基于XML语言甲骨文语料库元数据抽取的研究

吴琴霞[1,2],高 峰[1,2],刘永革[1,2]
(1.安阳师范学院 计算机与信息工程学院,河南 安阳 455002;
2.河南省甲骨文信息处理重点实验室,河南 安阳 455002)

摘 要:计算机辅助甲骨文考释首先要解决的是以统一的数据形式来管理甲骨文信息,文中提出了一种基于XML语言的元数据抽取方法,提出了将抽取出的元数据存放到元数据库中,使用视图对元数据库进行动态地增加或删除,来弥补元数据抽取不全或错误的现象。同时通过XML Schema文档检测保存的元数据的语法格式,为以后映射到甲骨文语料库中做了语法的检测。该方法为甲骨文语料库的建设提出了一种新方案,也为后期语料信息语义化和专家系统推理打下了基础。

关键词:语料库;元数据;信息抽取;XML;XML Schema
中图分类号:TP393.01 **文献标识码**:A **文章编号**:1673-629X(2012)05-0216-03

Study of Oracle Bone Inscriptions Corpus Metadata Extraction Based on XML

WU Qin-xia[1,2],GAO Feng[1,2],LIU Yong-ge[1,2]
(1. School of Computer and Information Engineering,Anyang Normal University,Anyang 455002,China;
2. Henan Province Key Laboratory of Oracle-bone Inscriptions Information Processing,Anyang 455002,China)

Abstract:In order to assist expert makes explanations for oracle bone inscriptions,first of all,to be solved based on a unified data format to manage the oracle information. It presents a meta-data extraction method based on XML to extract the information is converted to a unified XML format,extracted metadata stored in metadata database,used a view on metadata database for dynamically inserting or deleting,to compensate for metadata extraction not all or the wrong phenomenon. And at the same time through XML Schema document detect grammar format of preservation metadata,is later mapped to oracle bone inscriptions corpus for the grammar of corpus detection. The method for oracle corpus built a new program,as well as laid the groundwork for the later corpus semantic information and expert system reasoning.

Key words:corpus;metadata;information extraction;XML;XML Schema

0 引 言

文字是民族的文化之根,甲骨文是中国迄今发现的最早的成系统的文字,是汉字的源头。但是,在四千多甲骨文字中,已考释出的甲骨文字不到一千字。因此甲骨文考释在甲骨学研究领域一直是一个热门课题[1]。甲骨文经过近一百年的考释,容易考释的字所剩无几,所以现在的甲骨文考释已经成了一项极富挑战性同时也非常困难的工作。用计算机辅助甲骨文考释与甲骨文专家相比具有以下优势:如海量的存储数据、严密的逻辑推理等。因此随着甲骨文信息数字化研究层次的深入,目前使用计算机去辅助甲骨文考释未释字已经可行。语料库是指大量文本的集合,是使用计算机技术辅助研究语言学而产生的,甲骨文语料库是研究甲骨文的数字化平台,让计算机辅助甲骨文考释未释字首要任务就是甲骨文语料库的建设。语料库的建设首要任务是元素据的抽取,元数据对语料库的研究具有重要的意义。XML是万维网(W3C)推出的一种具有结构性的标记语言,可以组织数据、标记数据、存放数据和定义数据类型。它允许用户自定义标签,具有良好的可伸缩性和灵活性,XML不仅能提供对资源内容的表示,同时也能提供资源所具有结构信息。因此文中提出了一种基于XML语言的甲骨文语料库元数据的抽取技术。

收稿日期:2011-09-22;**修回日期**:2011-12-26
基金项目:国家自然科学基金资助项目(60875081)
作者简介:吴琴霞(1980-),女,河南扶沟人,硕士研究生,讲师,CCF会员,研究方向为语义Web、中文信息处理;刘永革,教授,研究方向为数据库与知识库系统、数据挖掘、中文信息处理。

1 XML与语料库

XML指可扩展的标记语言(Extensible Markup Language,XML)。XML是一套用来描述数据语义的标

记规则,其焦点是数据的内容。这些标签将文档分成许多部件,并对这些部件加以语义标识。XML 是描述万维网上数据的统一语言,已成为未来的 Web 语言和 Web 上数据交换的标准,XML 将会成为数据处理和数据传输的最常用工具。XML 具有良好的扩展性和灵活性,不仅能对资源内容采用统一的数据格式进行表示,同时也能对资源所具有的结构信息进行表示,XML 适合表示各种信息数据因而被广泛接受。

语料库(Corpus)是语料的仓库或者语料的集合,甲骨文语料库是针对甲骨文领域的多个应用而专门收集有一定结构可被计算机程序检索到的、具有一定规模的语料信息的集合。甲骨文语料库不仅是原始语料的集合,而是具有数据结构的、标注了语法语义和语用的语言信息的语料集合。语料库在各个语言学研究领域广泛地应用,它是语言研究数字化的重要基础,使用甲骨文语料库辅助甲骨文考释可以克服传统的甲骨文考释存在的问题,如:1)语料的真实性和客观性;2)语料所涉及的广泛性;3)知识共享等等[2]。

采用 XML 语言存储语料信息具有很多优点。

第一,XML 语言具有通用的国际标准,方便数据的集成、共享和交换;

第二,XML 是各种应用程序之间进行数据交换的最常用的工具,可以减少语料库的程序和语料库数据的依赖性,提高语料库数据的独立性,实现机器可读;

第三,XML 标记语言采用树形结构,语法形式简单,数语言工作者便于掌握,可以方便的组织和整理语料库;

第四,XML 标记语言具有可扩展性的特点,允许用户自定义标签,方便元数据的添加。

2　语料仓库元数据的抽取

甲骨文拓片上保存的数据为非结构化信息,计算机无法实现自动化分析和抽取。目前为止发现的甲骨片大约有十五万片左右,信息量大并且信息结构复杂。然而甲骨片上记录的信息都是有规律可寻的,这就对实现元数据的自动抽取带来了可能性。因此可以采用 XML 标签语言对甲骨片上的卜辞信息进行描述,对抽取出的元数据来定义对应的 XML Schema 文档。

2.1　元数据

元数据(Metadata)是用于描述数据及其环境的数据[3]。元数据一经建立,便可以共享。通过元数得出语料库中语料的分类信息、文本信息等各种各样的信息。这将为辅助甲骨文考释提供更多的可能性。元数据是对数据的一种描述方式,使用元数据可以提高系统的查全性和查准性。采用元数据以结构化的方式表达甲骨片上存放的卜辞信息,方便了信息的共享。这为后期考释甲骨字提供丰富的语料信息。目前关于元数据的抽取和标注方法中存在一系列问题。文中提出了将抽取出的元数据存放到元数据库中,使用视图对元数据库进行增加或删除,这样方便用户自由的添加或删除元数据。元数据抽取工作分为以下两步:

1)对元数据存储。语料库设计初期,由于设计者自身的知识水平有限,元数据的抽取很难达到其合理性[4]。因此对抽取出的元数据采用数据库中字段的信息进行存储。

2)保存模板。依据抽取出的元数据生成 XML Schema 文档,提供采用 XML 格式保存元数据的模板。

抽取的元数据可以表示为如图 1 的树形结构。图中的虚线框代表可以省略、* 代表可以有零个或多个。依据此树形结构,可以定义出元数据对应的 XML Schema 文档。

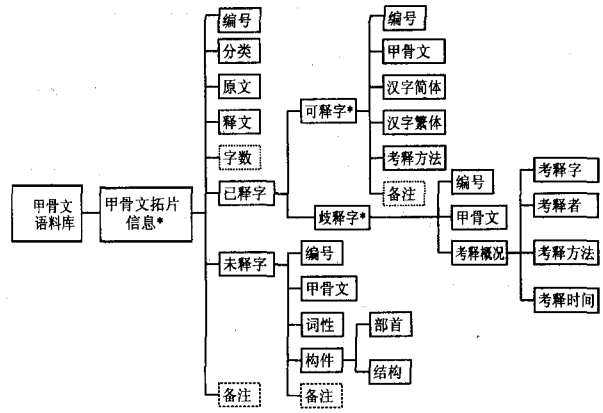

图 1　元数据树型结构图

2.2　甲骨文语料库元数据抽取模型

针对甲骨拓片上的卜辞信息进行元数据抽取,依据抽取出的结果定义出 XML 保存元数据的语法结构即 XML Schema,采用 XML Schema 对采用 XML 存储甲骨卜辞信息进行验证。最后保存为语料库(即 XML 存储到关系数据的存储)。元数据抽取模型如图 2 所示。各模块所对应的功能如下:

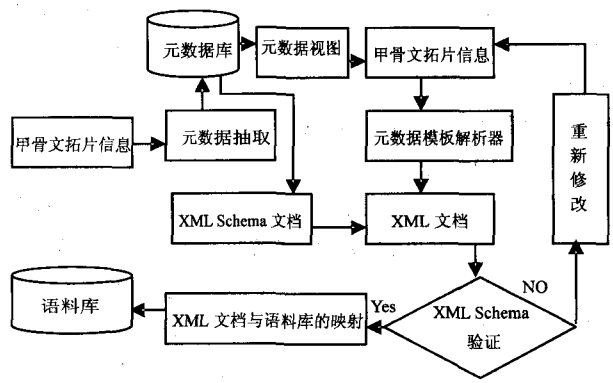

图 2　基于 XML 甲骨文语料库元数据抽取模型

甲骨文元数据的抽取模块:输入甲骨文拓片上的信息根据需要语料库的需求抽取元数据,根据抽取的

元数据建立甲骨文语料库中的存储表,元数据作为存取字段信息[5]。

视图模块:该模块主要是可以动态地增加和删除元数据,通过视图对元数据进行动态的管理,解决了设计者知识水平的局限性。保证了元数据抽取的完整性[6]。

元数据模板解析器:通过识别甲骨文拓片上的卜辞信息,对其格式进行归纳总结找出规律,进行批量的卜辞信息抽取,然后把抽取出的数据信息用XML的格式进行存储。对抽取的信息进行XML格式存储时,XML的标签即为元数据库的字段名。

XML Schema验证模块:利用XML存储甲骨文拓片上的信息数据后,为了检测抽取的元数据的语法格式,通过对应XML Schema文档进行语法验证[7]。

XML格式与语料库的映射:该模块主要是把XML存储的甲骨文卜辞信息映射到语料库中即转换到关系数据库存储。XML格式与关系数据库进行映射时,包括两部分即建立关系模式和解析XML格式的文档导入数据。在元数据抽取时已定制了语料库的关系模式,解析XML文档导入数据时只需注意XML文档与语料库的数据表相一致。在XML存储格式向关系数据库转换时,映射规则本身就存储在XML文档中,XML文档可以实现自解释。只需读取XML文档过程中的识别标记,并保存数据到语料库中[8]。

3 XML对元数据描述

通过XML Schema定义好XML语法结构后,就可以采用XML格式来保存抽取出的元数据,XML文档以树形结构来存储数据,依据保存的抽取的元数据的树形结构来构建XML文档。

3.1 元数据的类型定义

XML Schema是XML文档的结构和内容模式的约束。XML Schema完全以XML为描述手段,具有很强的描述能力、扩展能力和维护能力。XML Schema的元素都有name和数据类型属性。每条甲骨卜辞对应的XML Shema文档如图3所示[7]。

3.2 甲骨卜辞信息的XML存储

通过XML格式来存储甲骨文卜辞信息,不仅可以记录信息的本身数据,而且还可以准确的存储数据的语法结构[9]。另外,XML存储数据做到内容与形式完全分离,并遵循严格的语法格式,便于与不同系统间的数据传输,具有良好的保值性。另外XML允许粒度更新。设计人员不必在每次改动时都发送整个Schema数据集。使用XML格式保存甲骨卜辞信息更有利于知识的共享和数据的异步传输[10]。

图4是采用XML格式存储甲骨文卜辞信息的一个实例。

```
<xs:element name="甲骨文拓片" minOccurs="0"
maxOccurs="unbounded">
<xs:complexType>
 <xs:sequence>
 <xs:element name="编号" minOccurs="0">
 <xs:simpleType>
     <xs:restriction base="xs:string">
       <xs:pattern value="Z\d{5}" />
     </xs:restriction>
 </xs:simpleType>
 </xs:element>
  <xs:element name="分类" type="xs:string">
 </xs:element>
 <xs:element name="原文" type="xs:string">
 </xs:element>
  <xs:element name="释文" type="xs:string" />
   <xs:element name="字数" type="xs:int" />
         <xs:element name="已释字">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="可释字">
                <xs:complexType>
                  <xs:sequence>
                    <xs:element name="编号" type="xs:string"
minOccurs="0" />
                    </xs:element>
                    <xs:element name="甲骨文"
type="xs:string" />
                    <xs:element name="汉字简体"
type="xs:string" />
                    <xs:element name="汉字繁体"
type="xs:string" />
                    <xs:element name="考释方法"
type="xs:string" />
                    <xs:element name="备注"
type="xs:string" />
                  </xs:sequence>
                </xs:complexType>
              </xs:element>
```

图3 甲骨文元数据的XML Schema

```
<?xml version="1.0" encoding="gb2312" ?>
<甲骨文语料库
xmlns:jgw="http://www.w3.org/2001/XMLSchema-instance"
jgw:noNamespaceSchemaLocation ="jgw.xsd">
  <甲骨文拓片信息>
    <编号>B01723</编号>
    <分类>狩猎</分类>
    <原文>[image] ……</原文>
    <释文>令甲申令多射……</释文>
    <字数>6</字数>
<已释字>
 <可释字>
  <编号>198002</编号>
      <甲骨文>[image]</甲骨文>
     <简体>令</简体>
     <繁体>令</繁体>
    <考释方法>字形演变，繁文...</考释方法>
</可释字>
   <歧释字>
    ……
</歧释字>
</已释字>
    ……
  </甲骨文拓片信息>
  ……
</甲骨文语料库>
```

图4 甲骨文卜辞的XML描述

4 结束语

文中结合甲骨文语料库的建设提出了基于XML从甲骨文拓片上抽取元数据的方法,将抽取出的信息

(下转第222页)

所有案件。

(6)办结通知:当前审批申请者最近申办的办件经过相关某个部门的审批处理完成后,以通知形式告知申办者。

5 结束语

文中根据达县的实际情况,在分析和参考了众多现行的网上行政审批系统的基础之上,对达县网上行政审批系统进行了详细的需求分析,给出了核心业务的用例图和流程图。接着对系统进行了总体设计和数据库设计,给出了系统的体系结构图及数据库的E-R图,最后基于B/S架构,利用ASP.NET等技术实现了达县网上行政审批公共服务平台。该系统的实现对达县的电子政务建设和政府信息化建设有重要的现实意义,为审批系统的开发提供了参考和借鉴。

参考文献:

[1] 任雪松.电子政务现状与发展趋势[J].职业技术教育研究,2006(4):3-4.

[2] 张 雷.大庆市政府电子政务行政审批系统研究[D].哈尔滨:哈尔滨工程大学,2006.

[3] 吴 静,曹元大.针对网上审批系统的电子政务解决方案[J].微机发展(现更名:计算机技术与发展),2004,14(7):24-27.

[4] 郑 刚.基于浏览器服务器模式的网上教学系统[J].安徽机电学院学报,2001,16(1):58-60.

[5] 陆冬云,温 浩,许志宏.以客户为中心的浏览器/服务器(B/S)网络计算模型[J].计算机与应用化学,2001,18(4):336-337.

[6] 梁 娜,禹 农,杨国青.基于B/S计算模型的Web技术在电子商务中的应用[J].山东科技大学学报,2003,22(1):65-66.

[7] 杜根远.基于B/S模式的Web三层应用开发[J].河南城建高等专科学校学报,2002,11(1):37-39.

[8] 李书杰,李志刚.B/S三层体系结构模式[J].河北理工大学学报,2002,24(S1):25-28.

[9] 肖 丁,吴建林,周春燕,等.软件工程模型与方法[M].北京:北京邮电大学出版社,2008.

[10] 鲁 博,柴跃廷.关于统一建模语言-UML[J].计算机工程与科学,2000(4):13-19.

[11] 曲美霞.企业自动审批流程管理系统的实现[J].计算机技术与发展,2006,16(2):67-68.

[12] Lloyd D B. Integrating reporting services into ASP.net[J]. Dr-Dows Journal,2005(2):32-33.

[13] Anley C. Advanced SQL injection in SQL server applications [J]. IEEE Internet Computeing,2004(8):8-10.

[14] Sadiq S W. Handling dynamic schema change in process models[C]//Australasian Database Conference. [s. l.]:[s. n.], 2000.

(上接第218页)

以统一的XML格式进行组织,方便了甲骨文信息的管理和共享,更为以后的甲骨卜辞本体的生成和语义标注打下了坚实的基础[11]。为下一步甲骨文考释的知识表示与推理打下了语法基础[12]。

参考文献:

[1] 张德劭.甲骨文考释研究[D].上海:华东师范大学,2002.

[2] 何婷婷.语料库研究[D].武汉:华中师范大学,2003.

[3] Chan L M,Zeng M L. Metadata interoperability and standardization-a study of methodology part I:achieving interoperability at the schema level[J]. D-Lib Magazine,2006,12(6):121-123.

[4] 王 宁,郑振蜂.甲骨文字构形系统研究[M].上海:上海教育出版社,2006.

[5] 亓祥波,南 琳,张福顺.基于元数据和XML的信息抽取与集成技术研究[J].信息与控制,2008,37(1):52-57.

[6] 廖乐健,曹元大,李新颖.基于Ontology的信息抽取[J].计算机工程与应用,2002,38(23):110-113.

[7] 吴琴霞,刘永革.基于XML/Schema甲骨文语料库标注的研究[J].科学技术与工程,2009,17(9):5186-5188.

[8] World-Wide Web Consortium Xquery1.0:An XML Query-Language[EB/OL].2005-11-03. http://www.w3.ory/TR/Xquery,W3C Working Draft.

[9] 仲 华,催志明.基于XML的信息抽取和多层向量空间技术研究[J].计算机技术与发展,2007,17(7):49-52.

[10] Shamsfard M,Barforoush A A. Learning ontologies from natural language texts[J]. International Journal of Human-Computer Studies,2004,60(1):17-63.

[11] 梁晓涛,谢荣传.基于OWL描述本体的语义信息抽取[J].计算机技术与发展,2006,16(1):63-65.

[12] Li Man,Du Xiaoyong,Wang Shan. Learning ontology from relational database[C]//Proceedings of 2005 International Conference on Machine Learning and Cybernetics. [s. l.]:[s. n.],2005:3410-3415.