一种基于虚拟 XML 技术的 XQuery 查询器系统框架*)

朱征宇　朱庆生　王　茜
(重庆大学计算机学院　重庆400044)

An Architecture of the XQuery System based on Virtual XML

ZHU Zheng-Yu　ZHU Qing-Sheng　WANG Qian
(Computer Institute of Chongqing Universty, Chongqing 400044)

Abstract　Based on idea of virtual Web page for the HTML page[6], this paper proposes firstly a concept of VXMLF (Virtual XML File), which separates the structure of XML file from the content of that. By means of modifying data model and query language of Xquery proposed by W3C, the paper gives an architecture of XQuery system based on virtual XML. Comparing with the original system, this query system has many new characteristics and advantages such as simplifying memory process, easy to maintain consistency of data, optimization of query computation and simplifying user operation.

Keywords　Virtual XML file, Data model , XQuery system

1. 引言

XML 因具有可以直接在互联网(Internet)上使用、支持大量不同应用、与 SGML 兼容、XML 文件容易编写且能够让人直接阅读等优点，受到人们的普遍关注，应用日益广泛。

但因 XML 文件既包含内容也包含(通过标记表示的)内部结构，Web 数据资源采用 XML 文件来描述，在下述方面可能会给应用处理带来不利影响：1)当 XML 文件内容繁多时，文件庞大，影响内存的处理效率；2)当需要在 XQuery 查询的结果(XML 文件)上进行二次查询时，一般仅把其查询语句作为虚视图(view)保存，但存在两次查询计算和查询优化问题，降低了查询效率；3)若把查询的结果作为实视图保存来提高二次查询的效率，又会引起数据冗余和一致性问题。

经过仔细分析不难看出，引起上述问题的关键在于 XML 文件自身的特点：数据被直接包含在 XML 文件中。借鉴文[6]针对 HTML 文件的技术思想，本文讨论给出了针对 XML 文件存储与查询的虚拟 XML 技术，使这些问题得到了较好解决。

该技术的基本思想是：1)引入 VXMLF 概念，在符合 XML 规范[1]的基础上将 XML 文件的内容剥离出来；2)通过对 W3C 提出的 XQuery 数据模型和查询语言作细微调整，使其能够在 VXMLF 上工作；3)虚拟 XML 技术只在内部发挥作用，而对 XQuery 查询器外部的应用来说几乎没有影响。下面分别加以讨论。

2. 相关知识

2.1 XML

可扩展标记语言 XML(Extensible Markup Language)，专为适应 Web 上大量电子数据发布和不同应用间数据交换而设计，能够描述各种 Web 数据资源，包括结构化和半结构化文件、关系数据库和对象存储库。它不但包含内容数据(content)，而且还通过用户可自定义的标记来表示内容的结构和属性(语义)。

例1　一个 XML 文件举例如下(选自文[3]中1.1.2节，http//www.bn.com)：

```
〈bib〉
    〈book year="2000"〉
        〈title〉Data on the Web〈/title〉
        〈author〉〈last〉Abiteboul〈/last〉〈first〉Serge〈/first〉〈/author〉
```

*)本文工作得到国家863项目[863-511-910-102-2]资助。

起来比较容易，而且签名数据量比较少，速度也比 DSS 快。

结论　本文提出的方案可以说是"Fujioka 方案"、"Wei_Chiku 方案"、"谢金宝方案"和"姚立方案"的综合和发展。这个方案利用 RSA 体制下门限多重盲签名技术和可验证的选票序列码等技术，解决了"投票人中途退出"、"签证人欺诈"、"选票碰撞"三大难题，并简化了投票过程，因而很具安全性、实用性。

参 考 文 献

1　Fujioka A, Okatoma T, Ohta K. A Practical Secret Voting Scheme for Large Scale Elections[J]. Proceedings of Auscrypt, 1992, 92: 244～251
2　Ku Wei-Chi, Wang Sheng-De. A secure and practical electronic voting scheme[J]. Computer Communications, 1999, 22: 279～286
3　谢金宝，刘晖波. 基于盲、群签名和秘密共享的新型电子安全选举模型. 微型机与应用，2000，(9)：38～42
4　Riera A, JosepRifà, Borrell J. Efficient construction of vote-tags to allow open objection to the tally in electronic elections[J]. Information Processing Letters, 2000, 75: 211～215
5　段琪，孙淑玲. 电子选举研究概况[J]. 计算机应用，1998，18(4)：23～25
6　卢开澄. 计算机密码学. 北京：清华大学出版社[M]，1999
7　Shoup V. Practical Threshold Signatures. [EB/OL]. http: // citeseer. nj. nec. com/512. html.
8　Damgard I B, Koprowski M. Practical Threshold RSA Signature without a trusted dealer. [DB/OL]http: // www. brics. dk/RS/00/30/BRICS-RS-00-30. pdf
9　Frankel Y, Desmedt Y G. Prallel reliable threshold multisignature. [EB/OL]http: // citeseer. nj. nec. com/franke192parallel. html
10　姚立，李仲麟. 一个实用的电子投票协议的设计[J]. 华南理工大学学报(自然科学版)，1997，25(5)：96～99

```
    〈author〉〈last〉Buneman〈/last〉〈first〉Peter〈/first〉〈/
    author〉
    〈author〉〈last〉Suciu〈/last〉〈first〉Dan〈/first〉〈/author〉
    〈publisher〉Morgan Kaufmann Publishers〈/publisher〉
    〈price〉39.95〈/price〉
  〈/book〉
  〈book year="1999"〉
    〈title〉The Economics of Technology and Content for Digital
    TV〈/title〉
    〈editor〉
      〈last〉Gerbarg〈/last〉〈first〉Darcy〈/first〉
      〈affiliation〉CITI〈/affiliation〉
    〈/editor〉
    〈publisher〉Kluwer Academic Publishers〈/publisher〉
    〈price〉129.95〈/price〉
  〈/book〉
〈/bib〉
```

2.2 XQuery 数据模型[4]

为了满足 XQuery 查询的普遍需要[2,3]，W3C 推出了“XQuery 1.0和 Xpath 2.0数据模型”工作草案[4]。主要目的有二，一是精确定义包含在 XSL 和 XQuery 处理器输入中的信息格式，二是定义了在 XSLT、XQuery 和 Xpath 语言中表达式的所有可能值。在该数据模型中，XML 文件（通过将标记看作为节点）可以采用树结构来表示。对每种节点定义了一个构造器（constructor），用于产生具有唯一标识的该类的新节点；并定义了一组访问器（accessors），用于访问节点的有关信息。

例2 例1中 XML 文件的树结构表示为图1。

说明：1）◎表示文件节点（document node），●表示文本节点（test note），○表示元素节点（element node）；□表示属性节点（attribute node），虚线表示属性为元素的虚子节点；

2）标记加上标（如 $year^1$，$author^1$等）表示不同节点；

3）带引号的字符串表示（叶）节点的内容。

2.3 XQuery 查询语言[5]

W3C 工作草案[5]描述了 XQuery 查询语言，给出了各种表达式及其使用规则，以及查询序词（prolog）、函数、连接查询（join）、聚集（aggregate）、序列（sequence）查询、用户自定义函数等的说明。特别地，Xquery 允许使用构造器（constructor）表达式，使查询结果可以是 XML 元素和文件。Xquery 采用基于 XML 的语法，可供人们阅读，能够广泛地应用于各种 XML 数据源上。

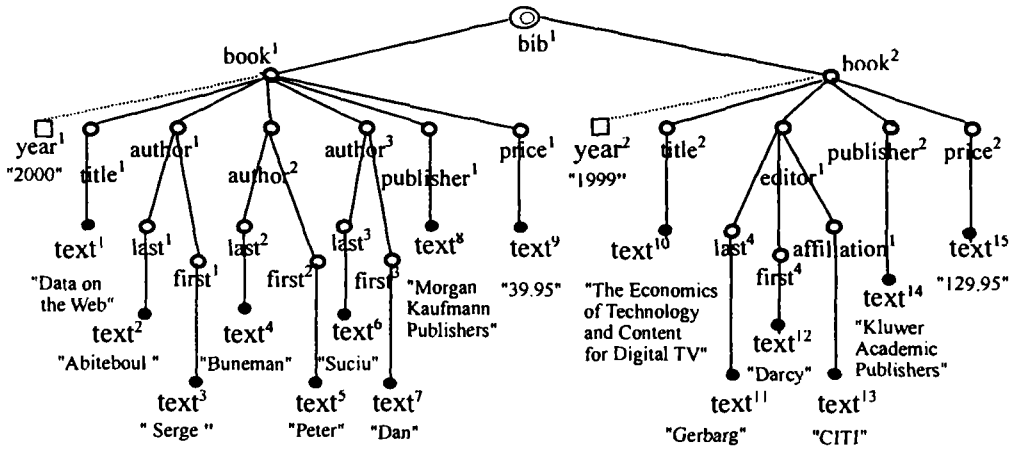

图1 XML 文件的树结构表示

3. 虚拟 XML 文件（VXMLF）

首先，将一个 XML 文件中各元素（elements）节点的、不含 XML 标记的、最大连续文本内容（简称内容段）都抽取出来，并对每个内容段赋予一个唯一标识（简称内容指针），以便能互相区分，而在原 XML 文件中，均用内容指针去替代相应内容段。

同样地，可将 XML 文件中的注释、处理指令、属性等节点的内容数据（内容段）也都抽取出来，而用相应的唯一标识（内容指针）去代替。为简化讨论（且不失一般性），以下讨论主要针对元素节点进行。

定义1 将在一个 XML 文件中去掉所有内容段、并以相应内容指针代替而得到的文件，称为虚拟 XML 文件，简记为 VXMLF（Virtual XML File）。

定义2 将从一个 Web 网站上所有 XML 文件中抽取出的内容段统一组织管理，称其为素材库，简记为 CDB（Contents Database）。

关于 CDB 的存储管理，以及如何根据内容指针从 CDB 获取相应内容段，不是本文重点，将在另文中讨论。

这里的“内容指针”是一种抽象概念，用于定位（即从 CDB 中提取）相应内容段，可解释为一种抽象标识或存取路径，其表示方式可因具体实现方式不同而异。

例3 例1中 XML 文件相应的 VXMLF 为：

```
〈bib〉
  〈book year=&year1〉
    〈title〉&title1〈/title〉
    〈author〉〈last〉&last1〈/last〉〈first〉&first1〈/first〉〈/author〉
    〈author〉〈last〉&last2〈/last〉〈first〉&first2〈/first〉〈/author〉
    〈author〉〈last〉&last3〈/last〉〈first〉&first3〈/first〉〈/author〉
    〈publisher〉&publisher1〈/publisher〉
    〈price〉&price1〈/price〉
  〈/book〉
    〈book year=&year2〉
    〈title〉&title2〈/title〉
    〈editor〉
      〈last〉&last4〈/last〉〈first〉&first4〈/first〉
      〈affiliation〉&affiliation1〈/affiliation〉
    〈/editor〉
    〈publisher〉&publisher2〈/publisher〉
    〈price〉&price2〈/price〉
  〈/book〉
〈/bib〉
```

其中，用前缀为‘&’的符号串表示各节点内容段相应的内容指针，所标识的内容如下：

表1 各内容指针所标识的内容段

内容指针	所标识的内容段（存放在 CDB 中）
&year1	2000
&title1	Data on the Web
&last1	Abiteboul

&first[1]	Serge
&last[2]	Buneman
&first[2]	Peter
&last[3]	Suciu
&first[3]	Dan
&publisher[1]	Morgan Kaufmann Publishers
&price[1]	39.95
&year[2]	1999
&title[2]	The Economics of Technology and Conte nt for Digital TV
&last[4]	Gerbarg
&first[4]	Darcy
&affiliation[1]	CITI
&publisher[2]	Kluwer Academic Publishers
&price[2]	129.95

应当指出，在新的 Web 系统中将不再保存 XML 文件，而是保存 VXMLF。系统内部的 Web 数据资源不再直接采用 XML 文件描述和存储，而是在 CDB 基础上间接地采用 VXMLF 来描述，与数据资源相关的所有原始数据统一存储在 CDB 中。

4. 对 XQuery 数据模型的调整

采用 VXMLF 描述 Web 数据资源后，需要对 XQuery 数据模型作必要调整，主要包括如下几个方面：

4.1 VXMLF 的树结构表示

XQuery 数据模型中，XML 文件可以采用树结构来表示，由于 VXMLF 与 XML 文件的不同之处仅在于将内容段替换为了内容指针，因此只需在 XML 文件的树结构表示中，对叶节点的组成成分作调整（将内容段改变为内容指针），即得到 VXMLF 的树结构表示。

例4 例3中 VXMLF 的树结构表示为图2。

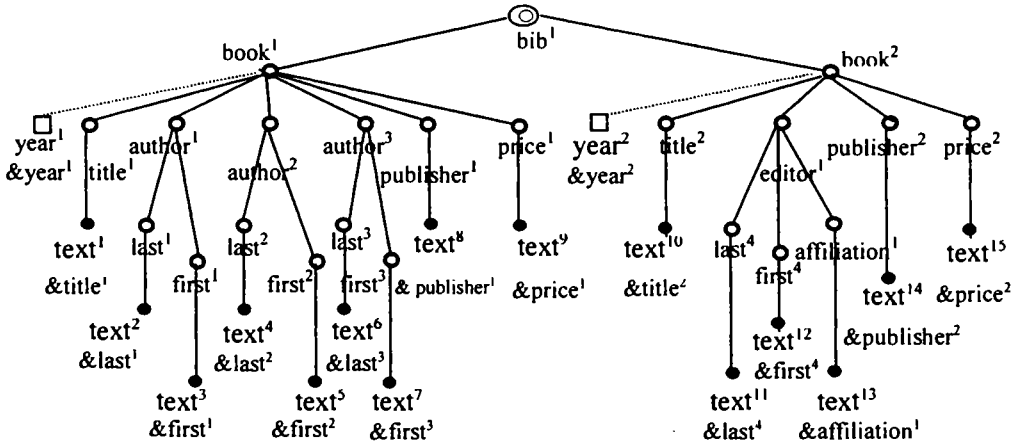

图2 VXMLF 的树结构表示

4.2 节点构造器和访问器的调整

由于将树中叶节点的组成成分从原来的内容段替换为内容指针，因此，在 XQuery 数据模型中，对各种类型叶节点的构造器和访问器也需作必要调整。现仅以文本类型叶节点为例加以说明：

1）对节点构造器 text-node 的调整：在构造文本节点时要求指明内容参数，原来为内容段，现应该为内容指针。

2）对节点访问器的调整：除原已有的一组访问器外，还应增加一个访问器 content-pointer，用于获取该文本节点的内容指针；原有访问器 string-value 仍用于获取该文本节点的内容段，只是在具体实现时应改为由内容指针从 CDB 中得到相应内容段。

5. 对 XQuery 查询语言的调整

为能应用于 VXMLF，XQuery 查询语言也需要作个别调整，包括如下两方面。

5.1 增加指针函数 pointer()

XQuery 查询语言[5]可以使用路径表达式(path)，其查询结果为 XML 元素或 XML 文件。为支持虚拟 XML 技术的需要，应允许查询结果为 VXMLF。为此可增加一个可作用于路径表达式之上的指针函数 pointer()，使查询结果中可以仍保存其输入文件 VXMLF 中的内容指针，而非相应内容段。

例5 如下 XQuery 查询：

```
for $b bib/book
return $b/title
```

返回查询结果为 XML 文件：

```
〈title〉Data on the Web〈/title〉
〈title〉The Economics of Technology and Content for Digital TV
〈/title〉
```

而使用指针函数 pointer()的如下 XQuery 查询：

```
for $b bib/book
return pointer( $b/title )
```

返回查询结果则为 VXMLF：

```
〈title〉&title[1]〈/title〉
〈title〉&title[2]〈/title〉
```

5.2 构造器表达式的调整

XQuery查询语言[5]可以使用构造器表达式(constructor)，使查询结果可以是 XML 元素或 XML 文件。为了支持查询结果(实视图)能作为 VXMLF 保存，应允许构造结果为 VXMLF 或虚拟 XML 元素，即构造器表达式中凡允许内容段之处也可以为内容指针。

例6 XQuery 查询：

```
〈results〉
  { for $a in distinct-values(document("http//www.bn.com")//
author)
    return
      〈result〉{ pointer( $a ) }
      { for $b in document("http//www.bn.com ")/bib/book
          where value-equals( $b/author, $a)
      return pointer( $b/title )
      }〈/result〉
}〈/results〉
```

其查询返回结果为 VXMLF：

```
〈results〉
    〈result〉
    〈author〉
        〈last〉&last1〈/last〉
        〈first〉&first1〈/first〉
    〈/author〉
    〈title〉&title1〈/title〉
  〈/result〉
  ……
〈/results〉
```

6. 基于虚拟 XML 技术的查询系统框架

6.1 系统框架

一般地，原支持 XML 的 XQuery 查询系统具有图3所示的系统框架。根据上述对 XML 文件及相关技术的调整要求，在原系统基础上扩展，即可得到支持虚拟 XML 技术的新系统框架如图4所示。

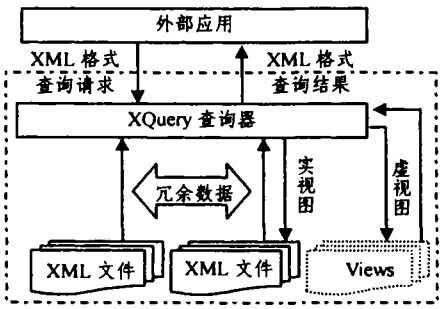

图3 原 XML 查询系统框架

二者的差别简述如下：

1) 原系统用 XML 文件描述 Web 数据资源，同时包含内部结构和内容数据，新系统改用 VXMLF 描述，仅包含内部结构，(Web 网站上)所有内容数据采用 CDB 统一存储管理；

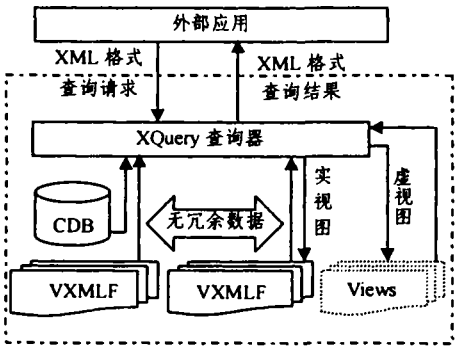

图4 VXMLF 新查询系统框架

2)原系统查询结果只能是 XML 文件，新系统查询结果则既可以是 XML 文件，也可以通过使用指针函数 pointer 返回 VXMLF(主要用于实视图保存情形)；

3)为二次查询需要，当一个查询结果需要作为实视图保存时，原系统作为 XML 文件存储，存在数据冗余和一致性问题，新系统则以 VXMLF 存储，避免了此类问题。

6.2 临时 XML 文件的管理

虽然采用虚拟 XML 技术必然增加从 VXMLF 生成实际 XML 文件或元素时提取内容段的额外开销，但对于较小的 XML 数据资源，增加的额外开销很小，可以忽略不计。而对于较大的 XML 数据资源，由于一个 XQuery 查询，无论是查询结果本身还是查询条件的处理过程，通常仅仅涉及其中很少一部分元素，因此在 XQuery 系统内部采用虚拟 XML 技术在相应 VXMLF 上工作时，也仅仅涉及从少量内容指针提取内容段的操作，增加的额外开销仍然很少。

此外，考虑到可能会有极少数 XML 数据资源，不但被频繁使用，而且采用虚拟 XML 技术处理时相应效率较低，不及直接在 XML 文件上进行处理，尤其是那些几乎总是整个地被调出和直接送往客户应用的 XML 文件，可以为系统设计一个辅助功能模块，专门用于监控这类资源，在需要时可及时产生实际的 XML 文件临时保存，在内容发生改变时及时更新临时文件，并定期清理不再频繁使用的临时 XML 文件。

结论 本文针对用“既包含内容，又包含内部结构”的 XML 文件来描述 Web 数据资源给 XQuery 应用查询带来的不利影响，提出了 VXMLF 概念，使其仅保存 XML 文件的内部结构，而将具体内容分离出来，统一组成一个素材库。然后，介绍了对 W3C 提出的 XQuery 数据模型和查询语言应作的必要调整。最后给出了基于虚拟 XML 技术的 XQuery 查询系统新框架。虚拟 XML 技术只在 XQuery 查询系统内部起作用，对外部应用来说是透明的。

与原系统相比，新系统具有如下特点：1)因不同应用需要，通常需要在查询结果基础上进行二次查询。对原系统而言，若将查询(语句)作为虚视图保存，则存在二次查询计算和优化的问题[7]，但若将查询结果作为实视图(XML 文件)保存，又存在数据冗余和一致性问题。而对于新系统，则可直接将查询结果作为实视图(VXMLF)保存，既可避免二次查询计算和优化，又不会引起数据冗余和一致性问题。2)一个 XML 文件包含的内容通常很全，但对于特定应用来说，往往仅涉及其中部分内容。利用 VXMLF 的特点，新系统可以为应用预先建立各种常用的查询实视图，既可以提高查询效率，又可以有效简化用户查询操作。3)由于 VXMLF 采用内容指针代替了内容段，在多数情形使文件变小，有利于减少内存占用，降低磁盘 I/O 交换开销，从而提高处理效率，尤其是在涉及路径搜索和涉及多个 XML 文件上的连接查询情形。

参 考 文 献

1 XML 1.0(Second Edition). W3C recommendation. Oct. 2000. http//www.w3.org/TR/2000/REC-xml-20001006

2 XML Query Requirements. W3C Working Draft. Feb. 2001. http/www.w3.org/TR/2001/WD-xmlquery-req-20010215

3 XML Query Use Cases. W3C Working Draft. Dec. 2001. http//www.w3.org/TR/2001/WD-xmlquery-use-cases-20011220

4 XQuery 1.0 and XPath 2.0 Data Model. W3C Working Draft. Dec. 2001. http//www.w3.org/TR/2001/WD-query-datamodel-20011220

5 XQuery 1.0: An XML Query Language. W3C Working Draft. Dec. 2001. http//www.w3.org/TR/2001/WD-xquery-20011220

6 朱征宇，等. 基于扩展标记图的虚拟网页技术. 计算机科学，2001，28(11)：80～82

7 Kato H, et al. A query optimization for XML document views constructed by aggregations. International Symposium on Database Application in Non-Traditional Environments. 1999