

XML 格式领域数据传输的优化技术研究

赵冲冲 王 鑫 胡长军

(北京科技大学信息工程学院 北京 100083)

摘 要 XML 由于有易用、可理解、跨平台的优点,在各个领域得到了广泛应用。它虽然有效地解决了跨平台数据的交换问题,但会带来传输效率问题。因为在传输 XML 格式数据时,由于 XML 文件完整性的要求可能传输大量的“无用数据”,这不仅增加网络负担,也降低系统数据服务速度。提出了一种 XML 格式领域数据传输的优化技术,其根本思想是在保持 XML 优点的同时,充分利用领域知识,按照领域数据的逻辑关系组织 XML 文件中的元素,将 XML 数据文件整体按照一定的规则分解为一系列的类型进行传输。通过这种划分可以实现数据交互过程中只传输修改过的类,从而减少数据的有效传输,有效解决了系统整体性能的问题。给出了不同粒度划分的原则和方法,给出了方法的适用条件和效率。本方法已在石油测井 XML 数据文件传输中得到应用,实用表明传输效率得到明显提高。

关键词 XML,数据传输优化,领域数据工程

中图分类号 TP311 **文献标识码** A

Optimization Technology of Domain-special Data Transmission Based on XML

ZHAO Chong-chong WANG Xin HU Chang-jun

(Information Engineering School, University of Science and Technology Beijing, Beijing 100083, China)

Abstract XML are widely used in various areas due to the usability,comprehensibility and cross-platform attribute. Although the XML technique can deal with the cross-platform transmission of domain data efficiently,the performance of this technique is limited by the fact that a great deal of additional data may be transmitted to keep integrality. This transmission problem not only increases the pressure of network,but also delays the response of data reserve. This paper proposed an optimized transmission technique for domain-special data based on XML. The thesis of this technique is to keep the attribute of XML,and to organize the elements of XML file in term of certain logical relationship and domain knowledge,i. e. based on certain rules,the XML data is divided into a series of classes to be transmitted without entire XML file. This approach can effectively resolve the problem arosen by the fact that the partial modification of XML data files lower the performance of system. Rules and method of the various granularity partitions were given. Application condition and efficiency of the method were given too. Proposed methods which have been applied in petroleum field can improve the efficiency of transmission.

Keywords XML,Data transmission optimization,Domain data engineering

1 引言

随着信息技术的发展,各领域的信息量呈爆炸性增长,其中包括众多公共有效的数据库资源,如生物、农业、材料、信息、交通、能源领域等。地理上广泛分布的用户都希望能够按需地、透明地交互访问和使用这些丰富的数据资源。但是每个领域中的数据资源分散在各地的多个数据库中,归属各个企业、部门或机构所有,往往缺少整体布局和管理,所采用的计算环境由不同平台组成。这些由不同核心技术构建的信息系统就像一个“信息孤岛”,各自有着不同的处理对象、操作方法和专用客户端。这就形成了大量业务逻辑类似但形

式分散且彼此缺少信息交互的系统。为实现各部门之间的协调工作、数据和数据之间的关联不易丢失,就必须解决在各个环节之间普遍存在的数据交互和部门协同的问题。XML 的出现为领域数据中异构数据源的交换带来了新思路和新技术^[8]。使用 XML 作为中间数据的交换技术克服了传统数据交换系统的缺点,具有以下优点:可扩展性,允许各个组织、个人建立适合自己需要的置标集合;结构性,数据存储格式不受显示格式的制约,把资料的内容与其表现形式合理地隔开,从而大大提高 XML 资料的可理解性、可交换性和可重用性;平台独立性,XML 文档是纯文本,独立于平台与应用^[4,7]。这些优点推动 XML 在各个领域和行业迅速得到推广和应

到稿日期:2008-09-23 返修日期:2008-12-18 本课题受国家 863 高科技发展计划项目(2008AA01Z109)、“十一五”国家科技支撑计划(2006038027015)资助。

赵冲冲(1976—),男,博士,副教授,研究方向为领域数据工程、软件工程等,E-mail:chongchongzhao@gmail.com;王 鑫(1981—),男,硕士,研究方向为领域数据工程、软件工程;胡长军(1963—),男,博士,教授,博士生导师,研究方向为领域数据工程与软件工程、并行计算模型与并行编译系统、网络存储系统等。

XML 技术虽然有效地解决了领域数据的交换问题^[1,9],有力地推动了领域信息化的进程,但是在实际的工程领域,如 CIMS、GIS 中有时会带来效率问题。因为 XML 文件的完整性和领域数据的特殊性,在传输 XML 格式数据时可能伴随着大量“无用数据”的传输。例如在一个 C/S 架构的石油测井数据系统中,当用户在客户端修改一份基于 XML 的 Well-LogML 数据文件时,需要上传至服务器并由服务器通知其它用户做相应的更新。如果 WellLogML 数据文件整体传输,不论客户端用户修改多少内容,都需要把整个文件上传至服务器并由服务器转发至其它客户端。这往往不仅影响传输效率,而且会浪费许多资源。因为修改的数据量占整体数据量的比例往往很低,所以造成大量未修改数据的重复传输。这是工程领域利用 XML 文件数据交换存在的普遍问题。特别是当文件较大、用户操作频繁且需要同步更新大量用户时,这问题就十分突出,甚至会影响系统的实用性。

针对这些问题,本文提出一种 XML 格式领域数据传输优化技术,核心思想是将 XML 数据文件分解为一组相关联的数据对象,在底层网络传输时只发送用户局部修改涉及到的数据对象而不必每次都传输完整的数据文件,从而允许局部修改 XML 格式数据,减少网络传输数据量,降低传输数据的无用率。

本文第 1 节是引言;第 2 节论述 XML 领域数据传输优化技术,详细分析该技术的优点和适用条件;第 3 节以石油领域的一个实际系统为对象,对优化方法进行实际验证和分析;最后是结论和下一步的研究工作。

2 XML 领域数据传输优化技术

2.1 优化技术模型

本文所提出的优化技术的核心思想是以领域数据模型规范为基础,充分利用领域数据的逻辑关联,实现更灵活的 XML 数据传输,如图 1 所示。

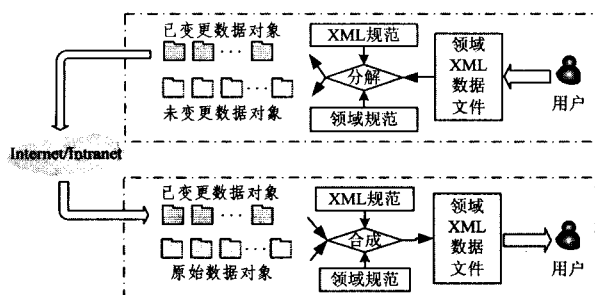


图 1 XML 领域数据传输优化模型

在模型中仍然采用 XML 作为用户的数据交互格式标准,这样可以充分利用 XML 的易用性和通用性。当用户修改 XML 数据并需要传送到其他节点以保持一致性时,不直接传送整个 XML 数据文件,而是首先参考领域知识和 XML 规范,以领域数据逻辑关联为依据,将 XML 数据文件分解成粒度更小的数据对象集合。然后确定修改的影响范围,将数据对象分为已变更数据对象队列和未变更数据对象队列。最后向其他网络节点发送已变更数据对象队列。其他节点在接收已变更数据对象队列后,替换原始数据对象集合中的相应部分。再将更新后的数据对象集成为 XML 文件,并展示

给用户。

如何分解领域 XML 数据文件,是优化的关键。合理的分解方法应在分解方式和分解粒度两个方面综合考虑。

- 分解方式:虽然不能预测用户的修改内容,但如果将关联紧密、容易同时修改的内容分解在同一个数据对象内,可以尽可能地减小已变更数据对象队列的数量和体积。

- 分解粒度:不宜过小或过大。粒度太大,失去优化意义,但粒度太小也会影响优化效果。首先,XML 数据文件分解和合成都需要额外时间,粒度越小花费时间越多。其次,将 XML 数据文件分解后必然需要附加信息,以记录数据对象和 XML 文件之间的映射关系以及数据对象队列信息,粒度越小附加信息越多。

考虑到领域特色,分解方式建议以行业数据模型标准提供的 XML Schema 或 DTD 文档为基础。确定分解粒度时,应综合考虑具体的数据使用模式、数据修改量等多种因素。下面将提供细粒度和中粒度两种划分方法。

2.2 定义

定义 1 有效传输数据是指 XML 数据文件进行网络数据交换时必须传输的 XML 元素,包括标签、标签属性以及元素值。

定义 2 无效传输数据是指 XML 数据文件进行网络数据交换时不必传输但由于其它原因仍须传输的 XML 元素,包括标签、标签属性以及元素值。

定义 3 XML 领域数据传输的数据有效比为:

$$De = \frac{\text{有效传输数据的体积}}{\text{XML 数据文件的体积}}$$

定义 4 XML 领域数据传输的数据无效比为:

$$Di = \frac{\text{无效传输数据的体积}}{\text{XML 数据文件的体积}} = 1 - \text{数据有效比}$$

数据有效比和数据无效比反映了 XML 数据文件的变化程度。数据有效比越低,表明数据变化相对整个文件体积越少。如果不采用传输优化技术,则传输的无用数据比例越高。

2.3 细粒度划分法

细粒度划分法按如下规则完成 XML 文件到数据对象集合的分解。

规则 1 把领域数据描述文件(即 XML Schema 或 DTD 文档)中的元素都逐一分解成对象,且规定对象名等于元素名。

规则 2 如果在 XML 元素标签中没有属性,则在对象中使用一个成员变量记录该元素的值。

例如元素 `<documentName>Rstm</documentName>` 映射为 `documentName` 对象后,在对象中使用一个成员变量 `value` 记录元素值 `Rstm`。

规则 3 如果在 XML 元素标签中有 N 个属性,则在对象中使用 $N+1$ 个成员变量,其中 1 个成员变量用于记录元素值, N 个与属性名同名的成员变量用于记录属性值。

例如元素 `<softwareName version="2.0">EzTools</softwareName>` 映射为 `softwareName` 对象后,使用成员变量 `value` 记录元素值 `EzTools`,同时使用一个名为 `version` 的成员变量记录元素属性 `version` 的值 `2.0`。

规则 4 XML 文件中父元素与子元素关系在对象集合中映射为聚合关系。

例 1 是石油领域标准数据模型规范 `Epicenter[x]` 中提供

的一个描述测井信息的 XML 文件 WellLogML。使用细粒度划分法分解例 1 并使用面向对象技术中的类实现数据对象, 可以获得图 2 所示的 UML 类图。图中每个类对应于 XML 文件中的每个元素, WellLogML 类对应 XML 根元素 WellLogML。

例 1 WellLogML 部分 DTD 示例

```
<WellLogML
  xmlns="http://www. posc. org/dtds/WellLogML. 1. 0"
  xmlns:xsi="http://www. w3. org/2001/XMLSchema-instance" xsi:
  schemaLocation = " http://www. posc. org/dtds/WellLogML. 1.
  0http://www. posc. org/dtds/WellLogML. 1. 0. xsd">
  <DocumentInformation>
    <dataOwner>N/A</dataOwner>
    <documentName>Rstm</documentName>
    <disclaimer>Oilware, Inc. is NOT responsible for the quality or
      correctness of the data contained in this file. </dis-
      claimer>
    <FileCreationInformation>
      <softwareName version="2. 0">EzTools</softwareName>
      <fileCreator>Oilware, Inc. </fileCreator>
    </FileCreationInformation>
  </DocumentInformation>
  <CurveData curve="DEPT" format="READABLE_TEXT">
    <data>11940. 5000 11941. 0000 ..... 11943. 0000 11943.
    5000</data>
  </CurveData>
</WellLogML>
```

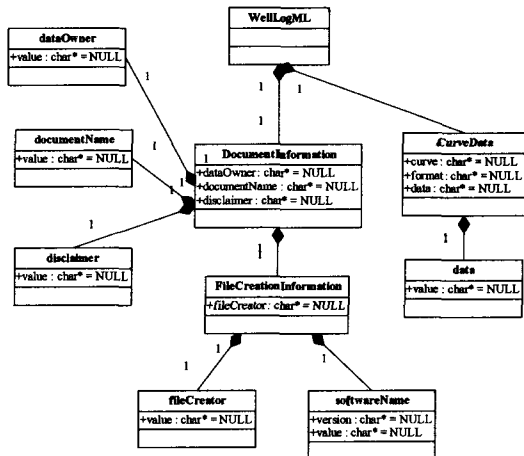


图 2 细粒度划分法示例 UML 类图

2.4 中粒度划分法

中粒度划分法按规则 5,3,4 完成 XML 文件到数据对象集合的分解。

规则 5 如果 XML 文件中除根元素外, 其它元素没有属性而只有元素值, 该元素就不单独映射成数据对象, 而是映射为父元素的一个属性, 其属性值就是该元素的值。为了保证从数据对象能够正确还原为 XML 元素, 可以在辅助信息中记录该属性是由于子元素转化来的还是真正的属性。

采用中粒度划分法重新分解例 1 并使用面向对象技术中的类实现数据对象, 可以获得图 3 所示的 UML 类图。例 1 中 XML 元素 documentName, dataOwner, disclaimer, fileCre-

ator 和 data 都没有属性而只有值, 根据规则 5 分别映射为父元素 DocumentInformation, FileCreationInformation 和 CurveData 的属性。softwareName 元素有自己的属性 version, 所以将该元素映射为一个独立的类。

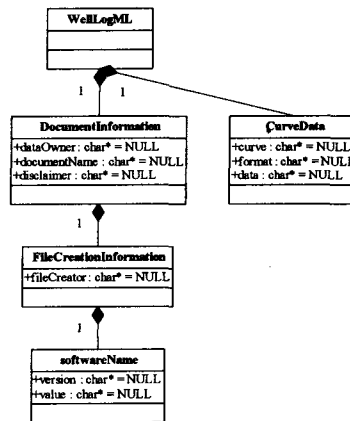


图 3 中粒度划分法示例 UML 类图

2.5 数据对象的关联关系描述和实现

无论使用何种分解方法将 XML 文件映射为数据对象集合, 都需要辅助信息记录数据对象之间的逻辑关联, 以便接收方正确地更新本地原始数据对象。

本文采用多叉树描述数据对象关联, 树上每一个节点代表一个数据对象。每个节点内的关联描述信息包括数据对象 ID、数据对象在多叉树中的路径 path、包含子对象的个数 childnum、指向子对象的指针 child 以及映射为父元素的属性名 transelement(适用于中粒度划分法)。

图 4 记录了使用中粒度划分法后 FileCreationInformation 数据对象的关联描述。其中根据规则 5 在 transelement 中保存了父元素数据对象 fileCreator。

FileCreationInformation
int ID=4
char * path= WellLogML/ DocumentInformation
int childnum=1
softwareName * child
char * transelement=fileCreator

图 4 附加信息示例

2.6 传输优化算法

基于上述说明, 本文提出如下传输优化算法:

- (1) 计算 XML 数据文件的数据有效比。如果数据有效比较高, 则直接传输整个数据文件, 否则执行以下步骤。
- (2) 选择 XML 数据文件划分法。
- (3) 根据所选划分法的规则, 把 XML 数据文件分解为一组已变更数据对象和一组未变更数据对象。
- (4) 计算关联描述信息多叉树。
- (5) 将已变更数据对象队列和关联描述信息多叉树发送到接收端。
- (6) 接收端根据关联描述信息, 使用已变更数据对象更新原始数据对象集合。
- (7) 接收端根据关联描述信息和更新后的数据对象集

合,合成新的 XML 数据文件。

3 分析与验证

本文选择石油领域 XML 数据文件传输进行实验验证。石油领域具有以下鲜明特点:(1)数据密集型行业。石油领域数据涉及地理信息、地质信息、油层信息、时间信息、井信息、工艺信息、设备信息等,内容复杂、格式复杂、数据量大、语义信息丰富。(2)数据交换频繁,XML 标准普及程度高。数据是石油领域生产和管理的基础,在石油勘探、开采、运输、生产、存储等各个环节需要频繁地交换数据。XML 标准比传统技术更适应石油数据交换的需求,已经成为行业标准之一。例如目前国际上广泛使用的石油测井数据 POSC 标准^[6]就是完全基于 XML 的。(3)体积庞大,数据集中。如 POSC 数据文件往往包含大量的地质数据,体积甚至可以达到几百 MB,而且大量的数据往往集中存储在 XML 文件的少量标签内。(4)数据操作频繁,一致性维护重要。石油领域部门多、用户多,不同用户操作数据后需要多次网络传输,通知其他用户,以保持数据之间一致性。石油领域的这些特点增加了网络数据交换的难度,按照传统方法直接传输 XML 数据文件将显著增加网络负荷。

实验以例 1 WellLogML 数据文件为对象,数据文件体积为 1MB。在<data>标签内集中保存大量数值,<document-Name>和<disclaimer>标签元素值为空。

测试 1 用户修改所有内容并传输给另一个节点,即使优化技术传输所有 XML 数据文件内容。

测试 2 用户把数据文件中的<dataOwner>和</dataOwner>标签属性值由“N/A”改为“一厂”,并传输给另一个节点。

测试 3 用户修改数据文件中<data>和</data>标签中的某个数值,并传输给另一个节点。

表 1 和表 2 分别列出了使用细粒度划分法和中粒度划分法的实验结果。

表 1 细粒度划分法结果

	测试 1	测试 2	测试 3
数据有效比	100%	0.0000038%	99.9%
数据对象数量	10	10	10
数据对象传输数量	10	1	1
传输数据量	≈1MB	800 B	≈1MB

表 2 中粒度划分法结果

	测试 1	测试 2	测试 3
数据有效比	100%	0.0000038%	99.9%
数据对象数量	5	5	5
传输数据对象数量	10	1	1
传输数据量	≈1MB	455 B	≈1MB

对比实验结果可见,本文提出的传输优化技术在数据有效比较低的情况下可以显著减少 XML 数据传输量,而在数据有效比较高时则作用不大。

细粒度划分法用简单的形式定义了 XML 数据文件到数据对象、数据对象关联、数据对象属性的映射,具有方法简单、容易理解和易于实现的特点。但该方法直接将 XML 元素映射到数据对象而不考虑元素的特点,因此可能会出现数据对象数量庞大、关联关系复杂,导致分解/合成速度慢和数据传输量增加的问题。中粒度划分法避免将 XML 简单元素映射成数据对象,虽然也可能会传输少量无用数据,但可以减小数据对象的数量,简化数据对象之间的关系,有利于减少数据对象之间辅助信息的体积。

XML 格式领域数据传输优化技术在石油领域某测井系统中得到了应用,有助于降低一致性维护时大体积 XML 数据文件的传输量。根据使用经验,当数据有效比低于 15% 时,建议采用传输优化技术。没有属性而只有值的元素数量少于 20 个时,采用细粒度划分法;多于 20 个时,采用中粒度划分法。

结束语 领域工程中普遍采用 XML 技术解决数据的交换问题,但是传输的效率问题往往是影响系统实用性的关键。依托领域数据模型,根据领域知识和 XML 规范将 XML 数据文件分解为一组逻辑相关的数据对象进行传输,减少无用数据的传输量,从而提高系统的数据服务速度,是一种简单、实用的优化技术。特别是对于传输数据量大、交互次数多而修改数据量较小的典型工程应用,此方法具有重要意义。此方法在石油数据传输系统中得到了成功应用。下一步的研究将在目前应用的基础上,在理论上进一步分析优化效率和数据传输量、带宽、交互次数、划分粒度之间的关系,提出更为普遍的理论模型。

参考文献

- [1] Extensible Markup Language (XML) 1.0.2 edition[EB/OL]. <http://www.w3.org/TRREC-xml>, Oct. 2000
- [2] W3CXML-Data[EB/OL]. <http://www.w3.org/TR/1998/NOTE-XML-data-0105/>, Jan. 1998
- [3] W3C XML schema. WC working draft[EB/OL]. <http://www.w3.org/TR/xmlschema-0/>, May. 2001
- [4] W3C XML Specification DTD [EB/OL]. <http://www.w3.org/XML/1998/06/xmlspec-report-19980910.htm>, Jun, 1998
- [5] Arenas M, Libkin L. A normal form for XML documents[A]// Symposium on principles of database systems (PODS' 02)[C]. Madison, Wisconsin, U. S. A: ACM Press, 2002: 85-96
- [6] POSC Specifications and Implementations [EB/OL]. <http://www.posc.org>, V2.2 1997
- [7] Zhou E Z. XML and Data Exchange for Power System Analysis [J]. IEEE Power Engineering Review, 2000, 20(4): 66-68
- [8] Brvan M. An Introduction to the eXtensible Markup Language (XML) [J]. Bulletin of the American Society for Information Science, 1998, 25 (1): 11-14
- [9] Goldfarb C F, Prescod P. The XML Handbook[M]. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998