文章编号:1000-5013(2014)04-0387-05 **doi**:10.11830/ISSN.1000-5013.2014.04.0387

一组相关XML数据文件的数据类识别方法

李赛男,余金山

(华侨大学 计算机科学与计算学院,福建 厦门 361021)

摘要: 为解决当前可扩展标记语言(XML)绑定框架普遍存在的由XML模式映射生成的数据类的冗余,以及数据类系统规模过大的问题,提出一种从一组相关XML数据文件的数据实体类识别方法.该方法先抽取这一组XML数据文件的XML模式树图,并将其每个节点表示成向量空间中的向量;然后,利用相似度和距离识别该模式节点对应的预定义模式节点类型;最后,按模式节点类型到类的映射规则得到数据类.结果表明:该方法能识别合并对应同一个实体的数据类避免类冗余,将集合类型的XML文件映射成泛型类和集合类减小生成类系统的规模.

关键词: 类识别;可扩展标记语言;数据绑定;模式树图;节点类型;相似度

中图分类号: TP 311 **文献标志码**: A

可扩展标记语言(extensible markup language,XML)数据绑定是指将数据从一些XML文件中取出,通过程序表示这些数据的过程.即把数据绑定到计算机能够理解且可以操作的某种内存结构中[1],大多数是绑定到类对象实例上.XML数据绑定隐藏了XML数据的具体结构,方便程序直接使用XML文档中的数据内容,使得XML数据能够直接转换为可处理的业务数据[2].目前,将XML数据绑定到Java对象的框架有XStream,JAXB,XMLBeans,Castor和JiBX[2-6].这些XML绑定框架的关键是XML数据对应数据类的获得,即根据XML模式文件按一定规则生成的,或用户自己编写绑定映射规则再字节码生成.它们可以很好地对遵循同个XML模式文件约束的一组XML文件进行XML数据绑定.但要解决来自相同应用系统中,遵循不同XML模式文件的XML文件的数据绑定,会有数据类冗余、生成类系统庞大等问题.基于此,本文提出一种从一组相关XML数据文件的数据类识别方法.

1 XML结构特点及相关定义

1.1 XML结构特点

每个XML文档有且仅有一个根元素,它是所有其他元素的父元素,而所有元素都可拥有子元素、文本内容和属性.从元素的嵌套关系可以看出XML文档是一棵文档树.相应地有一种抽象数据结构——文档对象模型(document object model,DOM).DOM将XML文档中各种类型的数据映射到相应的类型对象,构建出树形结构[7],分成文档节点、元素节点、文本节点、属性节点、处理指令节点、注释节点、文档类型节点、文档段节点、符号节点、CDATA段节点、实体节点和实体引用节点等12类.

XML文档可以分为以数据为中心和以文档为中心两大类.以数据为中心的XML文档常被用于机器的使用,而以文档为中心的XML文档则主要是为人类而设计的.文中涉及的XML数据文档均属于以数据为中心的XML文档.它着重于文档中的数据,而非文档格式.具有结构化的数据、数据粒度大小适中、很少或没有混合内容,以及文档顺序不重要等特点.故在考虑设计实现问题解决方案时,需要考虑的DOM节点对象只有属性节点、元素节点和文本节点.

1.2 XML模式节点定义

XML文档模式描述一类XML文档中数据的结构关系和类型信息,在内存可表现为一个树图.树

收稿日期: 2013-02-17

通信作者: 余金山(1952-),男,教授,主要从事软件工程和人工智能应用等方面的研究.E-mail:yjs@hqu.edu.cn.

节点对应元素节点、属性节点，节点的附加信息有元素或属性的名称、重数、数据类型等. 将模式树图中的模式节点进行分类，以建立 XML 模式到数据类之间的映射，并定义如下 6 个模式节点类型.

1）属性型. 该模式节点对应 XML 文档中的属性节点.

2）属性类元素型. 该模式节点对应 XML 文档中的元素节点，而该元素节点没有子元素节点或只包含文本节点，没有属性.

3）List 类元素型. 该模式节点仅包含一个子模式节点，而该子模式节点是非属性类节点且对应元素节点的重数大于 1.

4）包装类元素型. 该模式节点包含多个属性类节点和一个非属性类节点的子模式节点. 在多 XML 文件环境下，包装类元素型的模式节点有两种类型：一种是包含一个非属性类型的子模式节点，但具有多套不同的属性类型的子模式节点组；另外一种是只有一个属性类型的子模式节点组，但包装多套非属性类型的子模式节点.

5）包装类属性集型. 用于组织包装类元素型模式节点下的属性类型的模式节点.

6）实体类元素型. 该分类设置为非属性类型的模式节点类型的默认值，除前面 5 种类型节点外，剩下的模式节点就属于实体类元素节点. 实体类元素型模式节点中有一类节点具有明显特征，即该模式节点有子模式节点且这些子模式节点均属于属性类型.

上述 6 类模式节点类型中，属性型和属性类元素型统称属性类型. 至此，模式节点中包含的信息可以确定有：模式节点类型、模式节点名称、数据类型、映射到类的属性名称、重数、子模式节点列表、所属 XML 文件标识名、根节点标识、文本标识等.

2　XML 模式抽取

文中描述的是从一组相关 XML 数据文件识别实体类的方法，其总体思路是先得到 XML 文档的数据模式，再将 XML 数据模式映射到实体类. 方法的第一步是得到 XML 文档的数据模式，即为每个 XML 数据文件建立一个 XML 模式树图，并填充模式树图中每个模式节点的信息. 在抽取 XML 模式时，需要判断模式树图中的模式节点的类型. 预定义的 6 种模式节点类型中，属性型、属性类元素型、List 类元素型根据其说明的特征可以很容易判断出来，包装类属性集型不需要判断，而是最后再生成并填进模式树图的.

第一种类型包装型模式节点具有一个非属性类型和多套不同属性类型的子模式节点组. 对于这种类型，模式节点下的属性类型的子模式节点的相似度超过阈值 A 则为实体类元素型，否则为包装类元素型. 第二种类型包装类元素型模式节点具有一个属性类型的子模式节点组和多套单个非属性类型的子模式节点. 对于这种类型，模式节点下的属性类型的子模式节点的相似度超过阈值 B，并且非属性类型的子模式节点的相似度低于阈值 C 则为包装类元素型，否则为实体类元素型.

模式节点的相似度是利用空间向量模型 VSM[8] 将其表示成向量后再计算，具体计算方法如下：1）设在一组 XML 文件的模式树图中有 n 个同名的模式节点 N，分别记为 $N_1, N_2, \cdots, N_n$，$N_i (1 \leqslant i \leqslant n)$ 下的子模式节点记为 $[M_{i,1}, M_{i,2}, \cdots, M_{i,s_i}]$，其中 s_i 为 N_i 的子节点的个数；2）将所有的模式节点 $N_1, N_2, \cdots, N_n$ 的子模式节点合并，去除已重复名称的子模式节点，得到结果 Base$[M_{1,1}, \cdots, M_{n,s_n}]$. 将 Base 作为基，将模式节点 N_i 表示成由 0，1 组成的向量 $\boldsymbol{V}_i$. 如果 N_i 包含 $M_{i,j}$，则在 $\boldsymbol{V}_i$ 的 $M_{i,j}$ 位置上标为 1，否则标为 0. 模式节点 N 的相似度使用 $\boldsymbol{V}_i$ 中两两之间的 Jaccard 相似系数 $= \dfrac{f_{1,1}}{f_{1,1}+f_{1,0}+f_{0,1}}$ 的平均值来计算[9]. 具体操作时，还需将 Base 分成属性类节点的 Base 和非属性类节点的 Base，用以计算属性类型和非属性类型的子模式节点相似度.

在同一个应用系统环境下，从多个 XML 文档生成的多个 XML 模式树图中会有相同名称的实体类元素型模式节点. 因此，可利用它们的子模式节点的余弦距离进行基于密度聚类，把属于同一个聚类簇的模式节点的子模式节点合并当做同一个模式节点进行处理，以此避免模式节点映射成数据类时生成冗余的数据类.

对 XML 数据文档进行模式抽取，有如下 5 个主要步骤.

步骤 1 把 XML 数据转换为 DOM 树.

步骤 2 先根遍历 DOM 树,从中获得相关模式信息,创建相应类型的模式节点,并构建初步的模式树图,具体流程如图 1 所示.输入 XML 数据文档的 DOM 元素节点,输出结果是相应的模式节点,且能判断出相应的节点类型.最先获得元素节点的名称用于设置模式节点名,再判断元素节点是否符合属性元素型模式节点的特征,即元素节点没有属性且没包含子元素.符合特征,则根据元素文本内容判断其数据类型用于设置模式节点的数据类型,并返回属性元素型模式节点.处理元素节点及其属性列表创建的默认类型的模式节点,并设置模式节点的重数信息和文本标识等.最后再判断当前模式节点是否仅包含一个子模式节点,该子模式节点是非属性类节点且对应的元素节点的重数大于 1,如果是,则将节点类型设置为 List 类元素型.

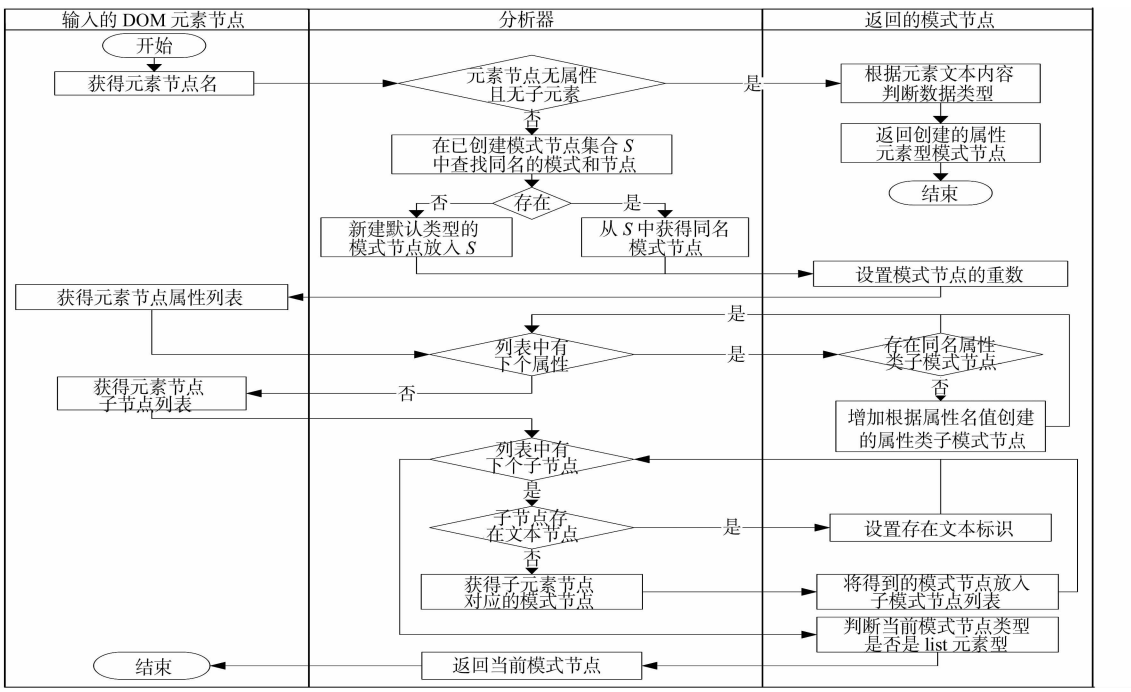

图 1 初步构建模式树图的流程

Fig. 1 Process of the preliminary constructing pattern tree diagram

步骤 3 对分布在不同 XML 模式树图中的同名模式节点进行相似度分析,识别包装类元素型和实体类元素型,并对实体类元素型模式节点进行聚类来完成合并工作,具体流程如图 2 所示.先进行一个快速优化的判断,即如果输入的模式节点列表只有一个模式节点,且该模式节点只包含一个多重的默认类型的子模式节点或一个 List 类元素型子模式节点,则把该模式节点的类型设置为包装类元素型后结束;否则,接着根据输入的模式节点列表建立属性和非属性类节点的 Base,以此计算出模式节点列表中每个模式节点的属性类和非属性类节点向量.如果非属性节点列表为空,则说明该模式节点一定是实体类型;否则,计算属性类和非属性类节点向量列表的相似度 A 和 E.根据两次条件判断模式节点是否属于包装类元素型,如果是,则将模式节点设置为包装类元素型;不是,则将属性类和非属性类节点向量合并得到模式节点向量列表,再对模式节点向量列表使用余弦距离进行聚类.把聚类得到聚类簇中模式节点的子节点合并成新的子模式节点列表,并替换簇中模式节点的子模式节点列表.另外,为区别同名的不同实体类节点的模式节点,在模式节点名称后加上序列数 1,2,….

步骤 4 从所有的模式树图中识别创建不重复的包装类属性集型模式节点.将所有模式树图中包装类元素型模式节点下的属性类子节点打包到一个新建的包装类属性集型模式节点的子模式节点列表中,再将此新建的模式节点替代原来的属性类子节点,插入到包装类元素型模式节点下作为新子模式节点.包装类属性集型模式节点是可以重用的,但必须保证包装类元素型模式节点下的每个属性类型模式节点的名称和重数是相同的.

步骤 5 后根遍历所有模式树图,填充模式节点中剩余的模式信息.包括映射到类的属性名称的设

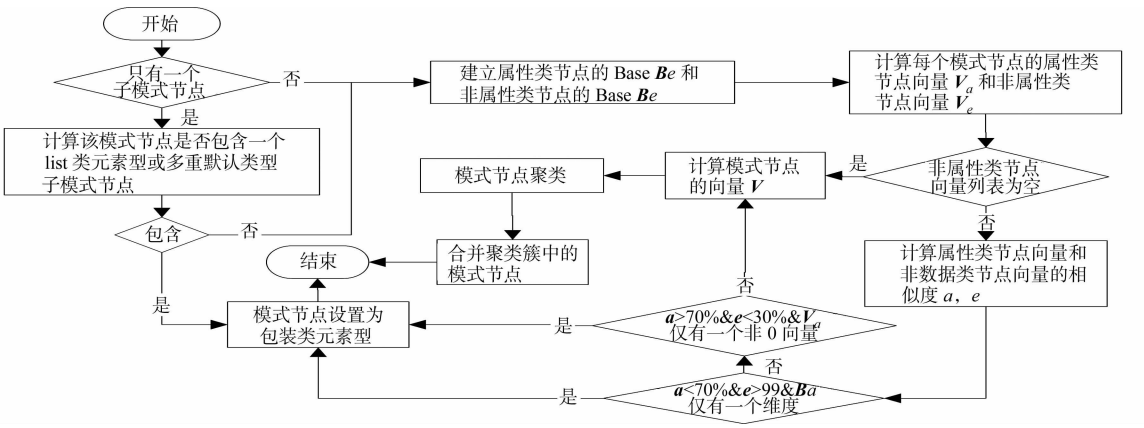

图 2　模式树图的模式节点的进一步分类流程

Fig. 2　Further classification process of the mode pattern tree node

置，数据类型的设置. List 类元素型和包装类元素型模式节点需要添加子模式节点的数据类型添加泛型信息，包装类属性集型和实体类元素型需要确定映射的数据类的类名.

将一组相关 XML 数据文档经过 5 个步骤处理后得到完整的 XML 模式树图. 图中的每个模式节点包含了完整的模式信息，作为 XML 数据模式映射到实体类的输入数据.

3　XML 模式树图到类的映射

从一组 XML 数据文件中识别数据类的最后步骤是，将 XML 模式树图中的模式节点按照一定规则映射到数据类. 把识别出的模式节点特征信息存储在模式节点的类型信息中，在 XML 模型到类层次的映射过程会根据模式节点类型，将模式节点映射成不同类型的数据类.

映射规则有如下 8 个方面：1）模式节点的模式节点名称映射成数据类的类名；2）模式节点的子模式节点映射成数据类的成员变量，该子模式节点映射到类字段名称映射成成员变量名，子模式节点的数据类型映射成成员变量的数据类型；3）如果是模式节点的文本标识，则新增一个名为_value，类型为字符串的成员变量；4）如果子模式节点的重数大于 1，则需将用集合类类名包装原数据类型作为其数据类型；5）若成员变量的数据类型是泛型，则还需在该成员变量之前添加注解的配置信息说明具体的泛型信息；6）给每个数据类的成员变量生成 setter 和 getter；7）将包装类属性集节点类型、实体类元素节点类型和包装类元素节点类型的模式节点分别生成放在 vo. attribute 包中的属性类、放在 vo. entity 包中的实体类和放在 vo. wrapper 包中的包装类；8）包装类元素节点类型的模式节点映射成的包装类是预先定义的泛型类 ListWrapper〈S，T〉和 SingleWrapper〈S，T〉. 这 2 个包装类的只有 2 个成员变量分别对应包装类属性集节点类型的子模式节点和非属性类节点类型的子模式节点.

经过上述映射规则，得到数据类从一组 XML 数据文件中识别出的数据类. XML 模式到类层次映射方法和目前 XML 绑定框架的模式编译器生成类文件的映射规则不同的是，还要根据模式节点的类型进一步区分映射成不同类型的类，如集合类、泛型类、实体类等，以减小生成类系统的规模.

4　应用实例

将提出的一组相关 XML 数据识别方法应用在一个 Last. FM OpenAPI 的数据实体类代码生成器的实现中. 该代码生成器先从网络上抓取每个 Last. FM OpenAPI 返回的 XML 数据，得到一组相关 XML 数据文件；然后，使用文中方法获得识别出的数据类的元数据；最后，在结合类模板输出类定义代码源文件. 该代码生成器从网络上获取的 Last. FM OpenAPI 返回的 XML 文件有 131 个，使用文中方法输入预定义的阈值 A、阈值 B、阈值 C 是经验估计值，分别是 70%，70%，30%. 经过计算，最终识别出 47 个实体类，38 个属性类，总共 85 个类. 如果使用传统的数据绑定框架对 131 个 XML 文件进行数据类生成，最终得到的数据类最少有 131 个. 使用文中方法生成类的个数比之前的方法降低 35%，而且生成的实体类和实体一一对应没有存在冗余. 该代码生成器的结果说明，文中提出的方法可以较好地解决

当前XML绑定框架类中生成器产生的类冗余和生成类系统规模过大的问题.

5 结束语

提出从一组相关XML数据文件的数据实体类识别的方法.该方法能有效解决当前XML绑定框架在绑定一组来自同一个应用框架下,遵循多个XML模式文件的XML文件时产生的类冗余和生成类系统规模过大的问题.但该方法存在的不足是,区别包装类元素型和实体类元素型的模式节点中的相似度阈值是预先给定的,后续工作需采用启发式及回归模型[10]对相似度阈值进行调整确定.

参考文献:

[1] MCLAUGHLIN B. Practical data binding: Get your feet wet in the real world [EB/OL]. [2004-05-20]. http://www.ibm.com/developerworks/xml/library/x-pracdb1/index.html.

[2] 焦春芳,罗晓沛.基于Castor的数据绑定技术[J].计算机工程与设计,2008,29(17):4550-4553.

[3] 吴翔,饶若楠.连接XML与对象的桥梁——XMLBean[J].计算机工程,2004,30(增刊1):69-71.

[4] BANGALORE R. Use XStream to serialize java objects into XML[EB/OL]. [2008-07-23]. http://www.ibm.com/developerworks/xml/library/x-xstream/.

[5] SIMEONI F, LIEVENS D, CONNOR R, et al. Language bindings to XML[J]. IEEE Internet Computing, 2003, 7(1):19-27.

[6] 许晖.应用XML实现Java对象序列化技术简述[C]//第七届中国Java技术及应用交流大会文集.北京:[s.n.], 2004:73-78.

[7] 李青山,陈平.对象层次上的XML数据绑定模型的研究[J].西安电子科技大学学报,2001,28(6):768-771.

[8] 庞剑锋,卜东波,白硕.基于向量空间模型的文本自动分类系统的研究与实现[J].计算机应用研究,2001,18(9):23-26.

[9] 李桂林,陈晓云.关于聚类分析中相似度的讨论[J].计算机工程与应用,2004(31):64-66.

[10] ZHANG Yi, CALIAN J. Maximum likelihood estimation for filtering thresholds[C]//Proc of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2001:294-302.

Class Identification Method a Group Related of XML Data File

LI Sai-nan, YU Jin-shan

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract: In order to solve the XML binding framework ubiquitous redundant classes generated by the XML schema mapping and data class system large scale, we presented a method of data entity class recognition from a group of related XML data file. The method first extracted XML mode tree a group of XML data files, and each node was represented as a vector in the vector space. Then used their similarity and distance to identify the mode node corresponded to a predefined mode node types. Finally by according to the mode node type to the class mapping rules to obtain the data classes. The results showed that: this method can identify and merger the class that correspond to the same entity to avoid redundant, mapped the collection of XML documents into a generic class and collection class to reduce the size of the generate class system.

Keywords: class recognition; extensible markup language; data binding; pattern tree diagram; node type; similarity

(**责任编辑**:黄晓楠 **英文审校**:吴逢铁)