Web数据仓库研究综述

韩京宇　徐立臻　董逸生
（东南大学计算机科学与工程系　南京210096）

摘　要　本文对Web环境下的数据仓库研究现状进行了综述。首先指出利用Web上的数据为决策支持服务的必要性，并针对半结构化数据如何集成到数据仓库、Web环境下的数据仓库体系结构、查询处理及几种具有代表性的Web数据仓库系统给出简要阐述。最后对相关问题的研究现状做了简要分析并加以展望。

关键词　Web数据仓库，半结构化数据，XML

An Overview of Web Data Warehouse

HAN Jing-Yu　XU Li-Zhen　DONG Yi-Sheng
(Department of Computer Science and Engineering, Southeast University, Nanjing 210096)

Abstract　This paper gives an overview of the research on Web data warehouse. First it points out that the necessity of making use of Web data to give support to DSS. Then it analyzes the methods of integration of semistructured data into web data warehouse, architecture of Web data warehouse, OLAP processing and the characteristics of several Web data warehouse systems. At last the paper summaries the problems related to Web data warehouse and the future reaearch.

Keywords　Web data warehouse, Semistructured data, XML

随着Web的发展，一方面Web上的数据量以前所未有的速度激增，如何充分利用Web上的数据为数据仓库的决策提供信息；另一方面数据仓库作为决策支持技术如何更好地支持Web环境下的分布式查询请求成为一个重要的研究课题。

1　研究背景

数据仓库是决策支持系统的基础，是一种面向决策主题的、时变的、集成的、不易失的、以读为主的数据集合[1]，是面向数据仓库应用如OLAP、数据挖掘、报表的基础。它作为一种高效地解决数据分析的技术，正在越来越多地应用到许多领域。然而随着Web的广泛应用，对数据仓库产生了巨大的影响：一方面以半结构化数据为代表的Web上的数据量飞速增长，如何有效地利用这些数据为企业的决策支持服务成为有意义的问题；另一方面由于企业的商务模式发生了巨大的变化，使得企业的各个部门有可能分散在世界各地，各个部门每天产生大量的数据，如何利用数据仓库有效地组织和管理这些分布于各地的数据，对传统的数据仓库的集中式的体系结构和查询处理方式提出了挑战。目前对Web数据仓库的研究尚无清晰的论述，本文就半结构化数据的集成和Web环境下数据仓库的体系结构、查询等相关问题的研究现状作了分析及展望。

2　半结构化数据集成

Web的飞速发展使得Web成为规模空前的数据库[2]，Web上的数据有HTML、XML及文本格式等，以半结构化数据如HTML、XML为主。目前用户利用Web上的数据主要有两种方法：浏览器方式和搜索引擎，浏览器方式由于在网页间导航式浏览是费时的，而且由于经常会出现“导航迷失”，因此是低效的；搜索引擎目前只能提供基于关键字的查询，而对用户发出的类似于结构化的查询如“查找高等数学成绩大于90分的1999级学生”则无能为力，而这种结构化的查询能力只有数据库技术才能提供。因此通过数据库技术对Web上的数据加以利用成为人们关心的问题。Web上的数据（本文仅讨论两种半结构化数据XML和HTML）集成有两种方法：数据仓库方法（warehousing）和虚拟方法（virtual）。前者是将Web上的数据装载到数据仓库中，查询在数据仓库中进行。后一种方法是基于一个“中间模式”[3]，数据仍然保存在局部的数据源中，通过数据源的“包装程序”（wrapper）数据虚拟成中间模式，用户的查询基于中间模式，不必知道每个局部数据源的特点，在本文中只讨论前者。数据仓库利用Web上的数据中关键一步所要完成的工作是如何将半结构化数据集成到数据仓库中。它面临着三个方面的挑战[4]：

（1）Web页面的不稳定性。（2）Web页面数据的结构和语义上的差异性。（3）半结构化数据和结构化数据的相互映射。

为了保证半结构化数据和结构化数据映射的完备性，设半结构化数据和结构化数据分别表达为$A=\langle S,Q\rangle(S\neq\Phi)$和$B=\langle S',Q'\rangle(S'\neq\Phi)$，其中$S$和$S'$分别表示半结构化数据元素集合和结构化数据元素集合，Q和Q'是相应的S和S'上的查询，则半结构化数据A和结构化数据B分别是一个代数系统，则寻求的两种代数结构上的理想映射$f:A\rightarrow B$应该是同构的，即满足双射的要求，从而使对XML半结构化数据的查询可以等价变换成SQL对关系数据的查询。下面对两种半结构化数据HTML，XML分别如何集成到数据仓库中相关问题加以分析。

2.1　HTML集成到数据仓库

HTML作为一种半结构化的网上标签语言，目前网上大约99%的Web文档仍是HTML格式，将HTML数据集成到

*）基金项目：江苏省十五高科技项目（编号BG2001013）。**韩京宇**　博士研究生，主要研究方向为数据仓库，XML；**徐立臻**　教授，主要研究领域为分布式数据库；**董逸生**　博士生导师，主要研究领域数据仓库，XML。

数据仓库中的关键一步是发现并抽取 HTML 页面中用户感兴趣的信息，大部分的研究方法建立在信息抽取技术(Information Extraction 简记为 IE)的基础上，IE 的主要任务可描述为：给定一个或多个数据源，利用相关技术，从这些数据源中获得一个结构元素的集合(又称为目标模式 Target Schema)。目前对 HTML 文档的抽取主要是通过包装器来实现的，包装器的任务是：对一给定的包含所需数据对象的页面 S，寻求一种映射规则将 S 中的数据映射到目标数据 R，并要求这种映射规则对其它相似于 S 的页面 S' 也是有效的。映射规则生成方法可以分成以下几种：

(1)手工方式

这种方式中抽取规则或抽取模式的获取是由用户来手工定义的，如 TIMMIS[5]对 HTML 页面的抽取任务是用一种可配置的信息抽取工具，它的抽取规则是用户按自己的需求和对相关网页的知识直接输入到包装器，然后包装器根据这些抽取规则执行信息抽取任务。ARANEUS[6]也采用类似的方法。

(2)启发式方式

这种方式中采用启发式的方法来确定 HTML 文档中用户感兴趣的数据片断和相应的数据片断中的“记录项”(record item)的边界，如文[7]中提出的五种启发式辨识“记录项”边界的算法来自动解析 HTML 文档以提取出“记录项”。

(3)基于本体方式

这种方式主要以文[8]为代表，在这种方式中首先由用户构造针对具体应用领域的本体概念模型，它描述了实体间的关系，实体常量及上下文关键字。通过解析本体概念模型，自动生成数据库模式来指导如何从 HTML 文档中辨识和抽取相应的数据信息。这种方法的优点在于不是用 HTML 的文档结构来指导数据抽取，而是由构造的应用领域的本体概念模型来指导数据抽取，因此只要针对应用领域的本体概念定义得恰当，就可以从具有不同结构的 HTML 文档中抽取用户需要的信息。

(4)基于机器学习的方法

这种方法首先通过机器学习的方法获得抽取模式后，然后利用规则或模式从相应的页面中进行数据抽取。根据是否需要训练样本可以分成有监督学习和无监督学习两种：(a)有监督学习：在这种方式中首先根据若干有代表性的样本页面训练出抽取模式和规则，然后再根据模式和规则进行页面数据的抽取。文[9]首先由有户选取样本页面，并对页面中将抽取的数据项作标记，以表明抽取模式，然后包装器根据页面标志来学习获得抽取规则。文[10]中每次通过比较两个属于同一页面类(page class)的 HTML 文档，从而产生相应的抽取模式，根据这个产生的模式来推断出抽取语法，如果有多个样本页面，则每次分别用已推断出的模式与样本页面进行比较，从而发现模式间的不同，进一步修正抽取模式，最后利用这种学习得到的模式来对新的页面进行信息抽取。(b)无监督学习：相比于有监督的学习方法不同，这种方法自动对将抽取的页面一边进行解析，一边根据一定的算法自动获取相应的抽取模式信息来完成抽取任务，不需要训练样本页面。如文[11]通过将一个 HTML 文档的标签结点和文本结点进行二进制编码成 PAT 树，利用 PAT 后缀树中内部结点对应的“重复模式”(repeated pattern)的性质来发现抽取模式。omini 系统[12]也是一个自动的 HTML 文档数据抽取系统，它首先将一个良构的(well-formed)HTML 文档解析成一棵标签树，然后根据启发式算法自动识将要抽取的数据区域并自动识别各个不同“记录项”的边界。这种方法可以完全自动化，但由于没有用户指导，有时抽取到的信息同用户的需求会有出入。

HTML 的信息抽取是以数据库的方式利用 HTML 数据第一步要完成的工作，其关键在于抽取规则的获取，在这方面还有许多工作要做，诸如如何实现增量抽取规则，如何用尽可能少的页面来学习获得抽取规则等。

2.2 XML 集成到数据仓库

随着 XML 的发展，它已成为公认的 Web 上的半结构化数据的代表，成为网上数据传输和交换的标准。相比于 HTML，XML 的语法结构更为严格，由于其可自定义标签因而具有很好的扩充性。目前将 XML 集成到数据仓库有两种方法：一种方法是物理上集成，即将 XML 数据物理存储到关系数据库中，相应的 XML 查询映射为 SQL 查询；另一种方法是逻辑上集成，从多维立方体的角度考虑如何利用 XML 数据来进一步丰富完善多维立方体。下面对这两种方法分别加以阐述。

(1)关系映射存储 XML 数据

如果想用 Web 上的 XML 数据为数据仓库的决策支持服务，其中一种重要的方法是首先将 XML 数据物理存储到关系数据库中，然后利用数据仓库技术，如 ETL、OLAP 等对这些数据加以处理、分析利用。关系数据库存储 XML 数据一方面涉及 XML 这种树状层次结构和关系表的平面结构之间的映射，另一方面涉及 XML 的查询如 Xquery、XML-QL、UNQL 和关系数据库的查询 SQL 间的相互映射。根据 XML 数据模式和关系模式映射的不同方法，可以分成三类。

(a)结构映射(structure-mapping)

在这类方式中，数据库模式代表了被映射的 XML 文档的逻辑结构，每个特定的 XML 文档结构会产生一种特定的模式。文[13]提出了依据 DTD 三种不同的映射策略：basic，shared 和 hybrid。其遵循的基本思想是每一元素的子孙元素尽可能与祖先元素放在同一表中(多值元素和元素间有循环嵌套的情况除外)。三种不同策略的主要区别在于数据冗余程度：basic 的冗余度最大，shared 没有数据冗余，hybrid 介于两者之间。该文同时给出 XML-QL 到 SQL 的映射。MONET[14]方法是一种与根据 DTD 映射截然不同的思路，它认为后者在同一个表中尽可能将其子孙元素也放入同一张表中的方法在查询时会扫描大量的无关数据，从而降低查询效率，为此 MONET 的方法基于边的方式进行关系映射，XML 树中每条具有相同标签的边作为一个二元关系存放，二元关系中的每一元组对应相应边的一个实例。这样基于关系数据库回答查询时，尽管 join 多一些，但由于每一个关系很小，因此效率是高的。但实际上[13]和 MONET 之所以形成这两种截然不同的观点，是因为其查询的模式是不同的。前者对数据发布式(published-orented)的查询更有效，而 MONET 对于查找式(lookup-orented)查询更有效。

(b)模型映射(model-mapping)

在这类映射方式中数据库的模式是固定的，所有的 XML 文档采用同一种数据模式被映射。Edge 方式[15]提出了一种基于边映射的方式，在这种方式中，将 XML 数据文件看作一个标签有序图，建立三张表，edge 表对图中的每条边相应建立一条记录，另外两张表分别存储图中结点元素所对应的 string 类型和 int 类型的值，这种方法的优点在于不需要 XML 模式信息即可以对 XML 文档实例自动进行转换，不足之处在于在重构一个 element 元素时为了获得其后代元素，要对 edge 表做多次自连接。文[16]提出的一种将 XML 数据存储到关系数据库中的方法，也要首先将 XML 文档存储成

固定格式的关系表。但其特点在于关系数据库中不仅存储着XML数据相应于关系模式的数据裂片，同时也存储着如何根据关系数据恢原XML文档的定义，这样在理论上，在关系数据库可以用一份物理数据同时向用户展示XML视图和关系视图。XREL[17]是一种基于路径映射的方法，在这种方法中为了保持XML文档的拓扑结构，在一张path表中存储了XML文档的所有路径，将结点区分为三种类型element，attribute和text，为每一种类型结点建立一张表来存储相应的结点值。同时它给出了xpath到SQL的映射算法。文[18]是另外一种基于路径的映射方式，它类似于XREL的方法，并给出由关系视图重构原来XML文档的NM(native method)和EM(efficient method)算法。

这种模型映射的方法的最大优点在于不需要任何XML文档的模式信息，对任何一个XML文档实例的映射模式是相同的。

(c)动态映射

这类方法没有固定的映射规则，而是采用动态的自适应的方法。STORED[19]采用一种对XML树进行数据挖掘的方法来同时实现模式发现和存储映射。logDB[20]提出利用XML schema结合工作负载(即各种不同应用查询的概率)和XML文档的相应数据元素的统计信息，采用启发式算法计算出在一特定的工作负载模式下(即各种查询的概率已经确定)的最优数据库模式。由于STORED采用数据挖掘的方法寻找映射模式，其搜索空间是指数比于整个XML数据实例的大小，因而寻优是不可行的，而logDB采用的方法是基于XML schema所提供的模式及相关数据统计信息，故寻优是可行的。文[21]提出根据XML文档的DTD建立关系模式后，再基于代价模型对模式进行寻优，类似于logDB的方法。这类方法的最大优点在于综合考虑了查询代价，存储代价来动态地寻优，不是单纯地从语法结构的角度寻求映射模式。

(2)逻辑上集成XML到数据仓库

一方面由于面向数据仓库的决策要求数据能够反映最近的趋势，而在物理上集成XML到数据仓库常常是滞后的；另一方面，面向数据仓库的应用如OLAP常常仅需要数据仓库中数据的一小部分，因此如果完全将XML数据物理集成到数据仓库中的方法有时不仅是低效的，而且不能满足及时性的要求。文[22]提出一种逻辑上集成XML到OLAP立方体的方法，即OLAP-XML联邦(OLAP-XML federation)——企业外部的XML数据可以作为虚拟维从三个不同的方面扩展OLAP立方体的表达能力：(a)OLAP的查询结果集中可以有XML数据；(b)可以用外部的XML数据作为产生OLAP立方体的查询谓词；(c) XML可以作为聚集维。文[45，46]对这种方法作了深入的分析，提出了查询的代价模型和查询优化方法。

这种逻辑上集成XML数据到OLAP立方体的意义在于提供了一种全新的集成XML到数据仓库的视角。

3 Web环境下数据仓库的体系框架及查询实现

传统的数据仓库一般采用集中式的体系结构，在这种体系结构中，ETL流程将所有业务处理系统中的数据按照统一的存储模型加载到一个中央数据仓库(物理上集中的)，中央数据仓库汇集了企业各部门业务处理系统的数据信息，同时也负责向各数据集市提供信息，其体系结构如图1所示。在这种体系结构中OLAP查询是在单一数据库上完成的。随着商业模式的变化，企业各部门跨地域分布，在广域网环境下数据仓库上进行联机分析时常常要跨网络上的多个结点，网络负载成为查询处理时要主要考虑的因素。如何充分利用现有的网络环境提高查询的响应时间，即一个查询从发出到第一个结果返回的时间，成为待解决的问题。

Web上的分布式数据库查询不同于传统的单一数据库查询除了异构数据的相互转换外还主要表现在：

1. 网络传输时间成为影响查询响应时间的主要因素。

2. 不同结点间的连接速度会有很大的不同。

3. 结点高度自治，很难找到精确的分割谓词使得全局数据U和各个结点的数据裂$v_i(1\leqslant i\leqslant m)$，满足$U=v_1\cup v_2\cdots\cup v_m$并且$v_i\cap v_j=\Phi(i\neq j)$。

而数据仓库OLAP查询不同于传统的数据库OLTP查询主要表现在：

1. OLAP查询是面向分析的，涉及的数据量非常大，响应时间长。

2. 多维数据集上查询所常用到的聚集函数(简记为aggr)如sum，count，min，max具有可分配性质，即设S是元组的集合，且被划分成n个水平裂片即$S=S_1\cup S_2\cup\cdots\cup S_N$，聚集函数$aggr$满足$aggr(S)=aggr(aggr(S_1),aggr(S_2),\cdots aggr(S_N))$。

针对OLAP查询的特点，Web上不同的数据仓库体系结构采取不同的查询解决方案：

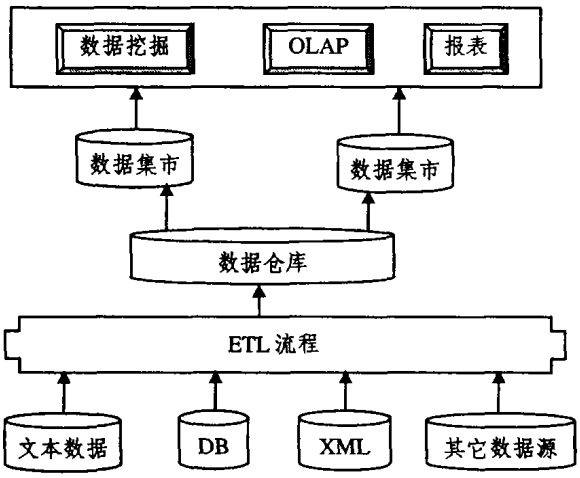

图1

(1)中央集成的树形体系结构

采用这种体系结构的Web数据仓库，跨地域分布的数据仓库结点形成一种树形结构，每一父结点的数据仓库包含其所有子结点的数据仓库的数据的信息。采用这种体系结构的优点在于每一结点的查询可以在本地实现，不需要跨广域网进行数据传输，传统的数据仓库查询优化方法都可以采用，从而可以大大提高查询处理的效率。这种方法的缺点在于存在大量的数据冗余，并且各个数据仓库结点的异构数据间其相互转换代价是很大的。

(2)层次体系结构

文[23]提出Web数据仓库采用一种DWH_Tree层次体系结构，父结点的数据仓库中不保存其所有子结点的全部数据，用户的查询需要本结点和其相应的所有从属结点联合完成实现。这种体系结构的优点在于：(a)具有小的数据冗余；(b)数据仓库的设计可以采用一种自下而上(先设计局部，后设计全局)的方式，总体开发成本小。针对这种数据部分重叠的层次体系结构，为了提高OLAP查询的响应时间，它提出了一种基于流水线并行的异步迭代查询方法。一般的同步查询，要求任一Web数据仓库结点只有在获得所有子结点的查

询结果集及本结点的查询结果集后才将合并后的查询结果集传给其父结点，从而在各个层次的父子结点间会有大量的相互等待时间。异步迭代查询方法主要基于数据传输和数据处理的并行，即某一 Web 结点只要获得了其结果集的一个数据块时，就将其传输给父结点，下一个数据块计算出再依次上传给父结点，从而实现各个结点的数据聚集计算和结点间的数据传输的并行。该文从理论上证明了异步相对于同步的效率要高出许多，为进一步分析提供了理论依据。文[24]提出的一种数据仓库集成的方法，也是基于这种层次体系结构来实现的。

(3)多层代理结构

为了满足广域网下多用户的 OLAP 分析需求，避免形成瓶颈，文[25]提出一种面向应用的多层代理 Web 数据仓库结构。采用这种代理结构的 Web 数据仓库体系结构分成三层：数据仓库层，OLAP 缓存服务器层(OLAP cache server 简记为 OCS)和客户端。OCS 不仅动态地存储 OLAP 查询结果集，并且可以对 OCS 上现存的数据进行聚集计算。采用这种体系结构的查询处理过程如下：客户端将查询发到最邻近的 OCS，OCS 比较本地计算的代价和根据基于启发式算法计算出通过其它邻近的 OCS 取得结果的代价来决定取得数据的途径。这种体系结构的意义在于：(a)利用缓存技术提出一种具有计算能力的 OLAP 代理服务结构来解决 Web 上的 OLAP 查询效率问题，OLAP 查询的数据从物理传输代价最小的 OCS 上取得。(b)对查询结果的路由采取了一种启发式的基于代价的优化选择算法。从而为 Web 环境下的 OLAP 查询提供了一种可行的途径。

4 几种有代表性的 Web 数据仓库及相关研究

目前 Web 数据仓库的研究主要有以下几种代表性的原型。

4.1 基于本体(ontology)集成半结构化数据

将 Web 上的数据抽取到数据仓库中的关键一步是首先要建立源数据(Web 上的数据)和目的数据(比如关系数据仓库)的映射关系。这是模式匹配研究的课题，目前模式匹配的实现分成两个层次，一个层次的模式匹配[26,27]是语法匹配：即根据两种数据模式元素的名字，数据类型，约束等来寻找源和目的数据模式元素的相似性建立匹配；另一层次模式匹配是语义匹配，即寻找源模式和目标模式中相等的语义对象来建立匹配。文[4,28,29]中的 Web 数据仓库是基于语义对象模型 MIX[30](metadata based integration model for data X-change)集成半结构化数据的数据仓库系统。MIX 中有两种语义对象模型：简单语义对象模型和复杂语义对象模型。一个简单的语义对象是一个三元组$\langle C,V,S\rangle$，其中 C 表示语义对象的本体概念，V 是语义对应的值，S 是上下文环境。一个复杂的语义对象表达成$\langle C,A\rangle$，其中 C 是本体概念，A 是简单语义对象的集合。依据语义对象模型。文[4,29]分别论述了两种不同的概念模式和数据库模式映射建立方法。前者对于每一个特定的 Web 源的 MIX 表达式和相应的目的数据模式，采用特定的 MRL(mapping rule language)语言来建立相应的映射文件(描述映射过程)，来实现特定的映射。而后者将 Web 数据表达为一个 MIX 树，目的数据模式的星型结构亦表达为一个等深的 MIX 树，从而根据树的重构来自动实现映射。这种基于 MIX 语义对象的映射实现具有如下的优点：(a)采用本体，解决了语义识别问题。(b)可应用于不同应用领域的数据集成，Web 数据源不须要针对某一特定的领域。

总之，语义等价是异构数据源数据集成的目标，基于 MIX 模型的数据仓库是这一方法的典型代表。

4.2 WHOWEDA[31](warehouse of Web data)

WHOWEDA 将整个 Web 的拓扑结构看成是若干个有向图的集合，有向图的每个结点代表 Web 站点的文档，边代表超链。目前大部分 Web 数据仓库中只存储 Web 结点上的 Web 文档数据内容，而对边即超链接却没有加以利用。WHOWEDA 中将 Web 文档的数据内容，Web 文档间的拓扑结构，元数据信息以视图的形式保存在数据仓库中。数据仓库中一个数据库由若干个表组成，每一个表中存储了一个 Web schema 所绑定的相关 Web 数据的元数据、结构和内容。这种利用 Webschema 对 Web 进行查询的意义在于：

(1) 为 Web 上的检索提供了模式信息，从而对 Web 的查询可以在数据库中查询执行。

(2) 可以在数据仓库中利用提供的模式信息对查询计划进行优化，加快查询。

它使传统的依赖于网页间拓扑结构的 Web 导航式浏览转变为对 Web 数据仓库的表的结构化查询，大大提高了查询精度。文[32]中提出的 Web 代数，是 WHEWODA 系统规范化的数据表达和操作的理论基础。文[33～36]对 WHEWODA 中的相关查询语言，检索数据的选择，增量处理，查询执行，模式获取及求精等进行了详细的论述。总之，WHOWEDA 的特点在于将 Web 的超链作为数据内容存储在数据仓库中，从而提取了 Web 的拓扑结构，从而为 Web 的结构化查询提供了可行途径。

4.3 XML 数据仓库

随着 XML 成为网上数据传输和交换的标准，人们在研究如何充分地利用 XML 来为数据仓库的决策服务，xyleme[37,38]是一个具有代表性的 XML 数据仓库系统，它是一个专门为 XML 这种半结构化数据定制的数据仓库，可以存储数亿的 XML 网页，它的主要意义在于用数据仓库的方式来存储 XML 数据，使得以往在网上不能实现的跨 XML 网页查询 XML 网页元素变为现实。其主要技术特色在于：

(1)以一种流字节和结构化相结合的方式存储 XML 文档。

(2)利用完全倒排索引实现树模式匹配来提高查询效率。

(3)网页选择，获取及刷新策略，及网页的增量变化处理方法。

(4)语义集成 XML 数据。

XML 作为一种半结构化数据的典型代表，如何以数据仓库的方式加以利用目前成为人们研究的热点，在查询优化，XML 网页增量处理，还有大量工作有待于进一步探索。

5 相关问题及展望

Web 数据仓库作为一种面向 Web 环境的数据仓库，许多相关的问题对于 Web 数据仓库有待于进一步研究：半结构化数据的模式提取作为集成 Web 上的半结构化数据到数据仓库中的前提条件。HTML 的信息抽取在前文已论述，对于 XML 的模式提取主要有基于图论的方法[39]和基于规则的方法[40]和基于概率[41]的方法等，这方面的研究是集成半结构化数据到 Web 数据仓库的前提；实视图的增量处理是关系型数据仓库中的一个重要问题，在 Web 数据仓库中，它既可以以关系视图的增量维护来解决，也可以用半结构化数据视图(如 XML view)增量维护方式来处理[42,43]，两类方法如何相互补充地解决 Web 数据仓库中的视图增量维护是值得进一步思考的；Web 数据仓库的查询既要利用已有的集中式数据仓库的 OLAP 查询的技术，也要充分考虑到这种查询的物理分布

特性[23,25,44]。

总之，Web 上的数据以指数的速度在增长，然而以浏览器，搜索引擎方式所获得的数据信息常常是费时的和不精确的，而数据仓库技术作为一种决策支持系统的基础，以往主要囿于结构化数据的管理，如何充分地利用 Web 上的数据信息，为决策分析提供更加完备的数据，成为一种要求。目前所见到的几个具有代表性的 Web 数据仓库原型都牵涉到如分布式查询处理，半结构化数据的模式提取，增量处理，模式匹配，模式求精等诸多问题，在不同的程度上给出了答案。从技术上讲 Web 数据仓库融合了数据库技术，WWW 技术，数据挖掘、信息检索和语义 Web 等领域，如何结合这些技术来更好地解决这些问题还有待于进一步探索。

参 考 文 献

1 Inmon W H. Building the Data Warehouse (Chinese verstion). John Wiley &Sons Inc, 1992. 20

2 The Asilomar Report on Database Research. SIGMOD Record, 1998, 27(4)

3 Ambite J L, Ashish N, Barish G. ARIADNE: a system for constructing mediators for internet sources. In: Proc. of the ACM SIGMOD intl. conf. on management of data, seattle, USA, 1998. 561～563

4 Zhu Y, et al. Materializing Web data for OLAP and dss. WAIM'00 shanghai, China, 2000, LNCS(1846): 201～214

5 Hammer J, Garcia-Molina H, Cho J, Aranha R. Extracting Semistructured information from the Web. In: proc. of the workshop on management of semistructured data, 1997. 18～25

6 Atzeni P, Mecca G. Cut and Paste. Journal of Computer and System Sciences, 99: 453～482

7 Embley D W, Jiang Y, Ng Y-K. Record-Boundary Discovery in Web Documents. In: Proc. of the 1999 ACM SIGMOD intl. conf. 1999. 467～478

8 Embley D W, et al. Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages. In: Proc. ACM SIGMOD Intl. Conf. on Management of Data, 1999. 467～478

9 Muslea I, Minton S, Knoblock C A. Hierarchical Wrapper Induction for Semistructued Information Sources. Autonomous Agents and Multi-Agent System, 1 1/2(2001), 93～114

10 Crescenzi V, Mecca G, Merialdo P. ROADRUNNER: Towards Automatic Data Extraction from Large Web Sites. In: Proc. of the 27th VLDB Conf. Roma, Italy, 2001. 109～118

11 Chang C-H, Liu C-C. IEPAD: Information Extraction Based on Pattern Discovery. In: Proc. of the ACM WWW10 Conf., 2001. 681～688

12 Buttler D, Liu L, Pu C. A Fully Automated Object Extraction System for the World Wide Web. In: Proc. of the 2001 Intl. Conf. on Distrubuted Computing Systems (ICDCS'01), 2001. 361～370

13 Shanmugasundaram J, et al. relational database for querying xml documents: limitations and opportunities. In: Proc. of VLDB, Edinbergh, Scotland, 1999. 302～314

14 Schmidt A, Menzo M K, Waas W F. Efficient relational storage and retrieval of xml documents. In: Workshop on the Web and Databases (WebDB), 2000. 47～52

15 Florescu D, Kossmann D. Storing and querying xml data using an RDBMS. bulletin of the IEEE computer society technological committee on data engineering, 1999, 22(3): 27～34

16 Shanmugasundaram T, Krishnamurthy R, Tatarinov L. A general technique for querying xml documents using a relational database system. ACM SIGMOD Record, 2001. 20～26

17 Yoshikawa M, Amagasa T. XREL: a path-based approach to storing and retrieval of xml documents using relational database. ACM Transactions on Internet Technology (TOIT), 2001, 1(1): 110～141

18 Khan L, Rao Y. A performance evaluation of storing xml data in relational database management systems. In: Proc. of the third intl. workshop on Web information and data management, 2001. 31～38

19 Deutsch A, Fermandez M, Suciu D. Storing semistructured data with STORED. In: Proc. of the 1999 ACM SIGMOD intl. conf. on Management of data, 1999. 431～442

20 Bohannon P, Freire J, Roy P. From xml schema to relations: a cost-based approach to xml storage. In: 18th Intl. Conf. on Data Engineering (ICDE'02), 2002. 64

21 Kim J, Lee W Y, Lee K. The cost model for xml documents in relational systems. IEEE, 2001

22 Pederson D, Riss K, Pederson T B. Xml-extended OLAP querying. In: 14th Intl. Conf. on Scientific and Statistical Database Management, 2002, 195

23 何震瀛，李建中，高宏. Web 数据仓库的异步迭代查询处理方法. 软件学报, 2002. 214～218

24 Triantanfllaks A, Kanellis P, Martakos D. Data warehouse clustering on the Web. In: 13th Intl. Workshop on Database and Expert Systems Applications (DEXA'02), 2002. 800

25 Kalnis P, Papadias D. Proxy-server architecture for OLAP. In: Proc. of the 2001 ACM SIGMOD intl. conf. on Management of data, 2001. 367～378

26 Milo T, Zohar S. Using schema matching to simplify heterogeneneous data translation. In: proc. of the 24th VLDB conf. New York, USA, 1998. 122～133

27 Beeri C, Milo T. schemas for integration and translational of structured and semi-structured data. ICDT'99, 1999; LNCS (1540): 296～313

28 Zhu Y. A framework for warehousing the Web contents, ICSC'99, Hongkong, 1999. 296～313

29 Zhu Y, Bornhovd C. Data transformation for warehousing Web data. Advanced Issues of E-Commerce and Web-Based Information Systems, 2001. 74～85

30 Bornhovd C. Semantic metadata for the integration of Web-based data for electronic commerce. WECWIS'99, satan Clara, USA, 1999. 137

31 Bhowmick S S, Ng W K, Madria S. Web schemas in WHOWEDA. DOLAP'00 USA, 2000. 17～24

32 NG W-K, LIM E, Huang C-T, Bhowmick S. Web warehousing: an algebra for Web information. In: proc. of IEEE intl. conf. on advances on digital libraries (ADL'98), 1998. 228～237

33 Sourav S, Ng B W. Anatomy of the coupling query in a Web warehouse. International Journal of Information and Software Technology, 2002, 44(9): 513～539

34 Sanjay K, Madria, Souras. Ranking of Web data in a Web warehouse. In: proc of the third intl. conf. on Web information systems engineering, 2002. 130～139

35 Bohmick S S. Detecting and representing relevant Web deltas in WHOWEDA. IEEE transaction on knowledge and data engineering, 2003, 15(2)

36 Bhowmick S S, Madria S. Controlling Web query execution in a Web warehouse. In: proc. of the 13th intl. workshop in database and expert system application, 2002. 805～809

37 Abiteboul S, Cluet S, Ferran G, Rousset M-C. The Xyleme Project. Computer Networks and ISDN Systems Journal, 2002, 39(3): 225～238

38 Xyleme L. A dynamic warehouse for XML data on the Web. IEEE Data Engineering Bulletin, 2001, 24(2)

39 Goldman R, Widom J. Summarizing and searching sequential semistructured sources: [Technical Report]. March 2000

40 Nextorov S, Abiteboul S, Motwani R. Extracting Schema from Semistructured Data. In: Proc. SIGMOD'98, SIGMOD Record. 1998, 27(2): 295～306

41 Florescu D, Koller D, Levy A. Using probabilistic information in data integration. In: proc. VLDB97, 97. 216～225

42 Liefke H, Davidson S B. Efficient View Maintenance in XML Data Warehouses: [Technical Report MS-CIS-99-27]. 1999

43 Cob'ena G, Abiteboul S, Marian A. Detecting changes in xml documents. ICDE, 2002. 41～52

44 Yerneni R, et al. Fusion queries over internet databases. In: Proc. of the Intl. Conf. on Extending Database Technology (EDBT), 1998. 57～71

45 Pederson D, Torben K R, Pederson B. Query processing and optimization for OLAP-xml federations. In: Proc. of the 5th ACM intl. workshop on Data Warehousing and OLAP, 2002. 57～64

46 Pederson D, Torben K R, Pederson B. Cost modeling and estimation for OLAP-xml federations. In: Proc. of the 5th ACM intl. workshop on Data Warehousing and OLAP, 2002. 57～64