XML技术最新进展综述

马逸行1

1(北京航空航天大学 北京学院, 北京 海淀)

通讯作者: 马逸行, E-mail: [18182648@buaa.edu.cn](mailto:18182648@buaa.edu.cn), Tel: +86-13141001978

摘 要: XML作为当今处理分布式结构信息的有效工具，在互联网社会中具有非常重要的作用。本文从XML的发展过程出发，介绍了XML的基本特征，进而介绍其在数据库、信息查询及处理领域和互联网服务领域的进展。

关键词: XML、XML数据库、XML查询、最新进展

# XML基本概念

可扩展标记语言，标准通用标记语言的子集，简称XML。是一种用于标记电子文件使其具有结构性的标记语言。

在电子计算机中，标记指计算机所能理解的信息符号，通过此种标记，计算机之间可以处理包含各种的信息比如文章等。它可以用来标记数据、定义数据类型，是一种允许用户对自己的标记语言进行定义的源语言。 它非常适合万维网传输，提供统一的方法来描述和交换独立于应用程序或供应商的结构化数据。是Internet环境中跨平台的、依赖于内容的技术，也是当今处理分布式结构信息的有效工具。早在1998年，W3C就发布了XML1.0规范，使用它来简化Internet的文档信息传输。

可扩展标记语言与Access，Oracle和SQL Server等数据库不同，数据库提供了更强有力的数据存储和分析能力，例如：数据索引、排序、查找、相关一致性等，可扩展标记语言仅仅是存储数据。事实上它与其他数据表现形式最大的不同是：可扩展标记语言极其简单，这是一个看上去有点琐细的优点，但正是这点使它与众不同。

XML的简单易于在任何应用程序中读/写数据，这使XML很快成为数据交换的唯一公共语言，虽然不同的应用软件也支持其他的数据交换格式，但不久之后它们都将支持XML，那就意味着程序可以更容易的与Windows、Mac OS、Linux以及其他平台下产生的信息结合，然后可以很容易加载XML数据到程序中并分析它，并以XML格式输出结果。

# XML发展历史

XML有两个先驱：SGML和HTML，这两个语言都是非常成功的标记语言，但是都有一些与生俱来的缺陷。XML正是为了解决它们的不足而诞生的。

## SGML

早在Web未发明之前，SGML(Standard Generalized Markup Language，标准通用标记语言)就已存在，正如它的名称所言，SGML是国际上定义电子文件结构和内容描述的标准。SGML具有非常复杂的文档结构，主要用于大量高度结构化数据的访问和其他各种工业领域，在分类和索引数据中非常有用。

虽然SGML的功能很强大，但是它不适用于Web数据描述，而且SGML软件的价格非常昂贵；另外，SGML十分庞大，既不容易学，又不容易使用，在计算机上实现也十分困难：不仅如此，几个主要的浏览器厂商都明确拒绝支持SGML，这无疑是SGML在网上传播遇到的最大障碍。鉴于这些因素，Web的发明者——欧洲核子物理研究中心的研究人员，根据当时（1989年）的计算机技术，发明并推出了HTML。

## HTML

1989年，HTML诞生，它抛弃了SGML复杂庞大的缺点，继承了SGML的很多优点。HTML最大的特点是简单性和跨平台性。

HTML是一种界面技术，它只使用了SGML中很少的一部分标记，例如HTML 4.0中只定义了70余种标记。为了便于在计算机上实现，HTML规定的标记是固定的，即HTML语法是不可扩展的。HTML这种固定的语法使它易学易用，在计算机上开发HTML的浏览器也十分容易。正是由于HTML的简单性，使得基于HTML的Web应用得到了极大的发展。

## XML的产生

随着Web应用的不断发展，HTML的局限性也越来越明显地显现了出来，如HTML无法描述数据、可读性差、搜索时间长等。人们又把目光转向SGML，再次改造SGML使之适应现在的网络需求。随着先辈的努力，1998年2月10日，W3C(World Wide Web Consortium，万维网联盟)公布XML 1.0标准，XML诞生了。

XML最初的设计目的是为了EDI(Electronic Data Interchange，电子数据交换)，确切地说是为EDI提供一个标准数据格式。

当前的一些网站内容建设者们已经开始开发各种各样的XML扩展，比如数学标记语言MathML、化学标记语言CML等。此外，一些著名的IT公司，如Oracle、IBM以及微软等都积极地投入人力与财力研发XML相关软件与服务支持，这无疑确定了XML在IT产业的重要地位。

# XML数据库

XML数据库是一种支持对XML（标准通用标记语言下的一个应用）格式文档进行存储和查询等操作的数据管理系统。在系统中，开发人员可以对数据库中的XML文档进行查询、导出和指定格式的序列化。

## XML数据库的类型

(1)XML Enabled Database(XEDB)，即能处理XML的数据库。其特点是在原有的数据库系统上扩充对XML数据的处理功能，使之能适应XML数据存储和查询的需要。一般的做法是在数据库系统之上增加XML映射层，这可以由数据库供应商提供，也可以由第三方厂商提供。映射层管理XML数据的存储和检索，但原始的XML元数据和结构可能会丢失，而且数据检索的结果不能保证是原始的XML形式。XEDB的基本存储单位与具体的实现紧密相关。

(2)Native XML Database(NXD)，即纯XML数据库。其特点是以自然的方式处理XML数据，以XML文档作为基本的逻辑存储单位，针对XML的数据存储和查询特点专门设计适用的数据模型和处理方法。

(3)Hybrid XML Database(HXD)，即混合XML数据库。根据应用的需求，可以视其为XEDB或NXD的数据库，典型的例子是Ozone。

## XML数据库作用

XML数据库是一个能够在应用中管理XML数据和文档的集合的数据库系统。XML数据库是XML文档及其部件的集合，并通过一个具有能力管理和控制这个文档集合本身及其所表示信息的系统来维护。XML数据库不仅是结构化数据和半结构化数据的存储库，像管理其它数据一样，持久的XML数据管理包括数据的独立性、集成性、访问权限、视图、完备性、冗余性、一致性以及数据恢复等。这些文档是持久的并且是可以操作的。

当前着重于页面显示格式的HTML标记语言和基于它的关键词检索等技术已经不能满足用户日益增长的信息需求。近年来的研究致力于将数据库技术应用于网上数据的管理和查询，使查询可以在更细的粒度上进行，并集成多个数据源的数据。但困难在于网上数据缺乏统一的、固定的模式，数据往往是不规则且经常变动的。因此，XML数据作为一种自描述的半结构化数据为Web的数据管理提供了新的数据模型，如果将XML标记数据放入一定的结构中，对数据的检索、分析、更新和输出就能够在更加容易管理的、系统的和较为熟悉的环境下进行，因而我们将数据库技术应用于XML数据处理领域，通过XML数据模型与数据库模型的映射来存储、提取、综合和分析XML文档的内容。这为数据库研究开拓了一个新的方向，将数据库技术的研究扩展到对Web数据的管理。

## XML数据库特点

与传统数据库相比，XML数据库具有以下优势：

1、XML数据库能够对半结构化数据进行有效的存取和管理。如网页内容就是一种半结构化数据，而传统的关系数据库对于类似网页内容这类半结构化数据无法进行有效的管理。

2、提供对标签和路径的操作。传统数据库语言允许对数据元素的值进行操作，不能对元素名称操作，半结构化数据库提供了对标签名称的操作，还包括了对路径的操作。

3、当数据本身具有层次特征时，由于XML数据格式能够清晰表达数据的层次特征，因此XML数据库便于对层次化的数据进行操作。XML数据库适合管理复杂数据结构的数据集，如果己经以XML格式存储信息，则XML数据库利于文档存储和检索;可以用方便实用的方式检索文档，并能够提供高质量的全文搜索引擎。另外XML数据库能够存储和查询异种的文档结构，提供对异种信息存取的支持。

## XML数据库管理方法

利用数据库管理XML数据主要有两种方法：一是已有的关系数据库管理系统或面向对象数据库管理系统的基础上扩充相应的功能，使得能够胜任XML数据的管理，称之为支持XML的数据库，亦称为XML使能数据库；二是为了管理XML数据而量身定做的数据库管理系统，称之为原生XML数据库，亦称为纯XML数据库。

### 支持XML的数据库

要使RDBMS支持XML存储和查询，必须有一个XML转换层，这个转换层可以是RDBMS中的一个模块，也可以是RDBMS之上的一个中间件。这个转换层完成XML数据/查询与关系数据/查询之间的转换，其中包括两个模块：一个分裂器和一个查询翻译器。

其工作原理如下：首先，根据XML模式，分裂器产生对应的关系模式，XML模式到关系模式的映射会保留下来；然后，当载入XML数据时，分裂器根据XML模式到关系模式的映射将XML数据分裂到关系表中存储；用户提出XML查询，经查询翻译器翻译后，产生SQL查询，提交关系给关系引擎执行，返回关系结果，这些关系数据经过查询翻译器包装后，再以XML结果形式返回给用户。

支持XML数据库主要是基于关系数据库的。这种方法的优点是可以充分利用已有的非常成熟的关系数据库技术，集成体现有的大量存储在关系数据库中的商用数据，但这种处理方法不能利用XML数据自身的特点，如结构化、自描述性等特征，使得在处理数据的时候要经过多级复杂的转换，如存储XML数据时要将其转换为关系表或对象，在查询时要将XML查询语句转换为SQL或OQL语句，查询结果还要转换为XML文档等，多级的转换必将带来效率的降低，而且还会导致一些语义信息的丢失。

### 原生XML数据库

原生XML数据库这个术语，首先是由Software AG 公司为其开发的XML数据库产品Tamino所做的营销宣传中提出来的，后来这个术语在同类产品中成了通用叫法。但它只是一个营销术语，从来没有正式的技术定义。在Ronald Bourret 的网站上，Ronald Bourret 给出了一个定义（但这个定义并不是Ronald Bourret 本人提出来的），即满足以下三个条件的XML数据库才能称之为原生XML数据库：

（1）为XML文档定义一个（逻辑）模型，XML数据的存储和查询都是基于这个模型，这个模型至少包含元素、属性以及PCDATA等，保持文档顺序；

（2）将XML文档作为（逻辑）存储的基本单位，正如关系数据库将元祖作为它存储的基本单位一样；

（3）不要求只能使用某一特定的底层物理模型或某种专用的存储格式。

第一个条件要求原生XML数据库必须基于某种模型，这就像关系数据库基于关系模型，层次数据库基于层次模型一样。第二个条件的意思是在原生XML数据库中，数据只有在一个XML文档中才有意义，当然这并不妨碍查询一个文档的片段，就像在关系数据库中仍然可以查询一个元祖的片段（属性值）一样。第三个条件的意思是底层的存储格式并不重要。事实上，很多原生XML数据库是采用一些传统的数据库作为底层存储引擎的，当然也有很多是采用专有存储格式的。

# XML信息检索

XML信息检索是指对多个XML文档进行查询以活得理想结果。目前XML信息检索的研究主要包括XML索引结构、XML检索与排序方法，以及XML检索评价方法等。基于XML的信息检索系统有两种信息检索单位：一种是以文档为检索单位，另一种是以文档中被标注元素为检索单位。

## XML文档信息

一般认为XML文档的结构信息包含两方面：一是XML树中的祖先-后代关系（含父-子关系）；二是兄弟之间的关系。

### 支持文档检索的索引结构

该方法只适合与以文档为检索单位的XML搜索引擎。

### 索引方法

1、基于路径记录和结构索引思想，本体索引包括所有元素名称，用来提高检索效率。

2、基于树节点编码思想的索引结构通过对XML文档树节点采用不同编码方案来快速判断两个节点是够存在前后代关系。

### 节点编码方案

1、祖先后代编码（先序，后序），不足之处是不能判定是父子关系，当文档结构变化时，节点的标识要重新计算，所需花销较大。

2、间隔编码，（pre，size），数据库更新特别是插入时，间隔编码只需要较少的重新计算量，节点间隔的大小仅在插入节点的祖先节点改变。

3、虚拟节点编码，虚节点码采用广度优先的编码方式，虚节点编码模式支持一定程度的插入增量更新。

4、兄弟编码，具有同一个路径标识的父节点的子节点由兄弟数来区分。更灵活，支持增量更新。

## XML检索方案

### 基于改进向量空间的检索方法

1、利用XML文档的结构和语义信息，考虑检索单元的结构语义，从而提高查询效率

2、将XML及路径语言与传统的向量空间模型想结合，基于简单XPath路径的向量检索算法来时间线对XML文档的检索。

3、此外，在改进的向量空间模型中，特征词的权重计算方法大都是对tf\*idf的修改，即tf和idf的计算均以XML元素为单位计算的。

### 基于域加权词频法的概率检索

对于XML文档来说，不同的结构往往蕴涵了不同的语义信息，即使是XML文档级的检索，也与普通文本信息检索有所不同。XML文档通常包含一些子元素，考虑文档的内部结构对提高检索性能是有帮助的。

### 基于XML查询语言的方法

这些查询语言实现的是精确查询，不支持检索结果排序、传统的关键词查询以及异构数据源的查询，并不适合于在Web这样的开放式环境下或实在大型企业的intranets环境下检索信息。为此，一些研究致力于对XML查询语言进行扩展，使其适用于上述环境下的XML检索。

## XML检索返回结果的确定

XML检索系统是以XML元素为粒度来返回检索结果的，只返回用户感兴趣且符合检索条件的元素集，XML检索不但可以使得检索结果更为准确，也使得传输的数据量大大减小。节点的权重与节点文本大小成一定的反比关系。

## XML信息检索存在的问题

1）不支持检索结果排序、传统的关键词查询以及异构数据源的查询。

2）没有从根本上考虑XML文档结构信息

3）在输入查询信息时需要按照查询语言的语法输入XML文档结构信息和关键词

4）不同的标签可能表示相似的概念

# 基于XML的Web Service协议

SOAP、WSDL(Web Services Description Language)、UDDI(Universal Description Discovery and Integration)是Web Service 三要素， SOAP用来描述传递信息的格式， WSDL 用来描述如何访问具体的接口， UDDI用来管理，分发，查询Web Service 。具体实现可以搜索 Web Services简单实例 ; SOAP 可以和现存的许多因特网协议和格式结合使用，包括超文本传输协议（HTTP），简单邮件传输协议（SMTP），多用途网际邮件扩充协议（MIME）。它还支持从消息系统到远程过程调用（RPC）等大量的应用程序。SOAP使用基于XML的数据结构和超文本传输协议(HTTP)的组合定义了一个标准的方法来使用Internet上各种不同操作环境中的分布式对象。

## SOAP及其核心技术

简单对象访问协议是交换数据的一种协议规范，是一种轻量的、简单的、基于XML（标准通用标记语言下的一个子集）的协议，它被设计成在WEB上交换结构化的和固化的信息。

SOAP采用了已经广泛使用的两个协议:HTTP 和XML（标准通用标记语言下的一个子集）。HTTP用于实现 SOAP 的RPC 风格的传输, 而XML 是它的编码模式。采用几行代码和一个XML 解析器, HTTP 服务器( MS 的 IIS 或 Apache) 立刻成为SOAP 的 ORBS。SOAP 通讯协议使用 HTTP 来发送XML 格式的信息。HTTP与RPC 的协议很相似,它简单、 配置广泛,并且对防火墙比其它协议更容易发挥作用。HTTP 请求一般由 Web 服务器软件(如 IIS 和Apache)来处理, 但越来越多的应用服务器产品正在支持HTTP。XML 作为一个更好的网络数据表达方式( NDR)。SOAP 把 XML 的使用代码化为请求和响应参数编码模式, 并用HTTP 作传输。具体地讲, 一个SOAP 方法可以简单地看作遵循SOAP编码规则的HTTP请求和响应, 一个 SOAP终端则可以看作一个基于HTTP 的URL, 它用来识别方法调用的目标。像CORBA/ IIOP一样, SOAP不需要具体的对象绑定到一个给定的终端, 而是由具体实现程序来决定怎样把对象终端标识符映像到服务器端的对象。

## WSDL及其主要元素

WSDL（Web服务描述语言，Web Services Description Language）是为描述Web服务发布的XML格式。W3C组织（World Wide Web Consortium）没有批准1.1版的WSDL，当前的WSDL版本是2.0，是W3C的推荐标准（recommendation）（一种官方标准），并将被W3C组织批准为正式标准。

WSDL 元素基于XML语法描述了与服务进行交互的基本元素：

Type（消息类型）：数据类型定义的容器，它使用某种类型系统（如 XSD）。

Message（消息）：通信数据的抽象类型化定义，它由一个或者多个 part 组成。

Part：消息参数

Operation（操作）：对服务所支持的操作进行抽象描述，WSDL定义了四种操作： 1.单向（one-way）：端点接受信息；2.请求-响应（request-response）：端点接受消息，然后发送相关消息；3.要求-响应（solicit-response）：端点发送消息，然后接受相关消息；4.通知（notification）：端点发送消息。

Port Type（端口类型）：特定端口类型的具体协议和数据格式规范。

Binding：特定端口类型的具体协议和数据格式规范。

Port：定义为绑定和网络地址组合的单个端点。

Service：相关端口的集合，包括其关联的接口、操作、消息等。

References: