数字图书馆

数字图书馆中的ETL应用研究综述

黄永文 李广建

（中国科学院国家科学图书馆 北京 100080）（北京师范大学管理学院 北京 100875）

【摘要】 总结数字图书馆领域中与ETL相关的研究，在此基础上提出数字图书馆中ETL的分类，最后结合数字图书馆的应用需求和发展趋势，从ETL在数字图书馆资源建设、数字图书馆用户服务、实现数字图书馆与其他系统之间互操作3个方面，详细分析和研究数字图书馆中ETL的应用方式。

【关键词】 数字图书馆 ETL应用 信息抽取 数据清洗 【分类号】 G250.76

Review on the Application Reasearch of ETL in Digital Library

Huang Yongwen

(*National Science Library, Chinese Academy of Sciences, Beijing 100080, China*)

Li Guangjian

(*School of Management, Beijing Normal University, Beijing 100875, China*)

【**Abstract**】 The paper introduces some researches on ETL application in digital libraries, and analyzes classification and application field of ETL in resources construction, user service, resources sharing, system interoperability of digital libraries.

【**Keywords**】 Digital library ETL application Information extraction Data cleaning

1 引 言

随着用户对资源集成访问的要求越来越强烈，对不同资源进行整合和集成成为数字图书馆面临的挑战。数字图书馆在获取和融合内部和外部资源时，在追求信息资源数量的同时，在质量上必须也要进行控制，涉及到了信息的抽取、转化、清洗和加载等问题，即ETL各个环节的问题。

ETL(Extract – Transform – Load)[1]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。将ETL运用到数字图书馆中的目的是提供更加丰富的信息资源，建立信息资源保障体系，实现信息的无缝连接；实现更大范围、更有深度的资源共享；利用整合的信息资源体系为用户提供一站式信息服务，满足用户全方位、多渠道地获取信息的要求。

收稿日期：2007 – 10 – 15
收修改稿日期：2007 – 10 – 23

2 数字图书馆中与ETL有关的研究

在数字图书馆领域中，虽然没有明确提出ETL的概念，但在一些研究和应用中已经体现了ETL的理念和思想，如CiteSeer、Citebase等关于引文数据的研究，涉及到了引文的抽取、转换、查重以及存储；Sergio Bolasco等人[2]在讨论文本挖掘技术时研究了ETL在出版发行、司法政治、药物保健等领域的应用，评价了ETL在对欧美出版物的概念检索、政治新闻中相关事件的识别和检索、Biomedical文摘数据库中的作用。目前，在数字图书馆领域中与ETL有关或者涉及ETL的相关技术的研究主要体现在以下7个方面：

(1)信息抽取技术在数字图书馆中应用领域方面的研究

一些学者从理论和实践上，探讨了信息抽取技术在数字

图书馆中的应用。张智雄、刘鲁红等人[3,4]认为,信息抽取技术可以在数字内容的自动标引、元数据获取、数据挖掘、情报研究分析、大型知识库数值库建设、参考咨询等方面发挥重要的作用;刘剑兰等人[5]研究信息抽取技术在情报监测中的应用,针对国防情报应用设计了一个信息抽取系统,对各国国防经费信息进行动态的监测;Steve Jones 等人[6]介绍应用机器学习技术从纯文本中抽取信息,描述了3个应用领域:等级短语浏览;使用可调整的压缩技术进行文本挖掘;关键短语抽取以及在数字图书馆中的应用。

(2)从数字图书馆中文献类型的角度研究元数据抽取问题

针对数字图书馆中存在的不同类型文献资源,如期刊论文、学位论文、图书等类型进行元数据抽取研究。Ben Wellner 等人[7]以学术型论文为研究对象,提出利用条件随机域(Conditional Random Fields,CRF)从学术论文的头部和参考文献部分中抽取元数据;李朝光等人[8]以学位论文为研究对象,利用学位论文所特有的结构进行元数据抽取研究;对于图书类型的资源,李向阳等人[9]提出一种基于竞争分类的网上图书信息抽取方法,通过信息片段对信息模板槽的竞争来实现信息片段的分类和噪声信息的过滤;国防技术信息中心(Defense Technical Information Center,DTIC)数字图书馆主要研究从技术报告中抽取报告中的元数据和文档的结构信息。

(3)从不同信息来源研究抽取机制、框架及抽取规则

针对数字图书馆中信息资源的不同来源,如文献数据库、网上论坛、Web网页等进行研究。胡金化等人[10]研究了文献数据库中文本的匹配模式抽取,并提出了一种面向文本数据库的信息查询机制;奚伟鹏等人[11]针对网上论坛,提出一套面向网上论坛的语义话题线索抽取框架;针对Web网页进行的研究比较多,如冯伟华、郭志红、王亮等人[12-14]分别研究了利用规则模式或者Wrapper归纳技术抽取半结构化网页中的信息以及基于扩展标记图模型的Web信息抽取器,张丙奇等人[15]专门对企业类网页进行抽取以获取企业竞争情报。

(4)针对资源的不同格式进行抽取方法的研究

针对数字图书馆中文献资源存在的格式,如HTML、DOC、PS、PDF等进行研究,目前的研究主要集中在对HTML和PDF格式文献的研究。对于基于HTML格式的Web文档的信息抽取,有的学者提出了基于样本实例的Web信息抽取方法、基于DOM的Web信息抽取方法等;Donna Bergmark 等人[16]研究了从PDF文献中获取引文信息,先将PDF文件转换成HTML格式,再根据HTML标识来识别和处理引文信息;Wende Zhang 等人[17]利用PDF的物理结构和逻辑结构的特征抽取文档元数据。

(5)从语义Web和Ontology角度研究语义层次的抽取

郭瑞华等人[18]研究了语义Web上元数据的描述和抽取技术,指出经XML和RDF/XML描述后的DC元数据具有语义标注,可实现语义网上数据的自动抽取功能。国内外还出现了一些利用Ontology进行信息抽取的相关研究,如KIM项目中的OBIE、hTechSight项目等。刘金红等人[19]提出先对网络资源进行元数据标注,然后基于Ontology对标注过的网络资源进行数据抽取;陆科进等人[20]提出对需要抽取的领域抽出关键词和信息片,将其组织成数据库的元组属性值;廖乐健等人[21]从知识表示与推理的角度研究了提高信息抽取智能性的途径,提出将Ontology与模板规则相结合。

(6)数据清洗方面的研究

数字图书馆在信息资源建设中,很容易发生输入错误、加工重复记录、滥用缩写词和惯用语以及存在缺损值等问题,容易造成脏数据。这些脏数据和信息可能会带来检索不便、决策制定失败甚至错误等。因此,需要将从不同信息来源抽取出来的信息进行清洗、转换。Sunita Sarawag 等人[22]研究了从物理出版物的仓储中(LATEX格式)创建引文,介绍了对抽取后的引文数据的清洗、匹配和加权过程;还有一些研究主要是在信息抽取前对页面进行过滤、清洗,即去掉与抽取无关的信息区域。

(7)重复记录检测方面的研究

由于同一篇文献可能被多个数据来源收录,或者对同一篇文献以不同的方式进行标引,都可能导致重复记录的出现。目前,已经出现了一些关于重复记录的研究,如Ayres[23]提出了基于子字段字符串进行重复记录的检测;有些学者[24]研究了在联盟型数字图书馆(FDL)环境下重复记录的检测算法,先将所有的记录按权值排序,然后比较记录的所有元数据属性,从而确认记录间是否重复,并从Arc中选取一定的记录集,在其上进行了实验。

3 数字图书馆中的ETL分类研究

目前,还没有关于数字图书馆中ETL分类的相关研究。本文为了系统全面地分析和梳理ETL,按照5种不同标准对数字图书馆领域中的ETL进行分类(见图1),以便于从不同的角度分析不同类型ETL所具有的特点,更好地认识ETL在数字图书馆中的应用。

(1)从数字图书馆的业务流程来对ETL进行分类,可以更好地分析ETL在数字图书馆不同业务环节中的作用、侧重点及特点,具体分为:资源采集和获取过程的ETL、资源加工过程的ETL、资源存储过程的ETL和服务过程中的ETL;

(2)从处理的信息来源对ETL进行分类,可以分析针对不同信息源ETL所采取的不同实现技术,具体分为:数据库ETL和Web ETL;

(3)从处理内容的角度对ETL进行分类,可以分析对于不同的信息内容ETL所采取的不同过程和步骤,具体分为:元数据ETL和引文数据ETL;

(4)从运行的时机对ETL进行分类,可以分析在不同的

环境和应用场景下 ETL 处理对时间的要求,具体分为:离线 ETL 和实时 ETL;

(5)从处理的方式对 ETL 进行分类,可以分析在不同的情况下 ETL 所采取的不同实现方法和手段,具体分为:集中式 ETL、拆分式 ETL 和重构式 ETL。

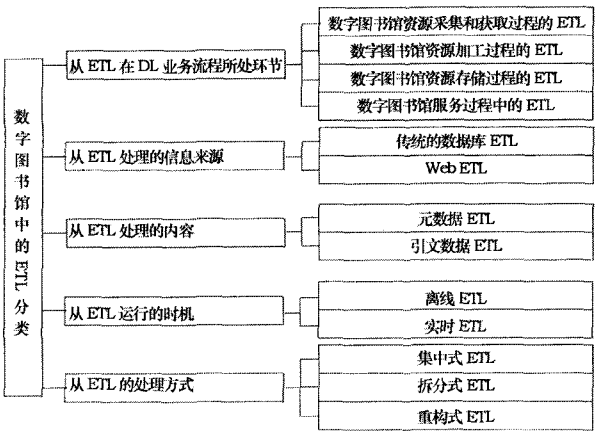

图 1 数字图书馆中的 ETL 分类

4 数字图书馆中的 ETL 应用方式

通过上述分析,可以看出 ETL 在数字图书馆中发挥着非常重要的作用,具体来说,ETL 主要可以应用在数字图书馆的资源建设、数字图书馆用户服务以及实现数字图书馆与其他开放系统之间互操作等过程中。

4.1 在数字图书馆资源建设中的应用

(1)ETL 在基于数据仓库的数字资源物理集成中的应用

基于数据仓库的物理集成主要是通过统一的数据模式对分布式、异构的内外部资源进行整合、重组与集成,消除异构性所带来的资源利用上的困难。在对异构信息资源的物理整合中,ETL 主要是从分布式异构的信息源中收集数据,映射成统一的结构,进行必要的转换和清洗,并将其存储在数据仓库中。采用建立本地数据仓库的方法进行整合,可以对整合的结果作更进一步的处理和分析,改善数字图书馆系统的服务效果。由于采用的是基于数据仓库的数据集成方式,整合后的资源相对稳定,既可供用户进行快速的统一检索,也可作为信息挖掘的资源基础,大大提高了数字资源的利用率。

(2)ETL 在信息处理和数据交换中的应用

在信息处理和交换过程中,目前主要的解决方案有:半手工方式和分布式数据库复制技术。随着需求的多变,半手工方式越来越呈现出弊端,如由于人工参与的成分太多,往往会造成工作的延误。分布式数据库复制技术主要通过为不同站点的数据引入适当的冗余,从技术上来讲,在解决数据抽取和导入问题时不适合采取数据库复制技术。采用 ETL 来进行信息处理和数据交换,可以克服半手工方式和分布式数据库复制技术存在的问题。将 ETL 统一部署或者分布式部署,由 ETL 来完成定期从各个数据处理或者交换单位进行数据抽取、转化、清洗等,尽量避免人工的操作,提高资源处理和交换的效率。

(3)ETL 在元数据获取中的应用

内容标引和元数据加工是数字图书馆区别于其他信息检索系统的一个重要方面,但随着信息资源的迅速增长,手工的元数据加工已远不能适应这一需要。探索有效的内容标引和元数据抽取已成为数字图书馆资源加工处理和数字图书馆可持续发展的瓶颈。实际上,国外的一些研究人员已经注意到了信息抽取在数字图书馆元数据建设方面的作用,出现了将信息抽取应用于数字图书馆的内容标引和元数据抽取的相关研究,也研制和开发了一些元数据抽取系统,如 CARA 系统、ViBSoz、CARMEN、MetaExtract 系统等。其中,MetaExtract 系统是 Syracuse 大学的自然语言处理中心(CNLP)利用自然语言处理为解决教育元数据获取问题而研制的。

(4)ETL 在引文数据建设中的应用

利用 ETL 可以实现在分布式数字图书馆环境下自动抽取引文数据,实现引文数据的识别以及引文数据之间的相互链接。例如,CiteSeer 从 Web 上下载文献,抽取引文和文章中引文的上下文信息,将这些信息存储在数据库中,实现了 Web 上文献之间的引文链接,提供“被引文献”、“引用的文献”、“同引的相关文献”等服务。抽取后的引文数据还可以扩展个性化推送服务的信息内容,主要是依据同引或者同被引文献之间具有相同或者相似的研究领域或者主题。例如,用户浏览了文献 A,如果以后出现了引用文献 A 的文章就进行推送。SCI 根据这个原理,提供个性化推送服务,但 SCI 的引文数据是由手工完成的。

4.2 在数字图书馆用户服务中的应用

(1)统一检索服务

在统一检索过程中,ETL 主要是受用户需求驱动的,对从各个数据源返回的检索结果信息进行转换、查重和集成,一般并不需要将检索结果装载在实际的数据(仓)库中。在检索结果返回过程中进行实时的 ETL 处理,对返回的结果进行抽取、去重、过滤、整合和加载到临时性存储区域,提供给用户整合后的结果。由于直接对用户服务,对服务的时间和效率要求比较高。因此,ETL 的处理不应该过于复杂。可以保留用户的检索结果,系统定时地根据用户检索历史启动 ETL,作更深层次的抽取和处理,为以后相同的用户检索或者用户推荐提供更好的服务。

(2)参考咨询中的问题解答服务

目前的参考咨询系统主要凭借馆员个人的学识解答读者的问题,而网络上丰富的资源为问答系统提供了另外一种良好的知识来源,对于回答简短、基于事实的问题非常有效。可以将 ETL 技术应用到参考咨询的问答系统中,在本地的参考咨询系统中嵌入 ETL 引擎,实现多途径的基于知识库的自动

问答。将表达问题的关键词转换为检索表达式，发送给搜索引擎并返回相关结果，系统直接利用网络搜索结果中的全文或者摘要内容进行答案的抽取，将答案提交给用户，并同时上载到本地的知识库，不断地扩展和丰富用于自动问答的知识库。以网络作为其知识库的自动问答系统主要有华盛顿大学的 MULDERL 系统、新加坡国立大学的 LAMP 系统等。

(3)个性化服务中的 ETL

ETL 可以应用到个性化服务中，对单个用户或用户群的需求提供有针对性的服务。对于个性化服务中的 ETL，主要需要处理两个问题：一是用户个人兴趣的获取，二是文献知识元的获取。对于用户兴趣的获取，主要通过用户的基本信息、检索和浏览行为来获得。关于用户的基本信息可以在注册时要求用户提供，而对于检索和浏览行为，则需要系统进行自动监控和记录，或者通过系统的日志进行分析和抽取，ETL 的重点在于后者，主要是从用户历史记录中抽取相关的信息；对于题名、作者等基本元数据，在对文献标引时都会生成，不过对于文献中的图、表、公式、引文、段落主题、章节主题等知识元，则大多不会标引，而这些是 ETL 处理引擎为用户提供个性化、知识化服务的重要基础，因此，需要对这些内容和信息进行抽取和处理。

(4)面向情报分析服务的 ETL

大量的相关数据和信息是进行情报研究的基础，ETL 技术提供了进行大规模数据及信息采集的思路和机制。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。ETL 可以应用到情报研究中，特别是在竞争情报中的应用。ETL 技术可以从文本中有效地抽取和表示信息，将分散在网络中的数据、数值、信息和知识进行集成，建立起可供研究分析的资源基础，使得分析人员可以更加全面深入地理解数据，支持未知事实的发现，提高分析趋势和发现潜在信息的能力。

4.3 在实现数字图书馆与其他系统之间互操作中的应用

开放性是数字图书馆一个重要的发展趋势，逐渐出现了一些开放式数字图书馆(Open Digital Library，ODL)，其主要目标是实现数字图书馆资源和服务的使用最大化，不仅为用户提供资源和服务，也可以为其他开放性的服务系统提供资源，如 VIS(可视化系统)、e - Learning 等。数字图书馆中具有丰富的信息资源，可以为这些系统提供高质量的资源。通过 ETL 从数字图书馆中抽取元数据、数字对象等，并经过封装后发送到具体的应用系统中，实现与这些系统和工具的互操作。

(1)实现数字图书馆与 VIS 之间的互操作

目前，大多数数字图书馆都只为用户提供检索和浏览服务，没有提供检索结果或者信息浏览的可视化服务。因为信息可视化具有直观、形象等优点，越来越多的研究关注于信息的可视化在不同领域中的应用，包括在数字图书馆领域中的应用，将数字图书馆中的检索服务和 VIS 系统中的可视化服务相互结合起来，为用户提供更好更直观的服务。

在 ENVISION MARIAN 项目[25]中提出了 VIDI 协议，尝试着将 VIS 和 DL 之间的相互结合。在 VIDI 的互操作框架中，包括 DL 提供者和可视化服务提供者。DL 提供者管理支持 VIDI 协议的各个 DL，并利用 VIDI 揭示各个 DL 中资源的元数据。可视化服务提供者通过 VIDI 向 DL 提供者发出请求，将返回的数据作为可视化的基础，以图形化的方式将结果展示给用户。支持 VIDI 协议的 VIS 可以看作是可视化服务提供者。不过，VIDI 协议只是定义了 DL 和 VIS 系统需要遵循的元数据格式，因为一些历史原因，在大多数情况下，DL 本身并不能直接提供符合标准的元数据，这时还需要对数据进行处理，可以利用 ETL 将 DL 的元数据转换成 VIS 可以接受的格式，需要进行元数据的映射和转化以及补充缺少的内容，再实时地传输到 VIS 系统中。

(2)实现数字图书馆与e - Learning系统之间的互操作

实现数字图书馆与 e - Learning 系统之间的整合方法有很多，如组织方法、系统方法、管理方法、基于内容的方法、以用户为中心的方法、过程方法和标准方法，在实际中通常采用系统整合、内容整合和用户整合策略来构造数字图书馆和 e - Learning系统的整合框架，可以通过协议或者在资源层实现 DL 与 e - Learning 系统之间的互操作。在关于教育管理系统(Instructional Management Systmen，IMS)和开放知识计划(Open Knowledge Initiative，OKI)之间关系的报告中[26]，研究了数字图书馆和 e - Learning 系统之间的数字仓储层次的整合。

利用 ETL 可以实现数字图书馆中的资源为 e - Learning 系统所使用的过程，在这个过程中，需要解决的核心问题是元数据的转换和封装以及用户信息的整合。在数字图书馆中一般采用 DC、MARC 等元数据标准，LCMS 通常采用 SCORM、LOM 等标准，需要实现两种元数据之间的映射，并增加 SCORM、LOM 等标准中必要的元数据。同时，为了整合两个系统中的每个用户数据，需要抽取使用者的属性和用户活动的模式，获得用户模型的知识表示，实现数字图书馆与 e - Learning系统之间更深层次的互动。

5 结 语

目前，在数字图书馆领域中已经出现了与 ETL 技术有关的探索和研究，运用抽取、清洗等技术来解决数字图书馆资源建设、资源共享和服务集成等过程中出现的问题。但总的来说，这些研究还是面向任务或者问题的，目前还没有从整体角度上对 ETL 进行讨论和探索的研究。

随着数字图书馆的不断发展,对数字资源的融合、对网络资源的自动获取以及服务集成等需求会越来越强烈。因此,有必要从整体出发对 ETL 在数字图书馆中的应用作全面的探索。

参考文献:

[1] Simitsis A, Vassiliadis P, Sellis T. Optimizing ETL Processes in Data Warehouses[C]. 21st International Conference on Data Engineering (ICDE'05), 2005 :564 - 575.

[2] Bolasco S, Canzonetti A, Federico M C, et al. Understanding Text Mining: A Pragmatic Approach[C]. In: Proceedings of the NEMIS 2004 Final Conference, 2005: 31 - 50.

[3] 张智雄. 信息抽取技术及其在数字图书馆中的应用前景分析[J]. 现代图书情报技术, 2004(6): 1 - 5, 23.

[4] 刘鲁红, 刘力强, 胡亚军. 信息抽取技术在数字图书馆中的应用研究[J]. 情报理论与实践, 2005, 28(3): 321 - 324.

[5] 刘剑兰, 朱东华. 信息抽取技术在情报监测中的应用[J]. 情报学报, 2004, 23(6): 661 - 666.

[6] Jones S, Paynter G W. Automatic Extraction of Document Keyphrases for Use in Digital Iibraries: Evaluation and Applications [J]. Journal of the American Society for Information Science and Technology, 2002, 53(2): 653 - 677.

[7] Wellner B, McCallum A, Peng F C, et al. An Integrated, Conditional Model of Information Extraction and Coreference with Application to Citation Matching[C/OL]. [2007 - 05 - 20]. Conference on Uncertainty in Artificial Intelligence (UAI), 2004. http://www.cs.umass.edu/~mccallum/papers/integrated04uai.pdf.

[8] 李朝光, 张铭, 邓志鸿, 等. 论文元数据信息的自动抽取[J]. 计算机工程与应用, 2002, 38(21): 189 - 191, 235.

[9] 李向阳, 张亚非. 一种网上图书信息抽取方法[J]. 情报学报, 2004, 23(6): 655 - 660.

[10] 胡金化, 胡运发, 周益群, 等. 面向中文文本数据库的信息抽取机制[J]. 小型微型计算机系统, 2002, 23(10): 1161 - 1164.

[11] 奚伟鹏, 李昕, 蒋饥, 等. 面向网上论坛的信息抽取技术[J]. 计算机工程, 2005, 31(4): 66 - 68.

[12] 冯伟华, 苗长芬. 基于 Web 的网页信息抽取方法的研究[J]. 洛阳工业高等专科学校学报, 2005, 15(3): 30 - 31.

[13] 郭志红. 基于 Web 资源的信息抽取技术[J]. 情报科学, 2002, 20(12): 1282 - 1284.

[14] 王亮, 朱征宇. 基于扩展标记图的 Web 信息抽取器[J]. 计算机工程, 2005, 31(8): 159 - 161, 191.

[15] 张丙奇, 姜吉发. 企业相关信息抽取技术研究与系统实现[J]. 微电子学与计算机, 2004, 21(1): 1 - 6.

[16] Bergmark D, Phempoonpanich P, Zhao S M. Scraping the ACM Digital Iibrary[J]. ACM SIGIR Forum, 2001, 35(2): 1 - 7.

[17] Zhang W D, Song Y J. Research on PDF Documents Information Extraction System Based - on XML[EB/OL]. [2007 - 05 - 20]. http://adt.caul.edu.au/etd2005/papers/057Zhang.pdf.

[18] 郭瑞华, 张玉莉. 语义 Web 上 DC 元数据的描述及抽取技术[J]. 现代情报, 2005, 25(6): 212 - 214.

[19] 刘金红, 夏阳, 陆余良. 基于 Ontology 的网络元数据抽取系统的研究与实现[J]. 安徽电子信息职业技术学院学报, 2004, 3(5): 10 - 13.

[20] 陆科进, 李新颖. 基于 Ontology 的文本信息抽取[J]. 计算机应用研究, 2003, 20(7): 46 - 48.

[21] 廖乐健, 曹元大, 李新颖. 基于 Ontology 的信息抽取[J]. 计算机工程与应用, 2002, 38(23): 110 - 113.

[22] Sarawagi S, Srinivasan S, Vydiswaran V G, et al. Resolving Citations in a Paper Repository[J]. ACM SIGKDD Explorations Newsletter, 2003, 5(2): 156 - 157.

[23] Ayres F H, Huggill J W, Yannakoudakis E J. The Universal Standard Bibliographic Code (USBC): Its Use for Clearing, Merging and Controlling Large Databases[J]. Program, 1998, 22(2): 117 - 132.

[24] Haseebulla M K, Kurt M, Mohammad Z. Similarity and Duplicate Detection System for an OAI Compliant Federated Digital Library [C]. The 9th European Conference on Research and Advanced Technology for Digital Libraries, 2005: 531 - 532.

[25] Shen R, Wang J, Edward A F. A Lightweight Protocol Between Digital Libraries and Visualization Systems[EB/OL]. [2007 - 05 - 25]. http://vw.indiana.edu/visual02/Shen.pdf.

[26] Griffin S, Merriman J. E - learning and the Digital Library——A Report on Collaboration Between IMS and OKI[EB/OL]. [2007 - 05 - 25]. CNI Fall Task Force Meeting, 2002. http://www.cni.org/tfms/2002b.fall/PowerPoint/PPT - E - Learning.ppt.

(作者 E - mail: hyongwen@mail.las.ac.cn)