·新技术应用·

ETL技术及其在数字图书馆中的应用研究

黄永文 (中科院文献情报中心 北京 100080)

李广建 (北京师范大学管理学院 北京 100875)

文 摘 简要介绍了ETL的背景,分析了ETL的实现过程和体系结构,对国内外关于ETL的研究内容和现状进行了详细论述,最后提出了ETL技术在数字图书馆领域中的应用。

关键词 信息抽取 数据转换 数字图书馆

Research of ETL Technology and Application on Digital Library

Huang Yongwen

(Library of Chinese Academy of Sciences, Beijing100080)

Li Guangjian

(School of Management, Beijing Normal University, Beijing100875)

Abstract: The paper introduces the background of ETL (Extraction - Transformation - Loading), analyzes its architecture and functions, and discusses its current situation and research focus. At last, presents ETL applications on digital library.

Key words: Information extraction, Data transformation, Digital library

1 引言

在网络环境下,存在着大量的异构系统、庞杂的资源和分散的知识,来源不同、分散和不清洁的信息与人们的信息需求之间的矛盾,以及不断集成化的信息服务,呼唤着信息集成和整合的有效方式和方法。具体来说,主要来自于以下两方面的驱动:

(1) 信息环境的驱动

① 存在大量非结构化信息且变化频繁。WEB上的信息隐藏在页面中,信息的类型、格式、约束、以及与其它信息的关联都没有明确的定义,呈现出非结构化或半结构化的特征,需要特殊的处理程序来分析和识别。同时,WEB上的信息始终在不断变化之中,这不仅仅是指信息的数量,也包括新的数据类型、数据格式以及包含这些信息的页面的结构(样式)。

② 存在大量异构的数据源和自治系统。这不仅表现在现有的各个WEB站点的软硬件平台、数据库各不相同,更重要的是对数据内容的表现方式也不相同。每个WEB站点都独立设计、实现并运行,具有完整的功能,而且相互之间关联很弱,具有相同语义内容的数据往往表现完全不同。

(2) 信息需求方面的动因

与传统信息需求相比,在网络环境下,用户的信息需求发生了很大的变化,用户的信息需求呈现出社会化、综合化、集成化和高效化的特征。信息需求的变化对网络化信息服务起着决定作用,未来的网络信息服务应该是以主动的方式、无缝地整合有效资源、为用户提供高效、贴切的服务。这种网络信息服务模式的实现要以信息资源整合和集成服务系统的建立为基础,没有完备的资源整合体系作后盾,没有集成信息服务的理念作支撑,这种信息服务模式是无法实现的。信息需求的变化,提出了网络环境下信息资源集成和整合的要求。

ETL技术正是在这一目标的引导和推动下发展起来的一种资源集成与互操作手段,ETL可以应用在电子商务、智能信息检索、数字图书馆、WEB信息挖掘、信息门户等诸多领域。ETL(Extract - Transform - Load)是一个来源于数据仓库的概念,指抽取(Extract)、转换(Transform)、清洗(Cleaning)、装载(Loading)的过程。ETL是按照特定的应用需求,将WEB上特定数据源中的信息抽取、识别、整理、规范和存

储,并在此基础上实现高效的查询和比较,乃至数据挖掘、知识发现等应用。文章主要对 ETL 的研究现状以及其在数字图书馆中的应用进行研究,希望能为基于 Cyberinfrastrure、网格(Grid)、E - science 等环境下的数字图书馆的建设和服务提供借鉴。

2　ETL 的实现过程及体系结构

ETL 的基本实现过程:数据抽取引擎从不同的数据源中进行完全或差异性抽取,这些数据来源可以是关系数据库、文件等,然后将抽取出来的数据存放在 DSA(Data Staging Area,据存储区),在上载到目标数据仓库之前进行数据的转换和清洗。在 ETL 的体系结构中主要包括[1-2]:通用数据接口、数据抽取、数据集成、数据清洗、数据装载、系统管理等,ETL 的体系结构如图 1。

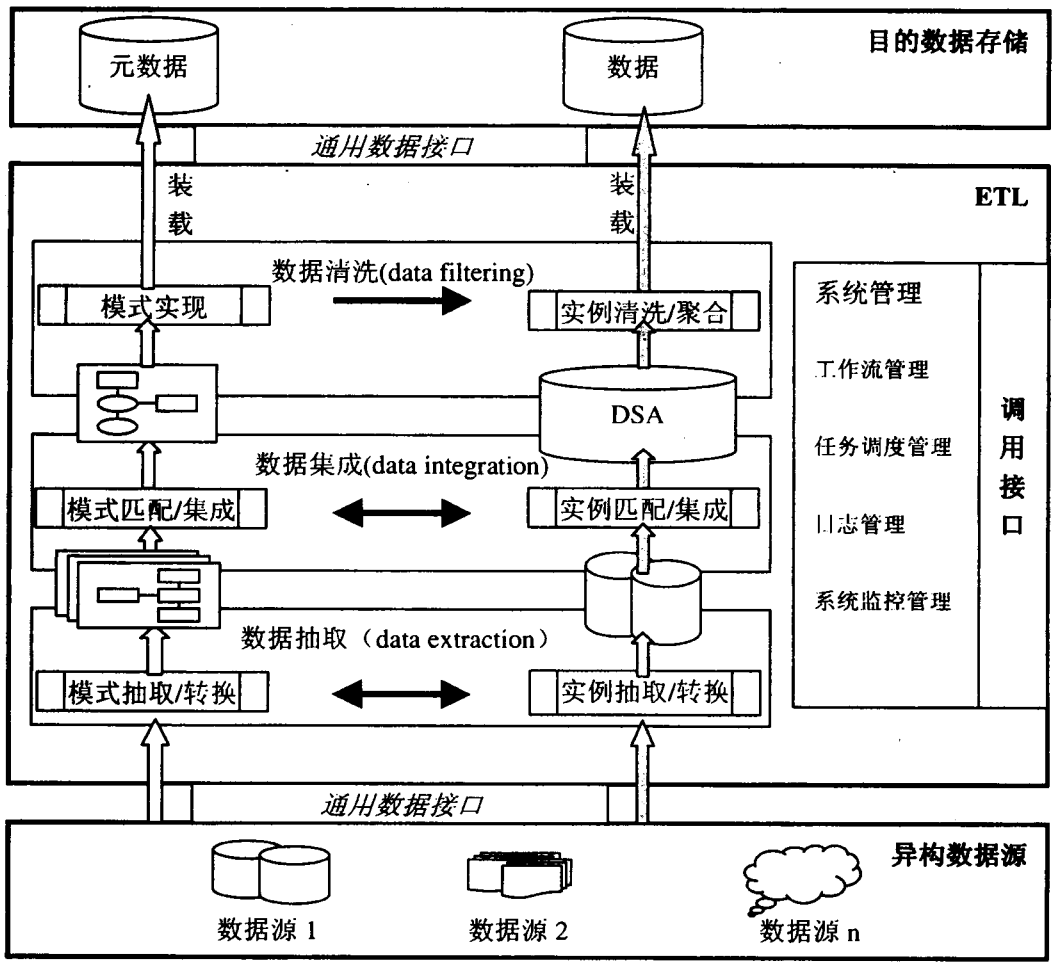

图 1　ETL 的体系结构图

(1) 通用数据接口:支持多种数据源,如各种关系型数据库、网站、各种格式的文件等。该接口能够跨平台访问数据源,支持在不同类型数据源之间建立连接。通过数据接口可以屏蔽各种数据源之间的差异,提供统一的数据视图。

(2) 数据抽取:包括模式数据抽取和实例数据抽取。先从数据源中抽取模式信息,然后进行人工分析或智能分析,形成实例数据的抽取策略,并将其存储在知识库中作为装载数据的依据。

(3) 数据集成:数据抽取后形成多个模式和实例数据集,但是最终需要的是经过集成的和语义一致的数据,因此必须对数据集进行映射,形成统一的结果集。在此过程中,需要进行数据格式统一、数据标准化、一致性校验、修改内容上的错误等,最后将处理后的数据存储在 DSA 临时区域中,等待进一步的清洗。

(4) 数据清洗:经过数据集成后的数据集中还包含许多相似重复记录,它会影响数据的语义一致

性,因此需要进行消除。在此过程中,针对数据集进行匹配,即发现重复异常,删除部分记录或者将多个记录合并为更完整的记录。

(5) 数据装载:此过程主要解决数据装载时机等问题,有选择地将经过清洗后干净的数据集装载到一个或多个目的数据表中,并允许人工干预。

(6) 系统管理:工作流管理主要负责任务与ETL步骤的定制;任务调度管理负责分配ETL任务与步骤的运行时机;日志管理提供ETL过程中的所发生事件备忘录;系统监控负责监视ETL过程的动态运行情况,报告运行时错误、系统异常。

(7) 调用接口:调用接口主要是使ETL可以方便地嵌入到不同的应用中。为了获得更好的交互性与可扩展性,ETL应该提供类SQL的描述性语言,方便定制高效执行的数据处理流程。

3 国内外研究现状

通过对国内外相关文献和项目的研究和分析,目前关于ETL的研究主要集中在以下几个方面:

(1) 关于ETL模型的研究

① 基于工作流的ETL模型

工作流是一套相关活动的集合,活动按照一定规则串联便形成工作流。工作流系统一般具有高度的灵活性,主要表现在对活动的构造和描述上,这些活动可以是可递归分解的,也可以是由子活动合成的组合活动。基于工作流的ETL模型的设计思想是将数据仓库的构造过程作为一个工程看待,将该过程拆分成具有不同操作和语义的活动,用工作流的概念定义它。活动之间通过事件进行协调,事件的定义和管理通过中心知识库完成。

文献[3]给出了一种以工作流为模式、以中心知识库为中心、基于元数据管理的ETL模型。中心知识库存放用户定义的有关数据抽取、数据转换、数据加载等过程的工作流组织方式,构成每个过程的相应活动的定义以及用户定义的商业规则、转换规则和变化规则等控制信息。

② 基于CWM的ETL模型

CWM(Common Warehouse Metamodel,通用数据仓库模型)是一组元模型,目的是为了在数据仓库工具、数据仓库平台和数据仓库存储之间建立一个商务智能元数据的交换机制。CWM覆盖了数据仓库应用的整个生命周期,包括数据源表达、分析、仓库管理以及典型的数据仓库环境基础组件。数据源元模型支持对多种数据源进行建模;分析元模型用于对数据转化、OLAP、信息可视化以及数据挖掘的建模;仓库管理元模型负责仓库处理、动态跟踪以及时间规划方面的建模;基础组件元模型支持多种通用的数据仓库元素和服务,例如数据类型、类型系统映射、抽象键值和索引、表达式等。

文献[4]介绍的ARKTOS系统提出ETL统一元模型包括:数据仓库架构、活动模型、偶然事件处理和质量管理。ARKTOS可以通过捕捉一般任务进行建模和执行ETL过程。ARKTOS提供三种方式来描述和实现ETL,图形化的前端工具和两种描述语言:XADL(XML变量)和SADL(类似于SQL语言)。文献[5]提出了统一的ETL元数据模型,对ETL过程中涉及的操作进行分类建模。ETL体系结构中主要包括ETL元数据对象模型、ETL任务建模以及ETL任务模型的描述语言(XTDL)。元数据对象模型主要分为:过程元数据模型、数据质量元数据模型和数据元数据模型,采用面向对象方法设计元数据对象实体。

(2) 关于ETL元数据的研究

元数据是关于数据的数据,对于ETL来说尤其重要。ETL中大量的数据源定义、映射规则、转换规则、装载策略等都属于元数据范畴,如何妥善地存储这些信息已经关系到ETL过程能否顺利完成而且影响到后期的使用和维护。目前研究和使用较多的标准主要有[6-7]:MDC(Metadata Coalition,元数据联合)、XMI(XML-based metadata Interchange,基于XML的元数据交换)、CWM等标准。

文献[2]主要研究了数据仓库的元数据管理,包括元数据的类型和模型;联合元数据架构;基于UML的元数据标准,如OIM(Open Informaiton Model,开放信息模型)、MDC、CWM等。描述了ETL中的数据流和元数据流,指出了在数据抽取、集成和聚合过程中数据和元数据的处理过程、步骤以及规则。

(3) 关于ETL模式集成问题的研究

ETL需要提供相应的功能来解决由于模型和结构不同而造成的数据源之间的冲突问题,模式冲突解决在ETL中称作模式集成,目前常用的模式集成方法主要有[8-10]:基于模型和启发式算法的模式集成方法、数据模式再工程/映射方法、人工智能(AI)方法、元数据方法、词汇语义法、面向对象法等。

① 基于模型和启发式算法的模式集成方法。采用通用模型来描述各种不同的概念,通过人工来分析各数据源的数据模型,预定义了一些模式变换过程、启发式算法,用户可参与集成过程,且具备一定的可扩展性。此方法可以解决大部分的结构冲突,但语义方面需要手工解决,因此其自动化程度很

低,难以运用到大型、复杂的应用领域。

② 数据模式再工程/映射方法。利用自动的映射工具比较不同模式之间的异同性,将待集成的各种模式映射到公共模型上以实现集成。该方法可实现各种概念的重命名、聚合、对象化等多种变换操作,可解决模式集成中大部分的语法和语义冲突,但对应用不透明,需要设计员做大量工作,且其中的启发性算法过于复杂,效率较低。

③ 人工智能(AI)方法。每个待集成的模式作为成员模式,其中的实体、关系、数据模型等知识都被集中到系统的知识库中。此方法的不足之处在于需要预先存储含有大量知识条目的知识库,而对其进行查询所需的时间非常可观;而且知识库中存储的知识往往不够严密。但 AI 方法对模式集成时模糊查询,以及语义分析等方面具有启发性意义。

(4) 关于 ETL 中的数据质量和清洗问题

信息和数据的正确性(Correctness)、一致性(Consistency)、完整性(Completeness)和可靠性(Reliability)尤为重要,然而现存信息和数据存在很多的问题,容易造成脏数据,主要原因有:滥用缩写词、惯用语、输入错误、数据中的内嵌控制信息、重复记录、缺损值、拼写变化、不同的计量单位和过时的编码等。这些脏数据和信息可能带来操作不便、决策制定失败甚至错误等。因此 ETL 过程中必须对脏数据和信息进行有效处理,确保数据和信息的质量。数据清洗有很强的领域相关性,过去一段时间内通用的清洗框架很少有人关注,目前数据 ETL 已经成功地应用到数据仓库(DW)、数据库中的知识发现(KDD)和总体数据质量管理(TDQM)等诸多领域,所以通用的集成清洗方案越来越受到重视。

国外关于数据质量问题的研究开始较早,且非常活跃,研究领域涉及以下几个方面:研究高效的数据异常检测算法以避免扫描整个庞大的数据集;在自动化异常检测和清洗处理间增加人工判断处理以提高处理精度;实例层次上基于语义的数据集成;数据清洗时对海量数据集进行并行处理;如何消除合并后数据集中的重复数据,建立一个通用的与领域无关的数据清洗框架。由法国 INRIA 开发的 AJAX 系统[11]可以用来处理典型的数据质量问题,例如:对象同一性问题、拼写错误和记录之间数据矛盾问题。文献[12]讨论了数据清洗的一个子问题,即将地址字段分为不同的元素,并且使用训练隐马尔科夫模型(HMM)来解决该问题。文献[13]介绍了 Potter's wheel 系统,该系统为用户提供了交互式方式实现数据清洗过程。

(5) 关于 ETL 相关技术的研究

在信息抽取方面的研究,例如 Olera 信息抽取系统[14],可以从半结构化 WEB 文档产生抽取规则,不需要详细的训练文档集的注释,Olera 的性能和效果比较好,只需要少量的由程序自动产生的网页和有限的用户干预;OntoBuilder 项目[15]中定义了输入到在线数据库的 HTML 格式的分层规则,采用表述本体(presentation ontology)来对深层网络(deep web)进行信息抽取;文献[16]主要研究 WEB 资源的抽取问题,构建了一个交互式系统,用来进行半自动化地构造 WEB 资源的分装器(Wrapper)。

在规则构建方面的研究,文献[17]提出了一个规则学习系统,它可以学习如何构建一系列的映射规则并将其应用到特殊的应用领域中;在语义冲突和数据转换方面的研究,文献[18]讨论了用半自动化/自动的方法来支持在数据集成中识别和解决语义冲突,提出了多个数据源之间的潜在语义冲突的发现方法,同时提出了解决问题的数据转换技术;在数据调整方面的研究,文献[19]实现了基于集成强制技术的数据调整工具,主要是用来辅助生物剂量学领域中数据整合的设计。

(6) 主流的 ETL 工具

目前市场上主流的 ETL 工具可以分为两大类:一类是专业 ETL 厂商的产品,这类产品一般都具备较完善的体系结构,具有复杂和详尽的功能;另一类是整体数据仓库方案供应商,他们在提供数据仓库存储、设计、展现工具的同时也提供相应的 ETL 工具,这类产品一般对自己厂商的相关产品有很好的支持,但结构相对封闭,对其他厂商产品的支持很有限。

专业 ETL 厂商的产品包括 Ascential 公司的 DataStageXE、Data Junction 公司的 DataJunction、SAS 公司的 Data Builder、Sagent 公司的 Solution、Informatica 公司的 ArdentDataStage 等,整体方案提供商的产品则包括 Oracle 公司的 Warehouse Builder、IBM 公司的 Visual Warehousing、Micmsoft 公司的 DTS 等。这些 ETL 工具各有特点,功能很强,但在开放性、伸缩性和集成性等方面有待加强。

4　ETL 技术在数字图书馆中的应用

随着数字图书馆采集技术、存储技术、检索技术的进步,数字图书馆的技术瓶颈日益集中在数字信息的深层次整合和分析处理之上。面对数量日益庞大的数字信息堆积,如何高效地对这些数字信息进

行加工处理,有效地实现这些数字资源的开发利用成为当前数字图书馆研究者必须面对的一个课题。可以将ETL技术运用到信息资源建设、参考咨询服务、情报研究等方面上,来解决和缓解数字图书馆的建设和服务中遇到的问题。

(1) 在网络信息采集系统中的应用

可以在网络信息采集系统的爬行和装载过程中运用ETL技术。网络信息采集系统在进行信息资源爬行和下载时,需要对WEB网页进行识别和抽取,主要是对网页的一些特定标识进行识别,由于网站是不断发生变化的,经常会发生格式变化或出现一些新的标识,需要修改程序,造成使用上的不便。另外,由于网络信息采集系统侧重于网页相关度的判断,对格式和质量的核实和检查相对弱一些。通过ETL技术对改善上述问题以及采集系统存在或忽略的其他问题,从而提高系统采集信息资源的准确性和效率。

(2) 在分布式资源建设中的应用

在目前的数字资源共同建设模式中,通常采用的是分布式加工方式,各个加工单位向数据中心提交加工后的数据,然后由数据中心定期进行数据更新,对外提供统一集成的服务。可以将ETL技术运用在资源的共建共享中,由数据中心定期向各个加工单位进行数据抽取(全部或增量)、转化、清洗等,简化资源建设的流程,提高资源建设的效率。

(3) 在信息整合中的应用

可以将ETL技术应用到信息整合系统中,如在异构信息资源的物理整合和逻辑整合的应用。在信息资源的物理整合中,ETL对信息进行抽取、转换并将其装载到实际的数据库中。在信息资源的逻辑整合中,ETL表现为封装器(Wrapper),是受用户需求驱动对信息进行整合,并不将其装载在实际的数据仓库中。在异构数据源的选择和定位上,可以结合"调焦查询探测"(focused query probes)、"基于提问取样"(Query-based sampling)等算法,提高整合的效率和速度。

(4) 在协作式参考咨询中的应用

可以将ETL技术应用到分布式知识库的协作机制中,实现知识库的自动问答。在协作式的参考咨询服务中,各个协作单位的咨询知识库是分布的,在回答读者用户时需要分别到各个知识库中进行匹配;在本地的参考咨询系统中嵌入ETL引擎,先在本地知识库进行问题的匹配,如何未找到答案,调用ETL引擎到协作网中的其他咨询系统中的咨询库中进行抽取匹配,如果成功进行答案的整理并上载到本地知识库中,从而实现基于分布式知识库的自动问答。

(5) 在情报研究中的应用

ETL技术可以应用到情报研究中,特别是在竞争情报中的应用。ETL并不仅仅是从WEB上获取数据并保存下来,更主要的是从结构化数据、非结构数据或自由的文本中发掘出更深层次的信息和内容、实现知识层次的发现和整合。利用面向知识抽取的ETL技术,将分散在网络中的数据、数值、信息和知识进行集成,并加载到本地的数据库中,作为支持决策服务、竞争情报分析、OLAP等的基础。

参考文献

1 周宏广等.数据ETL工具通用框架设计.计算机应用,2003,23(12):96-98

2 Erhard R. Metadata management for heterogeneous. Information Systems. [2004-12-10] http://dbs.uni-leipzig.de/Research/meta-Dateien/meta2.pdf

3 周志逵,徐先传.数据仓库中数据抽取、转换及加载工具研究.北京理工大学学报,2003,23(6):720-723

4 Vassiliadis P, et al. Arktos: Towards the modeling, design, control and execution of ETL processes. Information Systems, 2001, 26(8): 537-561

5 贾自艳等.面向数据质量的ETL过程建模与实现.系统仿真学报,2004(5): 907-911,914

6 吴建芹等.基于CWM的企业数据仓库体系结构设计.计算机工程与应用,2002(21): 202-204

7 李姗姗,宁 洪,彭绍亮.基于CWM的元数据管理系统中数据交换格式的研究.计算机工程与应用,2004(14): 192-195

8 刘 浩,周宏广.数据ETL过程中的模式集成技术研究.湘潭师范学院学报(自然科学版),2004,26(3):53-56

9 张 斌,黄中庸.面得对象多数据库系统的模式集成.东北大学学报:自科版,1997,18(4): 395-399

10 石祥滨,赖翔飞.SCOPE/CIMS系统中模式集成的形式化基础.计算机学报,1998,21(11):1015-1021

11 Galhardas H, et al. AJAX: an extensible data cleaning tool in Proceeding of the ACM SIGMOD. International Conference on the Management of Data. Dallas: TX, 2000

12 Vinayak R B, Kaustubh D, Sunita S. Automatically extracting structure from free text addresses. Bull. Techn. Committee Data Engineering, 2000, 23 (4): 27-32

13 Raman V, Hellerstein J, Wheel P: an interactive framework for data cleaning and transformation, Technical Report, University of California at Berkeley, Computer Science Division, 2000

(下转第54页)

度的制定、设备采购方案、人员使用培训等等,需要投入的资金是很有限的,目前,东莞地区所建立的九个分馆,除了支付网络费用、购买计算机及相关网络连接设备之外,基本没有其它的费用。这就使得分馆的建设大大降低了风险和门槛,使基层馆更容易接受。

(2) 客户端实现零维护。东莞图书馆网络集群系统采用了B/S结构,客户端的要求很低,只要能够上网,就可以连接到总馆的服务器上,实现分馆的所有图书馆业务工作。所谓"零维护",就是说处于客户端的各分馆,只要能掌握系统的一般操作和简单管理,就可以正常地使用该系统,而不需要有专门的图书馆专业和计算机专业的技术人员负责系统的维护工作,这无疑为基层图书馆解决了很大的专业技术人员缺乏问题,更符合目前基层图书馆的情况。

(3) 增加"馆藏",共享资源。东莞图书馆分馆可以通过区域通借通还的功能,使分馆读者可以借阅到东莞区域其它分馆的文献资源,这实际上就相当于增加了其"馆藏"资源,大大提高了所在分馆读者借阅文献资源的范围和数量。同时,还可以以分馆正式读者的身份享受中心馆的所有数字资源和东莞地区共享工程资源,以及中心馆提供的各种资源。

"东莞图书馆总分馆管理技术平台"可以提供图书馆集群管理服务,即服务器和系统管理由总馆负责,用户只需通过互联网登录到总馆服务器就可以使用图书馆集群网络管理平台,大大降低中小型图书馆使用自动化技术的经济和技术门槛,特别适合目前我国城市图书馆事业的现实状况和发展需要,可以较好地激活图书馆原有资源存量,提高整体工作效率,扩大图书馆服务范围和服务能力。该系统在东莞地区投入使用一年半以来,实践验证了其具有的经济、高效、实用、方便,初步实现了东莞地区动静结合的图书馆总分馆形态,突破了制约城市图书馆事业发展的瓶颈——关键性技术实现障碍,奠定了东莞图书馆总分馆集群化管理的基础,大大扩展东莞图书馆的服务效能,实现了基层分馆业务工作的跨越式发展,提升了东莞城市图书馆事业的整体发展水平,也为共享工程建设奠定了有利的基础。

参考文献

1 张 群.中心馆制的强大生命力.新世纪图书馆,2005(3)

2 何 战.新加坡图书馆事业的发展.东南亚纵横,2004(12)

3 薛晓枫.地级市公共图书馆发展初探.科技情报开发与经济,2004,14(2)

4 祖 央.图书馆自动化建设的新模式 甘肃省图书馆基于Unicom自动化管理系统的中心/成员馆方案.图书与情报,2005(3)

5 杨思洛.国内图书馆自动化系统分析研究.新世纪图书馆,2004(6)

钟新革 东莞图书馆副研究馆员。

(收稿日期:2005-10-17 编发:刘炜 赵亮)

(上接第50页)

14 Chia-Hui Chang, Shih-Chien Kuo. Olera: semisupervised web-data extraction with visual support. Intelligent Systems, 2004,19(6):56-64

15 Labsky M, Svatek V, Svab O, et al. Types and Roles of Ontologies in Web information Extraction. ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies, Pisa, 2004

16 Ling Liu. et al. An XML-Enabled data extraction tool for Web sources. Information Systems, 2001, 26(8):585-606

17 Tejada S,. Knoblock C A, Minton S. Learning object identification rules for information integration. Information Systems, 2001, 26(8): 607-633

18 Weiguo Fan, et al. Discovering and reconciling value conflicts for numerical data integration. Information Systems, 2001, 26(8): 635-656

19 Embury S M, et al. Adapting integrity enforcement techniques for data reconciliation. Information Systems, 2001, 26(8): 657-689

20 Hanmin Jung, et al. Information extraction with automatic knowledge expansion. Information Processing & Management, 2005,41(2):217-242

21 Nahk Hyun Sung, Yong Sik Chang. Business information extraction from semi-structured webpages. Expert Systems with Applications, 2004,26(4):575-582

22 缪嘉嘉,邓 苏,刘青宝.ETL综述.计算机工程,2004,30(3):4-5,21

23 王新英,陈语林.数据抽取、转换、装载综述.企业技术开发,2004,23(8): 3-5

24 陈 弦,陈松乔.基于数据仓库的通用ETL工具的设计与实现.计算机应用研究,2004(8): 214-216

25 王新英.数据ETL问题研究.湖南工程学院学报(自然科学版),2004(3): 63-65

(收稿日期:2005-10-12 编发:刘炜 赵亮)