维普资讯 http://www.cqvip.com

一种有效验证 XML 文档语义的验证方法*)

韩 婷[1,2] 李陶深[1]

(广西大学计算机与电子信息学院 南宁 530004)[1] (上海财经大学信息中心 上海 200433)[2]

摘 要 本文针对传统的文本类型定义(DTD)对 XML 文档验证存在的漏洞,将本体论思想引入 DTD 文档,提出一种基于本体论思想的 XML 文档语义验证的有效方法,并在框架逻辑上予以实现。分析和实验表明,提出的方法可进一步提高信息获取的准确率。

关键词 XML,DTD,本体论,框架逻辑,语义有效

1 引言

可扩展标记语言 XML 已成为万维网上信息交换的一种新的标准[1],它允许用户定义自己喜欢的标签,且文档的书写无语法要求,具有极大的灵活性。然而,可扩展性既成为了 XML 的优点,但是也成为了它自身最大的缺陷,因为可扩展性会使服务器无法理解用户自定义的标签。虽然传统的文本类型定义(DTD)对文档的结构约束给予了强大的支持,然而对文档的语义理解却没有任何帮助。为此,人们都在努力研究针对 XML 文档的验证机制,例如 XML Schema 就提供了更多的数据类型并支持名称空间,目前被认为最有前途的 RDF(Resource Description Framework)则采用 URI 准确地定位网络中的资源。这些机制都试图为文档中的元素或资源寻找归宿,然而却仍然无法在文档内部构建元素或资源间的语义关联。哲学上的本体论(Ontology)思想给这一问题的解决带来了契机[2]。

哲学中的本体论是对一种存在的系统化解释,人们将本体论的含义用到计算机领域中,赋予本体论更具体的意义。在计算机领域中,本体论是采用某种语言对概念化对象(Conceptualization)的明确表示和描述[3]。因此,本体论依赖于所采用的语言,按照表示和描述的形式化程度不同,可以分为完全非形式化的、半形式化的和严格形式化。本体论形式化程度越高,越有利于计算机进行自动处理。从概念化对象的定义来看,一个领域中的术语、术语的定义以及术语之间的语义网络应是一个领域本体应包含的基本信息。

本体论已经在基于网络的知识处理、共享,以及软件重用方面扮演关键角色[4]。如果用本体论定义某个特定领域内共享的概念层次,可以为人与应用系统间提供用于通信的简单且易于理解的主题。本文针对传统的 DTD 在 XML 文档验证中存在的漏洞,提出一种基于本体论思想的验证方法,并在框架逻辑上予以实现,使语义有效的 XML 文档成为可能。

2 基于框架逻辑的本体论描述

人工智能领域用语义网络和框架逻辑来描述概念间的层次关系和关联关系[4],前者采用图例的方式描述概念间关联的不同视图,后者是关于个体类的结构化表示法。考虑到与具有层次结构的 DTD 文档形成对应关系,我们用框架逻辑描述领域内的本体论。下面给出的 DTD 文档实例(example. dtd)是一个基于框架逻辑的本体论描述,它描述了一个人和企业域,内容分为三个部分,包括该域的概念层次,概念之间的关系以及一些公理规则。框架第一部分描述了概念在其领域内的层次位置,"::"表示一种所属关系,表示右边是左边的上位概念;第二部分描述了概念间的联合,将概念引入了属性的定义中,"[ ]"中定义概念的属性,=≫说明属性的数据类型定义;第三部分描述了概念实例间公认的规则或者关系,"<->"表示了这种关系。

(example. dtd)

```
域[ ].
  人 :: 域.
    雇员 :: 人.
      技术人员 :: 雇员.
        程序员 :: 技术人员.
          本科生 :: 程序员.
    学生 :: 人.
      本科生 :: 学生.
    企业 :: 域.
      国有企业 :: 企业.
      外资企业 :: 企业.
```

*)基金项目:广西"新世纪十百千人才工程"专项基金项目(桂人字 2001213 号)。

维普资讯 http://www.cqvip.com

IT 公司 :: 外资企业
合资企业 :: 企业.
人[姓名 =>> STRING; email =>> STRING;
单位 =>> 企业; 住址 =>> STRING].
雇员[雇员 ID =>> 本科生].
技术人员[监督 =>> 程序员].
程序员[合作 =>> 程序员].
本科生[主管 =>> 技术人员].
学生[学生 ID =>> NUM].
企业[法人 =>> 人; 名称 =>>STRING; 性质
=>> STRING; 概要 =>>STRING].
外资企业[联系人 =>> 人; 企业 ID =>>
NUM; 员工数目 =>> NUM; 行业 =>>
IT 公司].
IT 公司[性质 =>> 企业].
FORALL 张三, 李四
张三:程序员[合作 ->> 李四] <->
李四:程序员[合作 ->> 张三].
FORALL 张三, 某企业
某企业:企业[法人 ->> 张三] <->
张三:人[单位 ->> 某企业].
FORALL 张三, 某外资企业
某外资企业:外资企业[联系人 ->> 张三] <->
张三:人[单位 ->> 某企业].
FORALL 张三, 李四
张三:本科生[主管 ->> 李四] <->
李四:技术人员[监督 ->> 张三].
FORALL 企业 1, 企业 2
企业 2:外资企业[行业 ->> 企业 1] <->
企业 1:IT 公司[性质 ->> 企业 2].

3 DTD 语义验证的设计思想

example. dtd 展现了本体论思想在描述概念层次上的优势,利用这个优势可将两个看起来似毫不相干的概念通过它们的上位概念间的联系关联起来,从而在更高层次上把握语义。

本体论描述的实质就是表明一种概念间的继承关系,若一个概念拥有上位概念,那么它理所当然继承了其上位概念的属性。也可以做这样的推论,如果两个概念具有一些相同的属性,那么它们极有可能拥有同样的上位概念。如此一来,概念就被约束在与自己具有语义交集的链上,不管概念用何种表述方法,只要语义相同就一定能被找到。然而,是否需要定义一种新的语言来对本体论思想进行形式化描述呢? 这个任务显得相当庞大。考虑到传统的 DTD 已经拥有强大的文档结构有效性验证功能,如果能用 DTD 来对本体论进行形式化描述,从而把本体论的思想引入 DTD 中,那将是一种最有效、快捷的途径。本文就是采用了这种方法,图 1 给出了实现这种方法所要完成的任务的框架描述。

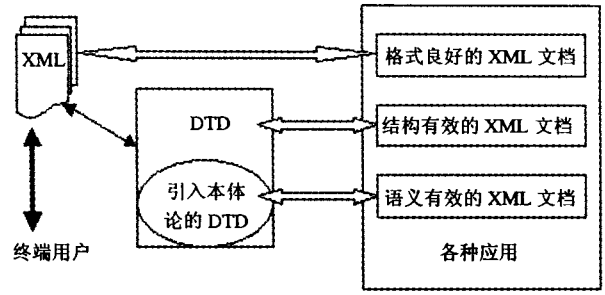

图 1 实现 DTD 语义验证的任务的框架描述

在将本体论思想引入 DTD 之前,首先要解决一个问题,即在 DTD 中用什么来定义本体。将每个本体都定义为一个元素也许是人们最先想到的方法,然而实际的情况是概念往往出现在元素的属性定义中,或者更本质地说某些概念包含了用概念定义的属性,因此仅简单地将每个本体定义为一个元素是不够的。为此,我们考虑使用这样的方法,即在用元素定义本体概念的同时,也将所定义的概念引入属性的概念定义中去,从而寻找它的语义根源。这种思想类似面向对象思想中类的定义过程,即概念的继承关系是在它被定义的时候产生的,是与生俱来的。

那么,在 DTD 中如何实现概念间的继承关系呢? 因为我们所要把握的是文档的语义内容,而这些内容只有在元素或属性的值中被定义,所以问题就转化为值的定义。DTD 中用 10 种内定的数据类型和一种实体声明的方法来对值进行定义,其中实体声明让值的定义变得比较灵活,它可以用来定义可重用的数据块或是引用非 XML 数据,以简化 DTD 和增强可读性。然而应该注意到,实体是一个代表内容的占位符,它具有天生的替代能力,而且替代的内容可以是有意义的某个短语或者概念。本文的着眼点是开发并强调 DTD 实体声明的这种强大的潜在能力,并将它应用到包含了用概念定义的属性定义中去,实现概念在本体论意义上的关联。在这个意义上,实体作为本体元素的优良载体在概念之间传递共有信息,从而产生含义更加丰富的本体元素,实现了概念的进化。当然要想实现概念间的继承关系,实体传递的概念与被定义的概念在本体论中必须处在同一条语义链中。

本文提出的语义验证方法的特别之处在于:实体声明的是已定义的本体元素,然后用它再去定义与之语义相关的元素,形成新的本体,如〈! ELEMENT branch (# PCDATA | % tree;) * 〉中用已定义的本体 tree 定义了新的本体 branch,从而向

维普资讯 http://www.cqvip.com

branch 传递了自己的属性。如此将所有的概念进行关联，形成概念间的语义链。可见，整个文档的描述过程其实就是本体论的构造过程。

4 语义有效的 DTD 实例

基于上述的设计思想，我们做了相应的研究与实验，下面是实验的部分代码和说明。在这些转化后的 DTD 文档中，example. dtd 的本体论描述的三个部分也同样可以表现在 DTD 文档中。

```
〈! --entities for realizing the is-a hierarchy --〉
〈! ENTITY % 人 "人 | 雇员 | 学生 | 技术人员 | 程序员 | 本科生" 〉
〈! ENTITY % 程序员 "程序员 | 本科生" 〉
〈! ENTITY % 企业 " 企业 | 国有企业 | 外资企业 | 合资企业 | IT 公司" 〉
〈! ENTITY % 外资企业 "外资企业 | IT 公司"〉
〈! -- element declarations for ontology concepts --〉
〈! ELEMENT 人 (# PCDATA | 姓名 | email | 单位 | 住址) * 〉
〈! ELEMENT 程序员 (#PCDATA | 姓名 | email | 单位 | 住址 | 雇员 ID | 监督 | 合作) * 〉
〈! ELEMENT 企业 (#PCDATA | 法人 | 名称 | 性质 | 概要) * 〉
〈! ELEMENT 外资企业 (#PCDATA | 法人 | 名称 | 性质 | 概要 | 联系人 | 企业 ID | 员工数目 | 行业) * 〉
〈! ELEMENT IT 公司 (#PCDATA | 法人 | 名称 | 性质 | 概要 |联系人 | 企业 ID | 员工数目 | 行业 | 性质) * 〉
〈! -- ATTLIST declatation for ontology attributes --〉
〈! ATTLIST 人
  姓名 CDATA #IMPLIED
  email CDATA #IMPLIED
  单位 CDATA #IMPLIED
  住址 CDATA #IMPLIED〉
〈! -- element declaration for ontology attributes --〉
〈! ELEMENT 法人 (#PCDATA | %人;) * 〉
〈! ELEMENT 联系人 (#PCDATA | %人;) * 〉
〈! ELEMENT 住址 (#PCDATA) 〉
〈! ELEMENT 单位 (#PCDATA | %企业;) * 〉
〈! ELEMENT 合作 (#PCDATA | %程序员;) * 〉
```

该 DTD 文档的第一部分是实体声明部分，它将本体论中概念层次间的继承关系转化为概念间的替代关系；第二部分是上位本体元素的值与属性定义部分，它们与实体声明中声明的实体对应，属于框架逻辑的最外层；第三部分真正实现了概念的继承，它用已定义的本体元素的实体声明定义了新的本体，从而使新的本体继承了已定义本体在第二部分中定义的值和属性，新的本体就成为已定义本体的下位概念。

尽管上面的 DTD 只是该领域描述的一个子集，但是也可以清楚地看到：通过参数实体的声明，下位概念很好地继承了上位概念的属性，既涵盖了其所属领域的共性又不失其本身的个性，从而为抽取概念间的共有信息或通过某种信息寻找拥有它的概念提供有效的途径。应该指出的是，DTD 文法中并不能识别中文，上面例子只是为了方便说明起见而采用了中文，在实际应用中应采用有意义的英文标识。

结束语 本文通过将本体论思想引入 DTD 文档，实现了对 XML 文档语义的有效验证。实验表明，它大大提高了信息获取的准确率。最简单地，在 example1. dtd 中，若对第 2 节中的查询语句稍做如下修改就可得到理想的结果。

```
〈! ENTITY % 名称 "名称 | 技术名称" 〉
… … …
〈! ELEMENT 合作伙伴 (#PCDATA | %名称;) * 〉
```

另外，用 DTD 描述的本体论具有很好的可维护性。新的本体的引入是在它的元素定义中完成的，并不影响已定义的本体元素，即新的本体只是简单的链入语义链中，链入点即是与它直接语义相关的上位概念。此后，由它又可以定义属于它自己的后代分支，扩展语义链。当然，本体论与 XML 模式的结合决不仅限于 DTD，但是 DTD 的简单和在描述结构化数据方面的特长不失为本体论提供了一个最简约和容易实现的承载体。随着 XML 相关规范的不断完善，相信本体论与 XML 模式的结合将不断深入，对基于内容的 XML 文档的语义理解也将不断深化。

参 考 文 献

1 硕网资讯编著. 洞悉 XML [M]. 北京：北京大学出版社，2001
2 Desmontils E, Jacquin C, Simon L. Ontology enrichment and indexing process [R]: [RESEARCH REPORT]. 2003, No 03. 05 Mai 2003
3 Chandrasekaran B, Josephson J R, Benjamins V R. What Are Ontologies, and Why Do We Need Them? [J]. IEEE Intelligent Systems, 1999, 1094-7167: 20～26
4 Hori K. An ontology of strategic knowledge: key concepts and applications [J]. Knowledge-Based Systems, 2003, 13: 369～374
5 廖明宏. 本体论与信息检索[J]. 计算机工程，2000, 26(2): 56～58
6 万捷，滕至阳. 本体论在基于内容信息检索中的应用[J]. 计算机工程，2003, 29(4): 122～123, 152