为一个知识应用案例。图1是某一知识资源的知识对象和知识点组成的树型知识结构图。

图1中的知识资源由{1级知识对象，2级知识对象，3级知识对象，知识点}组成。其中1级知识对象包括了3个2级知识对象；2级知识对象a包括了2个3级知识对象；整个知识资源的知识点一共有4个。知识对象之间有包含关系和一定的相关关系。它们的包含关系通过树型结构体现，相关关系通过知识点体现。图2所示的是2级知识对象a和2级知识对象b之间的相关关系。

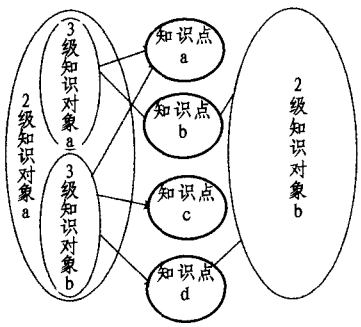

图2 2级知识对象a和b间的相关关系

从图2中可以看出：2级知识对象a与2级知识对象b之间的相关关系是依赖于它们所拥有相同的知识点b和d。两个知识对象更深层次的相关关系，通过这种树型描述方式很难观察出来，因为这种表达方式缺乏语义特征。

3 基于扩展语义网的知识资源组织

通过上面对基于知识点的知识资源组织的了解，我们发现采用传统的树型结构对各级知识对象和知识点的整体关系（主要有知识资源间的包含关系、传递关系、依赖关系、适应关系等）的描述能力有限，不利于知识资源分布式存储后知识资源的查找和定位。但是，语义网本身就可以详细描述本体间比较复杂的关系，并且具有天然的分布式特性的技术。因此将语义网技术应用于基于知识点的知识资源组织，必将增强对知识资源整体关系的描述能力，有利于知识资源分布式存储后对知识资源的查找和定位。

然而，一般的语义网本身不是按照知识点的结构进行构建的。为此，需要对传统的语义网进行扩展，使其适用于描述基于知识点的知识资源组织。扩展策略描述如下：

(1)对概念进行分类。分为：各级知识对象、知识点、一般概念。这些概念在语义网中通过属性“概念类型(AType)”表明。

(2)定义概念之间的关系。以上的几种概念都可能相互间发生不同的关系。为此，我们定义：

1)相关关系 R(概念1，概念2，x)，x 表示相关度。最大相关度表示同义概念(根据具体应用可以将相关关系细化成相似关系、反义关系、递进关系等)；

2)部分关系 P(概念1，概念2)，表示概念1包括概念2；

3)继承关系 K(概念1，概念2)，表示概念2由概念1继承而得；

4)实例关系 I(概念1，概念2)，表示概念2是概念1的实例。

(3)对概念增加两种属性：

1) 适应层次属性 ALevel(概念)，表示概念适合的接受层次；

2) 内容属性 AContent(概念，内容实体)，对概念(尤其是知识点)进行解释、说明。内容实体包括两个属性：文件格式、元数据地址(即该内容实体在元数据目录中的位置)。

在具体的应用中，还可以根据需要定义更多的概念、关系和属性。有了上述规则，就可以利用语义网按照知识点对知识资源进行组织，从而根据知识点语义网查找和定位知识资源并生成应用案例。

4 应用案例

在国家发改委的下一代互联网示范工程(CNGI)中，建设开放式教育网络是构建终身学习体系的重要组成部分，而配套的远程教育资源建设是其主要内容。远程教育资源分为媒体素材、题库、案例、课件与网络课件、网络课程等。为了便于分布在各地的学习者有效地学习，这些零散的教育资源需要重新进行组织。下面以课程《计算机组成原理》为例说明扩展语义网对知识资源进行组织的可行性和有效性。

4.1 利用扩展语义网描述课程

图3是通过知识点组织的课程《计算机组成原理》。它是按照{课程、篇、章、节、知识点}这几个层次进行组织的。其中每一篇包括的章可能具有相关关系，每一章包括的节也可能具有相关关系，同一节或章中间的知识点也可能存在不同程度的相关关系。

图3只是一个示意图，并不完整。事实上，每个知识点都包含一些文字、图片、声音、动画等说明内容；而且每个知识点都与一些不是知识点但又常用的概念相关。如“计算机软硬件概念”知识点与概念“硬件”、“软件”相关；“计算机系统层次结构”知识点与“早期的计算机”、“两级层次结构计算机系统”等概念相关。

为了增强对课程《计算机组成原理》中篇、章、节、知识点整体关系的描述能力，利用扩展的语义网规则将《计算机组成原理》重新进行组织，如图4所示。

图4中方框表示属性，椭圆和圆圈表示概念。因为图4只是示意图，所以还有一些属性和概念没有画出。

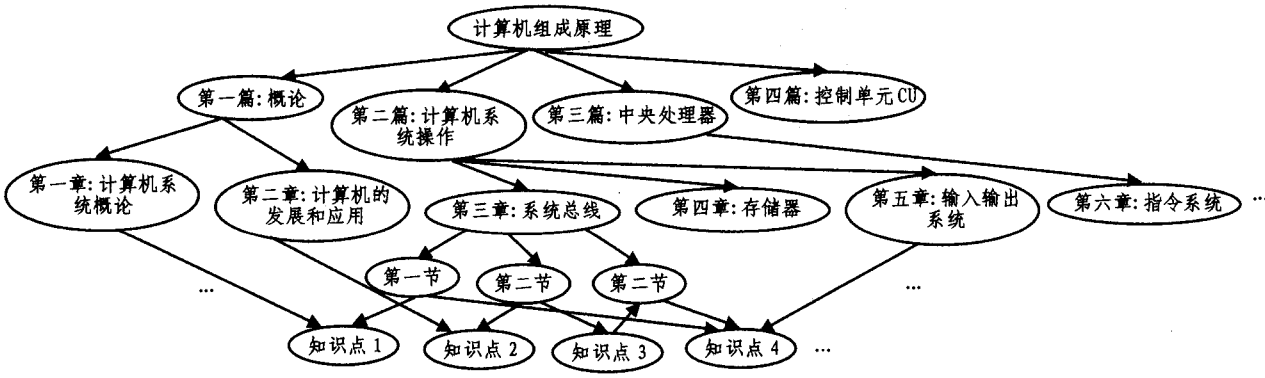

图3 通过知识点组织的课程《计算机组成原理》

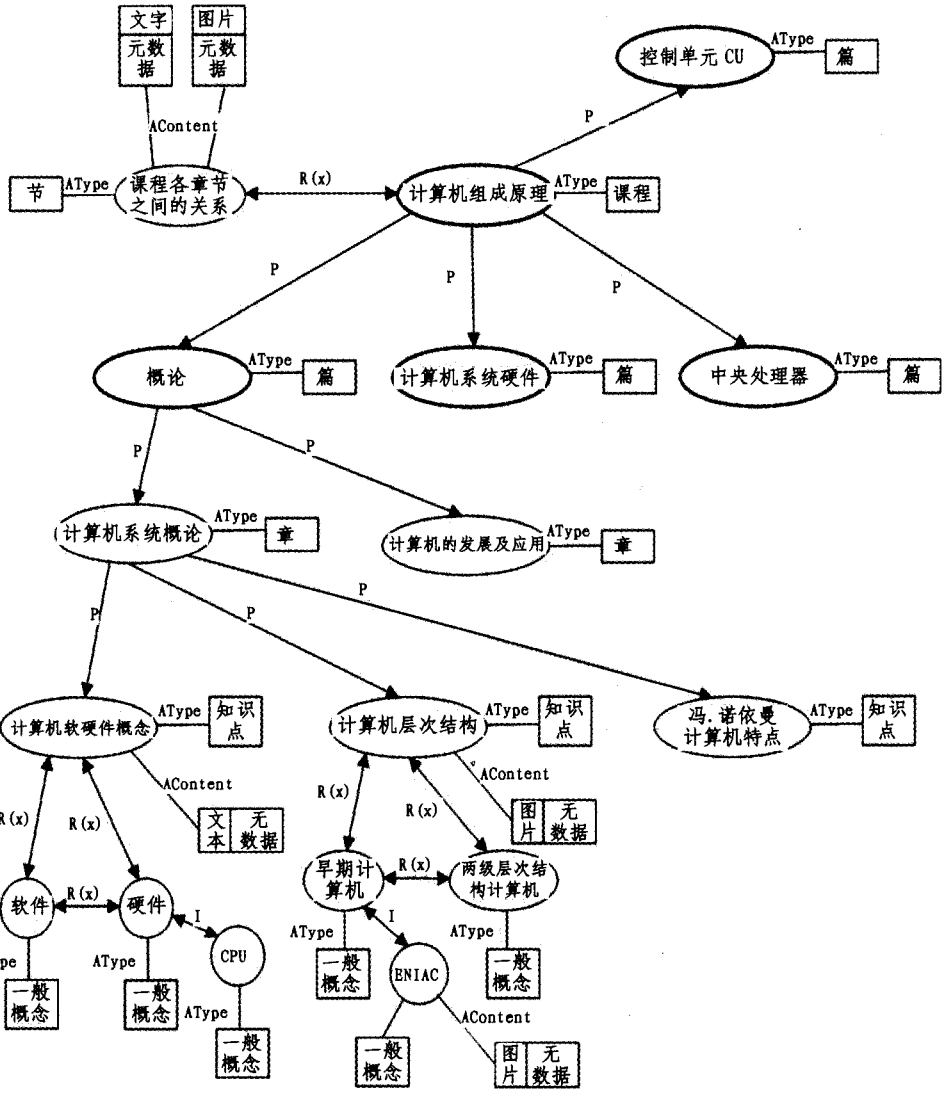

图4 《计算机组成原理》的扩展语义网示意图

4.2 利用扩展语义网进行查询

首先对用户选择的概念进行相关性扩展，具体的扩展范围由系统阈值σ决定，即要求$\sum_{i=1} R(x_i) \leqslant \sigma$。具体过程如下：

(1)用户选择兴趣领域或课程，以及知识点或概念；

(2)系统选择相关的课程，在该课程的语义网中根据用户输入的知识点或概念，基于$R(x)$连接进行相关概念查询；

(3)汇总相关概念，根据这些概念的内容属性AContent，得到它们的具体内容并按照相关性排序后返回用户，用户可以根据返回的结果进一步查询。

4.3 利用扩展语义网生成课件

由于语义网中对概念定义了适应层次属性ALevel(概念)，因此可以据此针对不同的待培训人员层次生成相关课件。

具体步骤：

(1)教师输入课程名(与语义网的课程名相同)和面向的学习者层次；

(下转第185页)

法对 XML 文件分组，就可以为每个聚类创造一种倒排文件。这个倒排文件可将不同片段进一步划分为频繁路径及树的代表，整个树对应一条路径的后缀，路径上各节点的子节点和父节点合并。这能够加快 XML 数据库文件的搜索。

4 实验结果分析

实验数据来自于 SigmodXML 数据集[12]，在实验中，已知 XML 数据集中的文档分属于 4 类(根据 DTD 确定)，每一类中的文档数目也已计算出来。并和 Tag-based 算法[6]、TreeFinder 算法[13]进行了比较。本文采用了通用的查准率和查全率来评价这些聚类结果，并引入一个内部度量标准来考查已发现的聚类簇的紧凑性，即

$$IC(P)=\frac{1}{n}\sum_{C_i\in P}\frac{1}{n}\sum_{d\in C_i}Dist(d,rep(C_i))$$

式中，$P=\{C_1,\cdots,C_i,\cdots,C_n\}$，$P\in D$，$C_i=\{d_i,\cdots,d_m\}$，$rep(C_i)$是$(C_i,C_j)$的聚类代表。

表 1 实验结果

文档	平均大小	DTD	算法	聚类	Precision	Recall	IC
84	3.2k	4	Tag-based	10	0.873	0.983	0.310
			TreeFinder	10	0.965	0.976	0.215
			XPath	10	0.974	0.983	0.331
84	2.9k	4	Tag-based	10	0.804	0.907	0.345
			TreeFinder	10	0.854	0.869	0.298
			XPath	10	0.865	0.968	0.318
84	4.3k	4	Tag-based	10	0.680	0.916	0.384
			TreeFinder	10	0.852	0.785	0.343
			XPath	10	0.856	0.922	0.391

三个文档集组成了测试数据，每个包含 84 个文档(每个类别 21 个文档)。表 1 就表明了该算法区分不同大小文档的准确性。可以看出，Tag-based 算法限制比较宽，聚类的内聚度很低，对较大的 XML 文档效果较好；TreeFinder 算法要求比较严格，聚类的内聚度比较高，但是文档的查全率比较低，对于较小的 XML 文档效果很好。基于 XPath 的算法，则显示出了良好的适应性。

结束语 基于路径的 XML 文档聚类是当前文本挖掘研究中的一个全新领域，由于其处理对象是结构化的 XML 文档，所以具体的聚类方法和一般的文本聚类有着较大差别。特别是 XML 文档的语义信息可借助于 XPath 描述的文档结构得以表示，通过增加文档间相似度比较的准确度和精确度，可以更加便利地操纵 XML 文档为用户的查询和检索提供服务。如何利用聚类得到的路径信息来提高 XML 信息检索效率，是今后继续研究的重点。

参 考 文 献

1 Cheng J, Yu G, Yu J X, et al. An EfficientClustering Based Indexing Method for XML PathExpressions. In: Proc. 8th Int Conf on Database Systems forAdvanced Applications, Kyoto, March, 2003
2 Clark J, DeRose S. XML Path Language (XPath), version 1. 0, 1999. http://www. w3. org/TR/1999/ REC- xpath-19991116
3 Georg G, Christoph K, Reinhard P. Efficient algorithms for processing XPath queries. In: Stéphane B, Akmal B. eds. Proc. of the VLDB 2002. Heidelberg: Springer-Verlag, 2002. 95～106
4 Sihem A, SungRan C, Laks V S. Minimization of tree pattern queries. In: Walid G A, eds. Proc. of the SIGMOD. Santa Barbara, 2001. http://www. research. att. com/～ sihem/publications/SIGMOD01. pdf
5 Frank N, Thomas S. XPath containment in the presence of disjunction, DTDs, and variables. In: Diego C, Maurizio L, eds. Proc. of the ICDE. Heidelberg: Springer-Verlag, 2003. 315～329
6 Guillaume D, Murtagh F. Clustering of XMLdocuments. Computer Physics Cmmunications, 2000, 127(2-3): 215～227
7 Gerome M, Dan S. Containment and equivalence for an XPath fragment. In: Lucian P, ed. Proc. of the PODS. ACM, 2002. 65～76
8 Agrawal R, Srikant R, Mining Sequential Patterns. In: Proceedings of the Eleventh International Conference on Data Engineering, Taipei, Taiwan, March 1995. 3～14
9 Leung H P, Chung F L, Chan S C F. On the use of hierarchical information in sequential mining-based XML document similarity computation. Knowledge and Information Systems, 2005, 7(4)
10 Kaufman L, Rousseeuw P J. Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley, 1990
11 Lee M L, et al. XClust: Clustering XML schemas for Effective Integration. In: Proc. 11th Int Conf on Information & Knowledge Management, McLean, Nov. 2002. 292～299
12 Sigmod XML DataSet. Available at : http :/ / www. acm. org/ sigmod/ record/ xml. 2005 - 7
13 Termier A, Rousset M C, Sebag M. TreeFinder: A First Step Towards XML Data Mining. In: Proc. of the 2002 IEEE Int Conf on Data Mining, Maebashi TERRSA, Maebashi City, Japan, December 2002. 450～257
14 潘有能. XML 文档自动聚类研究. 情报学报, 2006, 25(2): 215～220

(上接第 141 页)

(2)系统根据课程名找到该课程概念，然后根据部分关系 P(概念 1，概念 2)和适应层次属性 ALevel(概念)在语义网中扩展，直至与知识点连接的概念，结束扩展；

(3)根据导出结构生成课件，在课件中可以同时列出每个知识点的相关概念链接；

(4)教师对生成的课件进行增、删、改操作，得出最终稿。

结论 通过应用案例，按照知识点进行扩展的语义网提供了一整套描述知识资源、知识资源查找和定位以及自动生成知识资源应用案例的方法，增强了知识资源整体关系的描述能力，查找和定位知识资源也更加方便，知识资源应用案例的生成也更加简易、灵活。若将其应用于各类组织(企业、院校及科研结构等)的知识管理系统，必将提升人们知识获取、分享、分配和存取的能力，为知识的理解和利用提供了一种有效的途径。

参 考 文 献

1 王晓蓉. 知识管理中的语义网方法[J]. 情报技术, 2004, 5(24): 65～66
2 沈军, 顾冠群. 面向网络教学的互动式体系模型[J]. 东南大学学报, 2002, 32: 6～10
3 Berners-Lee T. Weaving the Web [M]. San Francisco, CA: Harper, 1999
4 Berners-Lee T, Hendler J, Lassila O. The Semantic Web [J]. Scientific American, 2001, 284(5): 34～43
5 Berners-Lee T, Hendler J. Publishing on the semantic Web [J]. Nature 410, 2001, 26(6): 1023～1024
6 刘柏嵩. 基于知识的语义网：概念、技术及挑战[J]. 中国图书馆学报(双月刊), 2003, 2(3): 18～21
7 Staab S. Ontologies' KISSES in standardization [J]. IEEE Intelligent Systems, 2002, 6(24): 70～79
8 曲敏, 冯志勇. 基于 OWL ontology 的制造业知识管理[J]. 制造业自动化, 2006, 28(1): 8～12
9 张屹, 祝智庭. 知识管理在现代远程教育中的应用研究[J]. 中国远程教育, 2003, 3(182): 17