

基于 WEB 的比价交易代理模式的研究

胡纯蓉, 刘新华, 陈世平

(湖南工业大学, 湖南 株洲 412008)

摘 要:文中提出了一种为了检索万维网上的信息机制并构建了一个关系数据库。解决这个问题分三步:处理了基于 HTML 的 WEB 页面的困难;从 WEB 页面上抽取指定的信息并整合成结构化的文档;给出了把结构化的文档转换成相关的数据表的算法。满足了用户以最小代价、最短时间买到适合自己的商品。

关键词:比价交易代理;相对 URI;映射

中图分类号:TP393.08

文献标识码:A

文章编号:1673-629X(2007)03-0228-03

Research on WEB - Based Comparison - Price Transaction Agent Model

HU Chun-rong, LIU Xin-hua, CHEN Shi-ping

(Hunan Polytechnical University, Zhuzhou 412008, China)

Abstract: Proposed an architecture in order to retrieve some information on 3W, and populated a relational database. To solve it by three steps: dispose with those specific difficulties of Web pages based on HTML; extract and marshal data from Web pages into structural document; present an arithmetic converting valid XML document into relational data tables. It satisfied user by virtue of the shortest time and the lowest cost.

Key words: comparison - price transaction agent; relational URI; mapping

0 引 言

随着 Internet 技术的发展,网络商店也开始流行起来,它以节省时间、节约费用、操作方便吸引了越来越多的网民。然而,网上商店的多样性、产品的多样性、页面的动态性特征使参与网上购物的用户会有很多时间浪费在网上商店的浏览上,为了满足用户以最小代价、最短时间买到适合自己的商品的要求,目前网上出现了一种新型的交易方式——比价交易(comparison-price transaction)。所谓比价交易,就是消费者利用比较网站提供的搜索引擎技术,整合并集成众多在线销售商品的信息,对各种商品的性价比进行比较,从而选择物美价廉的商品。这就是比价交易的最大特点。

目前中国的比价交易有几十家,然而 WEB 信息的不断增长和异构数据源集成的应用,导致了大量半结构(Semi-structured)数据^[1,2]的产生。目前这些数据多是通过 HTML 语言来展现,而 HTML 语言的一个显著特点是结构模糊、不规则或不完整,使得 WEB

上的数据处于杂乱无序的状态,数据集成性极差。这对应用程序而言,无法直接解析、获取并利用 WEB 上海量的信息造成了极大的困难^[3]。然而,人们从 WEB 上获取和利用有用信息的要求却与日俱增。如何从浩繁的 WEB 数据中抽取出有用的信息成为众多研究工作希望解决的问题,比价电子交易代理就是其中的应用之一。

1 当前比价电子交易的技术特点

一个完整的比价交易代理系统必须包含信息的抽取模块,以发现信息为导向,以信息规则抽取技术为手段,为比价交易代理系统内核算法提供干净、准确、更贴切的信息,从而减少比价交易代理的数据处理量。在信息抽取过程中,通过识别并理解用户的抽取要求来确定所抽取的任务及相关的数据源,把 WEB 页面信息整合成结构化的页面信息,根据领域知识中的约束规则对页面信息进行合法性检查,将其转化为规范的 XML 文档,据语义的依赖性和句法规则来实现页面信息到数据表的映射。

2 比价交易代理的体系结构

基于 WEB 的比价交易代理(见图 1)是一个用

收稿日期:2006-05-26

作者简介:胡纯蓉(1977-),女,湖南双峰人,硕士研究生,研究方向为信息管理、软件工程;陈世平,教授,研究方向为信息管理。

XML及相关技术去抽取从WEB页面上的指定信息的系统,且能对关系数据库进行操作,是一个无人干预的独立域。通过它可以把WEB页面转换成结构规范的XML文档,映射成数据库的数据表,并存储在RDBMS里。

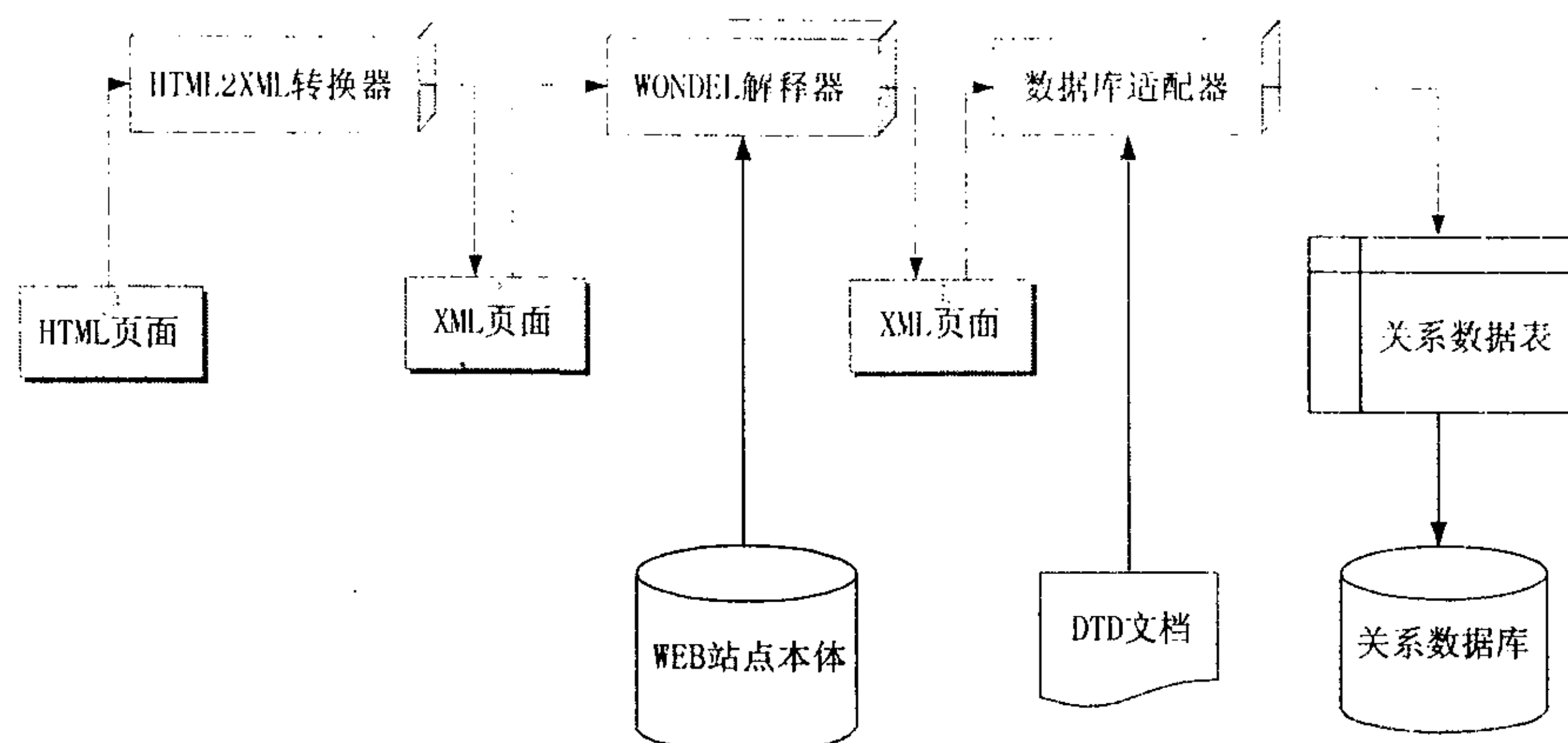


图1 比价交易代理的体系结构图

为了处理基于HTML的WEB页面,系统首先处理的是频繁出现在HTML页面里的句法错误和taglib规则的模糊性。这项任务是由HTML2XML转换器来完成的。这个组件是为了把所有的WEB页面经系统处理后转换成结构化的XML文档^[3]。WONDEL解释器是基于XML的WEB代理系统的核心,WONDEL解释器是用XML解析器和XPointer^[4,5]解释器来解释存在WEB站点本体的WONDEL文档,WONDEL站点本体是WONDEL文档的树。WONDEL解释器是一个能抽取且能集成来自WEB页面信息的组件,能把页面信息转换成规范的XML文档。这些XML文档遵守DTD文件所提供的声明,为了完成这项任务需要一些WEB站点的内容信息,这些信息存储在WEB站点的本体里。最后一个组件是数据库适配器,它是据DTD文档把规范的XML文档映射成关系数据表存储到关系数据库。

2.1 HTML2XML转换器

这个组件是用来把WEB页面转换成结构化的XML文档而设计的。完成这项转换涉及到如下操作:从HTML到XML的句法映射;解决了由于HTML标签规则所引起的模糊性;处理了WEB页面中句法错误。

2.1.1 HTML的WEB文档结构和格式

HTML的WEB文档是树型结构,树型结构里的元素通常有三部分:开始标签、内容、结束标签。HTML规范定义了两种元素类型:空元素和非空元素。HTML很流行,因为它很容易使用且它的标签规则很自由。然而这种自由性会带来如下问题:

1) 没有提供检测空元素的规则,这样解析器不得

不清楚地了解每个空元素。另外,结束符对一些非空元素而言是可以选择的,这样就要增加许多特殊事例代码到解析器里。

2) 目前的浏览器允许构建HTML文档的某种自由性,这就意味着文档不能严格遵守HTML规范也能正确显示,因此那些含有错误的文档可能会被编辑器所忽略。

2.1.2 把WEB页面转换成结构化的XML文档

1) 把标签名改变成超类:不像HTML那样,当处理元素名时,XML解析器是一个敏感的组件,这就需要在同一元素的开始和结束标签名必须是相同的。因此,在开始和结束符名都必须被转换成一个超类。

2) 改变空元素的标签句法:HTML4.0声明了如下的作为空元素的元素:IMG, BR, META, HR, AREA, BASE, COL, FRAME, ISINDEX, LINK, PARAM。对这些元素而言,必须在结束标签前插入字符‘/’。

3) 从属性声明里移除模糊性:在XML里,属性值必定是由两个相似的引号标记所包围,然而,在HTML文档里,属性值可能出现三种模糊性:a. 属性值可能没带任何引号标记,因此,需要检查每个属性的声明;如果没有带引号标记就增加引号。b. HTML文档可能没有包含它们所指定任何值的属性,因此,就需移除。c. 进一步而言,在WEB文档里,一个属性可能含重复多次声明,在这种情况下,移除重复多次的属性。

4) 固定嵌入元素的错误:在HTML文档里,有一些非空元素就其结束标签来说是可随意选择的,另外,HTML文档里出可能会忽略一些结束标签。在XML文档里,所有的非空元素必定要有一个结束标签,因此把WEB文档转换成HTML文档时,需要在正确的位置插入一些结束标签。

5) 解决相对URI问题:HTML的最重要的特点之一是能参考给定的文档里的其它信息源,这个特点通常被认为是超文本(Hypertext)。给定文档里的其它资源可能是HTML文档、图像、视频或任何其它文本或二进制文件。URI有两种类型:绝对URI是唯一定位资源所在位置;相对URI是被抽取资源的位置,这就必须根据基本的URI来解析。在转换处理过程中,HTML2XML转换器解决了把所有相对URI转换成绝对地址,解析相对URI的算法如下:

(1)如果元素 BASE 存在,那么基 URI 是 HREF 属性的值,否则基 URI 是所给定文档的 URI;

(2)检查文档里的 URI:

a. 检查是不是绝对 URI,如果是,就不需作任何转换了;

b. 如果不是,检查相对 URI 是否是以字符“/”开始的。i):如果是以“/”开始的,这就意味着是文件的绝对路径名,在这种情况下,新解析 URI 是:基 URI 里所给定的协议名 + “//” + 主机名 + 相对 URI;ii)如果不是以“/”开始,这就意味着是文件的相对路径,在这种情况下,新解析 URI 是:基 URI 里所给定的协议名 + “//” + 主机名 + 基 URI 里的绝对路径 + 相对 URI。

2.2 WONDEL 解释器

WONDEL 解释器是比价交易代理系统时最主要的组件,它是用来抽取和整合来自 WEB 页面上信息的组件。将所抽取的信息格式化成规范的 XML 文档,这些规范的 XML 文档必须遵守由 DTD 所指定的规约。用 XML 解析器和 XPointer 解释器去解释存储在 WEB 站点本体的 WONDEL 文档。WEB 站点本体是 WONDEL 文档的树,因此,可把 WONDEL 文档分成两种类型:叶子文档和结点文档。WONDEL 叶子文档是一个或多个 WEB 页面用户的观点,并给出了 WEB 页面所在的位置且描述了其语义。然而,所需的数据不会总存在一个页面里,可能分散在多个页面里,WONDEL 结点文档就是在这种无人干扰的情况下整合所抽取的信息机制。一方面 WONDEL 结点包括 WONDEL 叶子文档,另一方面,WONDEL 结点可以调用其它结点的结点,因此 WONDEL 文档能被组织成层次结构,这样层次结构更便于管理,易于维护。

2.3 数据库适配器

所给定的 DTD 表示的是一个数据库模型,规范的 XML 文档是遵守所给定的 DTD 规范且包含从 WEB 页面里所抽取的信息,数据库适配器根据所给定的 DTD,把规范的 XML 文档映射为 RDBMS 的数据表,

并存储在 RDBMS 里,见图 2。

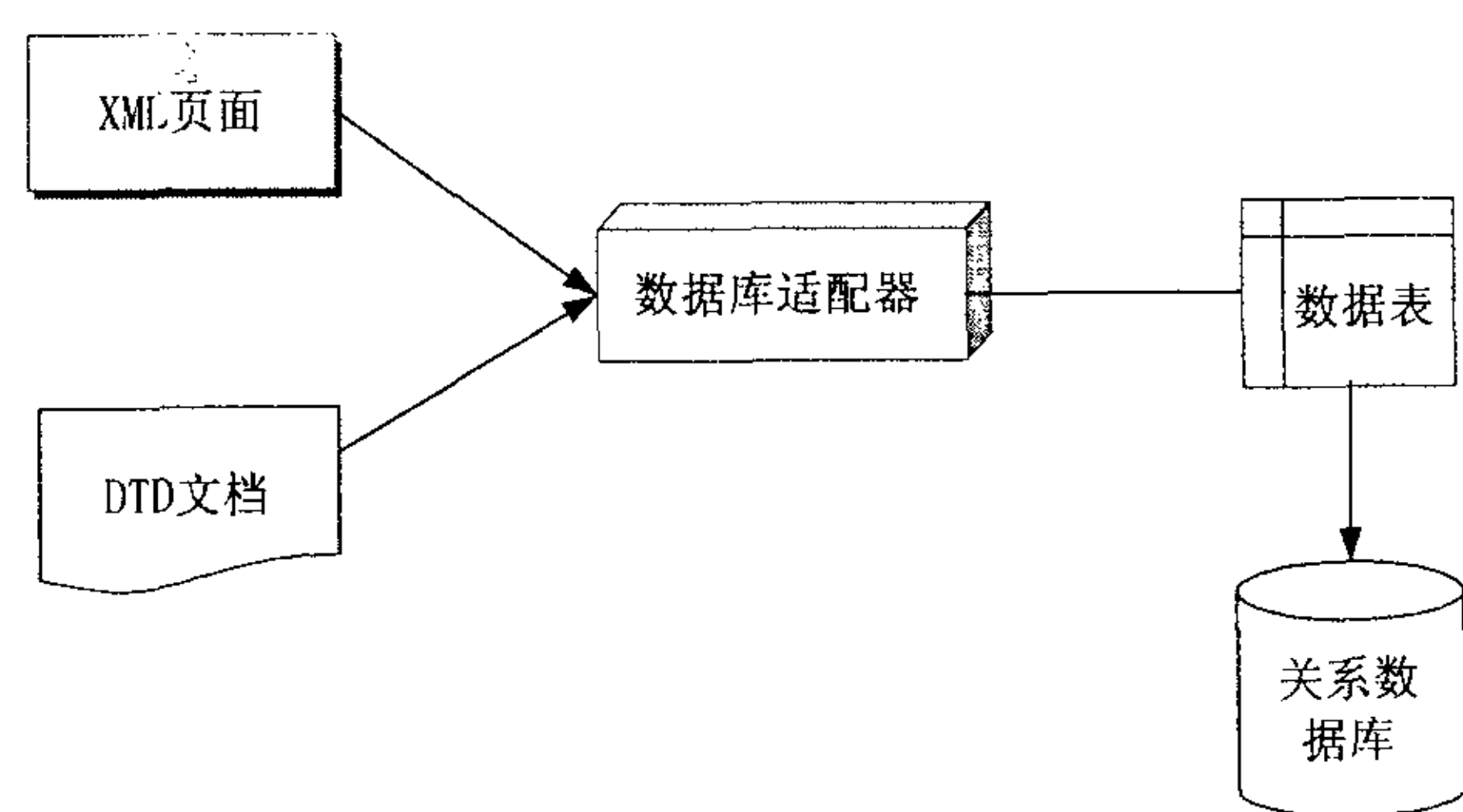


图 2 数据库适配器关系图

3 总结与展望

比价交易作为电子商务的一种强而又新生的营销方式,它不仅可以降低交易成本,而且还很便利地抓住商机,完成每一单业务,可以真正意义上地完成坐在家购物。随着比价交易日益成为人们的期盼时,网上交易、支付、配送等种种问题又着实令人头疼,在这个领域,定位、技术与市场开拓,将成为进一步的工作。

参考文献:

[1] Floresu D, Levy A, Mendelzon A. Database Techniques for the World- Wide Web: A Survey[J]. ACM SIGMOD Record, 1998,27(3):35-39.

[2] Buneman P. Semistructured Data[C]//In Proceedings of the ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems. Tucson, Arizona:[s. n.],1997:117-121.

[3] XML Schema(World Wide Web Consortium W3C)[EB/OL].2004-10-28. <http://www.w3.org/XML/Schema>.

[4] 王景皓,陈锦辉,XML 工作室.XML 与 JAVA 程序设计大全[M].北京:中国铁道出版社,2002.

[5] Berners-Lee T. Realising the Full Potential of the Web. Based on a seminar given the W3C meeting, London 1997 [EB/OL]. 1998. <http://www.w3.org/1998/02/Potential.html>.

(上接第 227 页)

提高了业务受理能力,增强了企业的竞争力。

Microsoft .NET 作为一种全新的技术,通过应用 ASP.NET,ADO.NET 及 XML Web Services 等技术,极大地改变了软件的开发模式^[6]。

参考文献:

[1] 孟军,王宝学.精通 ASP.Net 网络编程[M].北京:人民邮电出版社,2002.

[2] Anderson R,Francis B. Professional ASP.NET 1.0[M].王

毅译.北京:清华大学出版社,2002:65-102.

[3] 詹文军,王新程.安全应用程序开发[M].北京:清华大学出版社,2003.

[4] Kaufman J,Matsik B. ASP.NET 数据库入门经典——C# 编程篇[M].张哲峰,黄翔宇译.北京:清华大学出版社,2003.

[5] 王宝祥.基于 ADO.NET 的数据库访问技术研究[J].计算机应用与软件,2004,21(2):120-122.

[6] 毛德祥,罗荣阔.基于 ASP.NET 技术的 Web 应用程序三层设计模型[J].微型电脑应用(开发应用),2002,18(3):26-28.