第 35 卷 第 8 期 计 算 机 工 程 2009 年 4 月
Vol.35 No.8 Computer Engineering April 2009
·软件技术与数据库· 文章编号：1000—3428(2009)08—0065—03 文献标识码：A 中图分类号：TP393

基于 XML 的信息系统集成的可视化与匹配

施 莹，蔡 勇
(江南大学信息工程学院，无锡 214122)

摘 要：近年来，信息架构的匹配显示了在诸如数据库集成等领域的重大应用。针对这一具有挑战性的问题，提出一个可视化匹配的模型，并开发一个半自动的集成工具。此模型主要用于基于 XML 的信息系统中的数据集成。该模型把一个新颖的可视化界面和以算法为基础的推荐引擎结合在一起。经过用户的实际操作和使用，证明该模型对于基于 XML 的信息系统中的数据集成既简单又快速。

关键词：集成；可视化；推荐引擎

Visualization and Mapping for Information System Integration Based on XML

SHI Ying, CAI Yong
(School of Information Engineering, Jiangnan University, Wuxi 214122)

【Abstract】Schema mapping is applied in many important areas like database integration in recent years. With schema mapper, a semi-automatic tool is demonstrated for schema integration that combines a novel visual interface with an algorithm-based recommendation engine to cope with such a challenging problem. Formative evaluation and user's study suggest that schema mapper may be usefully employed for performing schema mapping based on XML efficiently and quickly.

【Key words】integration; visualization; recommendation engine

1 概述

不同信息系统由于缺乏统一的规划和管理，存储和描述信息的架构不一样，格式也不一样，同类型的信息系统之间采用不用的语意来构建自己的信息架构，这些由不同语意规范的信息系统就像一个个“信息孤岛”。信息的发展日新月异，每天都在以惊人的速度增长。而信息爆炸的背后，人们面临的巨大的挑战是如何来共享这些信息。要共享信息，必须把不同的信息系统中的信息集成起来，提供给用户一个统一的信息系统。因此，希望能有一个可视化的匹配工具来帮助实现信息系统中数据的集成。

如今，XML 技术已成为网络间信息表示和交换的标准格式，并逐渐成为许多网络资料的存储方式。但是由于语意上的差别，不同的信息系统之间制定的 XML Schema 有差异，缺乏一个统一的标准。因此，不同的信息系统之间要相互协同工作变得尤其困难。为了集成数据，必须把描述同一种元素的不同的 XML Schema[1]进行匹配。

2 处理流程

针对不同信息系统的信息架构都有着各自的特点，信息系统集成的可视化匹配处理过程如图 1 所示。

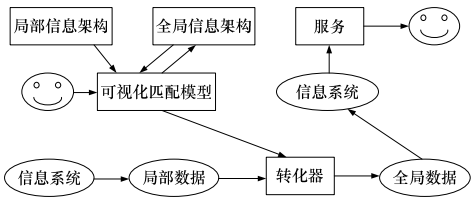

图 1 信息系统集成的可视化匹配处理过程

以图 1 为例，完成数据的集成通常需要以下 4 个步骤：

(1)针对不同信息系统中的局部 XML Schema，创建一个全局的 XML Schema，作为信息架构匹配的一个标准模式。

(2)利用局部 XML Schema 和全局的 XML Schema 在描述同一元素的语意上的差别，来建立信息架构匹配的算法模型，完成信息架构的匹配工作。

(3)在匹配的过程中，通过可视化匹配工具来自动生成转化器，使得局部的 XML 文档转换为全局的 XML 文档。

(4)把所有的局部 XML 文档转化为全局的 XML 文档之后，可以形成一个统一的信息系统，从而实现不同信息系统中的数据集成。

3 体系架构

要实现不同信息系统中的数据集成，必须完成 2 个 XML Schema 在信息架构上的匹配。可视化匹配的体系架构如图 2 所示。

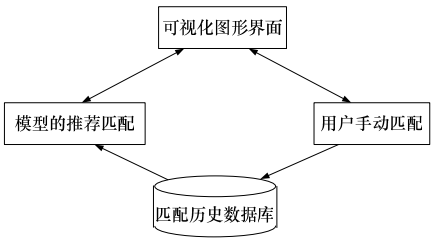

图 2 可视化匹配模型的体系结构

基金项目：国家自然科学基金资助项目(60573056)；浙江省自然科学基金资助项目(Z106335,Y107293)

作者简介：施 莹(1980－)，女，硕士研究生，主研方向：计算机网络技术应用，数据库，数据挖掘；蔡 勇，副教授

收稿日期：2008-07-30 **E-mail：**shiying@hutc.zj.cn

可视化匹配工具主要由 4 部分组成:

(1)用 Java 来搭建可视化图形界面。在可视化图形界面中用双曲线树，同时呈现出所有的全局信息架构和局部信息架构，使用户在匹配的过程中能主动参与进来，从而使得匹配活动更具交互性和有用性。

(2)在匹配的过程中，模型的推荐引擎能帮助用户找到与局部的信息架构可能相匹配的全局信息架构。当用户选择了一个局部信息架构中的元素，推荐引擎通过算法找到一个与之相适合的全局信息架构中的元素，同时以列表的形式清楚直接地呈现在屏幕上。

(3)用户的手动匹配能依用户的意愿来进行局部信息架构和全局信息架构的匹配。当用户决定保存该匹配，它将会被保存在匹配历史数据库中，并利用 XSLT[2]自动生成 XSL 样式表。

(4)匹配历史数据库存放了所有的匹配历史。当用户更改了匹配信息，又保存了被更改的信息时，匹配历史数据库中的内容就会被更新。当用户选择了局部的信息架构，推荐引擎将会从匹配历史数据库中找出与之相匹配的全局信息架构。

4 关键技术的实现方法

4.1 可视化角度

信息可视化，就是“计算机支持的、交互的抽象数据图像化方法，从而帮助用户增强识别信息的能力”[3]。对于信息架构的匹配处理，笔者认为最好的表现方式应该是呈图形显示。图形对分析问题提供了最直接的手段，从视觉上给了用户很大的冲击，留下了一个深刻的印象。

针对 XML Schema 特有的层次机构模式，最适合的方法是采用双曲线树(hyperbolic tree)来进行描述。双曲线树[4]是一种交互的、多维的可视化方法，它能使用户毫不费力地找到他们正在寻找的精确信息。双曲线树的结构是把一个文件作为一个节点的形式来呈现的。用户可以通过攥住某点进行拖动来浏览下层文件。与传统的文件结构相比，正是由于双曲线树的这种柔韧性特点，用户不必为了看到下层文件中的内容，经常地打开或者关闭文件。因此，据目前的研究表明，人们用双曲线树来定位信息比标准的导航方法快 62.5%。

4.2 算法角度

为了实现局部和全局的 XML Schema 匹配问题，又基于描述同一元素的语意上的不同，笔者考虑从算法的角度来加以解决。

通过局部 XML Schema 的元素来查找与之相匹配的全局 XML Schema 的元素，可通过可视化匹配工具的推荐引擎来实现。这种推荐匹配的实现主要由以下 3 种算法来自动完成。

算法 1 是以元素的名字为匹配基础的一个算法。对同一个元素因语义上的差异，用 XML Schema 的元素的名字做一个字符串比较的匹配，然后通过调用距离函数来计算任意 2 个字符串的差异，把这种差异用一个整数作为匹配与否的参数标志，只要返回的整数值小于设定的一个阈值，就认为匹配成功了。

算法 2 是以用户定义的规则作为匹配的准则。用户自己定义了一些局部 XML Schema 和全局 XML Schema 相匹配的准则，因为这些准则总是希望在用算法 1 进行匹配之前被执行，所以要把算法 2 的优先级设置得比算法 1 高。

算法 3 是以匹配历史为基础的。当用户把一个局部信息架构和全局信息架构相匹配好之后(这是在用户的图形界面中由用户完成的)，这种样式的匹配将作为匹配历史保存到存放匹配信息的数据库里。推荐匹配引擎将从匹配历史里找相应的推荐。

5 人才信息系统集成的个案研究

下面给出 2 个有关人才信息的 XML Schema 的信息架构:

(1)局部 XML Schema

局部 XML Schema 代码如下:

```
<?xml version="1.0" encoding="UTF-8"?>
 <xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified" xmlns:xs="http://www.w3.org/2001/XMLSchema">
   <xs:element name="object">
     <xs:complexType>
      <xs:sequence>
           <xs:element name="people_type" type="xs:string"/>
           <xs:element name="ID" type="xs:string"/>
        <xs:choice>
           <xs:element name="teacher" type="TeacherDescription"/>
        </xs:choice>
      </xs:sequence>
     </xs:complexType>
   </xs:element>
     <xs:complexType name="TeacherDescription">
      <xs:sequence>
        <xs:elemext name="name" type="xs:string"/>
        <xs:element name="age" type="xs:string"/>
        <xs:element name="sex" type="xs:string"/>
        <xs:element name="speciality" type="xs:string"/>
        <xs:element name="pay" type="xs:string"/>
        <xs:elemext name="rank" type="xs:string"/>
      </xs:sequence>
     </xs:complexType>
   </xs:schema>
```

(2)全局 XML Schema

限于篇幅关系，截选与局部信息架构的语意有关的全局信息架构片段如下:

```
…
<xs:complexType name="TEACHERDescription">
 <xs:sequence>
    <xs:element name="NAME" type="xs:string" />
    <xs:element name="AGE" type="xs:string"/>
    <xs:element name="SEX" type="xs:string"/>
    <xs:element name="SPECIALITY" type="xs:string"/>
    <xs:element name="MONTHLY_PAY" type="xs:string"/>
    <xs:element name="RANK" type="xs:string"/>
 </xs:sequence>
</xs:complexType>
…
```

(3)用于人才信息系统集成的工具的生成

对于本文提出的模型，再结合人才信息系统的特点，用 Java 和 XSLT 生成了一个工具。这个工具可以实现基于 XML 的人才信息系统的集成，主界面如图 3 所示。

在此工具中，利用双曲线树，用户可以很清晰地看到局部的 XML Schema 和全局的 XML Schema 的结构。用户若选中局部的 XML Schema 中的 SEX，SEX 节点立刻变深。同时，该工具的推荐引擎使全局的 XML Schema 中的 SEX 也变成深色。在该界面下方的列表中也能清楚地看到该匹配。用户若

希望局部的 XML Schema 中的 pay 和全局的 XML Schema 中的 MONTHLY_PAY 相匹配，可以在左右的窗格中分别单击这 2 个节点。击中后，这 2 个节点立刻变成深色。用户也可以选择是否把该匹配永久地保存在历史数据库中。

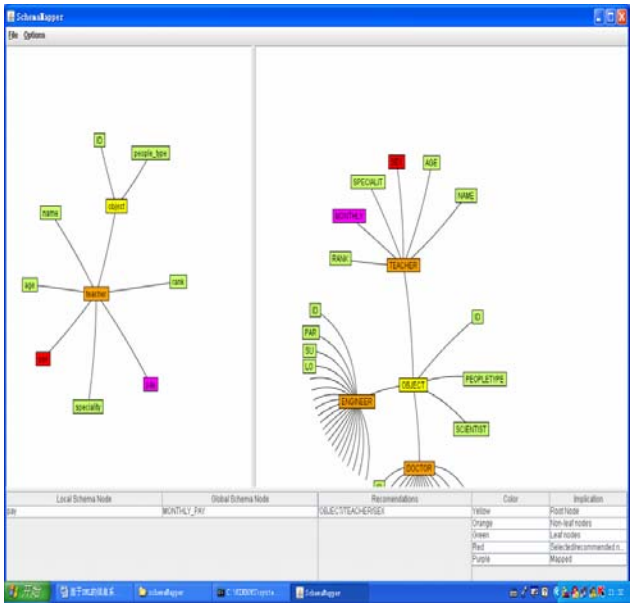

图 3　基于 XML 的人才信息系统集成的工具

(4)转化器

将局部的 XML Schema 中的 pay 和全局的 XML Schema 中的 MONTHLY_PAY 匹配成功后，自动生成的转化器的代码如下：

```
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:template match="/">
    <OBJECT>
      <FIGURINE>
        <MONTHLY_PAY>
          <xsl:value-of select="/OBJECT/FIGURINE/pay"/>
        </MONTHLY_PAY>
      </FIGURINE>
    </OBJECT>
  </xsl:template>
</xsl:stylesheet>
```

通过转换器，就可以把源 XML 文件转换为目标 XML 文件，完成整个人才信息系统集成的工作。

(5)可用性评量

实用性工程学是计算机科学的一个分支，它是从用户的角度，而不是从软件编程员的角度来体现人机之间的交互[5]。可用性评量是实用性工程学的一个组成要素。通常有 3 种评估方法：测试，检查和询问[6]。本文选择测试法来完成可用性评量。

目前在信息架构匹配的过程中，最常使用的软件是美国 Altova 公司推出的商业工具 Mapfoce。它的突出优点是可视化，但是商业价值高，作为学术上的研究，可用性不大。笔者随机选了 10 个用户，使用本文模型和 MapForce 对人才系统中的 5 个元素进行了可视化匹配工作，测试结果如表 1 和图 4 所示。

表 1　本文模型与 MapForce 比较结果　min:s

用户	本文模型花费时间	MapForce 花费时间
1	1:27	1:40
2	1:03	1:17
3	1:05	1:26
4	2:10	2:35
5	2:23	2:55
6	1:35	1:53
7	1:20	1:33
8	1:10	1:38
9	2:25	2:52
10	2:10	2:27

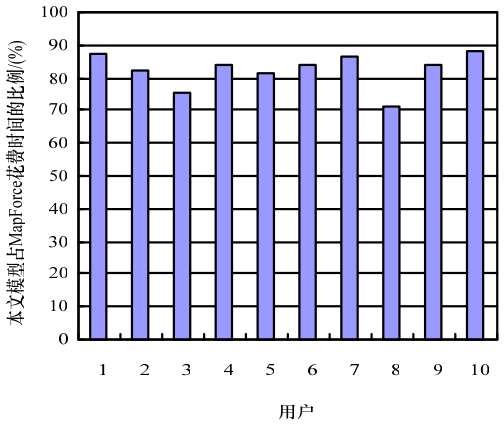

图 4　本文模型与 MapForce 花费时间的百分比柱状图

6　结束语

随着 XML 技术的不断成熟，XML 语言的共享与可视化需求不断增强。基于这一点，本文以一个全新的视角提出了一个可视化匹配的模型，该模型为基于 XML 的信息系统的数据集成提供了一种有效的解决方案。并由基于 XML 的人才信息集成的个案加以论证，用户只需该模型的推荐引擎模块和手动匹配模块即可快速有效地完成匹配工作来实现基于 XML 的信息系统的集成。与商业软件 MapForce 相比，该模型具有方便、灵活、可扩展性强等特点，因而使用它作为学术上的研究，可用性是非常大的。

参考文献

[1] Fallside D C. XML Schema Part 0: Primer[EB/OL]. (2001-05-02). http://www.w3.org/TR/xmlschema0.

[2] XSL Transformations(Xslt)[EB/OL]. (1999-11-16). http://www.w3.org/TR/xslt.

[3] Card S K, Mackinlay J, Shneiderman B. Readings in Information Visualization Using Vision to Think[M]. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1999.

[4] Zylab. Visualization Module[EB/OL]. (2007-10-30). http://www.zylab.com/products_technology/productsheets/Visualization.pdf.

[5] Hartson H R. CS5714 Usability Engineering[EB/OL]. (2007-11-03). http://courses.cs.vt.edu/~cs5714/spring2004/Class%20notes/01-Intro.pdf.

[6] Usability Evaluation[Z]. (2007-11-30). http://www.pages.drexel.edu/~zwz22/UsabilityHome.html.

编辑　顾逸斐