●王兰成

主题信息检索应用数据库技术的研究现状与展望*

摘 要 文献领域主题信息检索应用数据库技术的系统研究是数字图书馆建设的关键之一。当前,主题信息检索应用数据库技术的研究可在两个方面展开:XML 的关系数据库应用研究,XML MARC 的信息描述和主题检索。参考文献 30。

关键词 数字图书馆 信息检索 数据库技术 主题概念 可扩展标记语言
分类号 G350

ABSTRAET The author thinks that the researches on subject information retrieval application databases are the key for the development of digital library. At present, the researches on the technologies of subject information retrieval application databases are reflected in two aspects, i.e. the researches on the applications of XML relational databases and the information description and subject retrieval of XML MARC. 30 refs.

KEY WORDS Digital library. Information retrieval. Databese technology. Subject concept. XML.
CLASS NUMBER G350

1 数据库新技术与数字图书馆

从狭义理解数字图书馆,它是在传统图书馆职能基础上的扩充与发展,不但具有传统图书馆的各种功能,而且增加了服务对象广域化、资源形式多元化、服务形式虚拟化等新特点和新功能。数据库技术是研究数据的科学组织与存储、高效检索与处理,因此它是数字图书馆的基础技术。

当前数据库新技术的主要应用有:

(1)采用 WWW、Java 和 Active X 的 Web 数据库技术,建立数字图书馆的前台 Web 信息发布系统平台,并通过 CGI、API 接口和脚本语言等技术实现与书目库、全文库等馆藏信息资源数据的交互处理。

(2)利用 CORBA 技术和 Z39.50 协议的联邦数据库技术实现网上的多个异构数字图书馆集成,不仅解决资源共享问题,而且有利于实现资源的合理调配和各联邦馆的资金有效利用。

(3)增加并行处理控制软件或采用高效并行算法的并行数据库技术,以提高数字图书馆系统(特别是处理全文库)整体服务的质量。

(4)面向对象数据库技术是处理多媒体数据的一种实效性强的方法,能够很好管理数字图书馆中的多元化数据源。

(5)包括关联规则和分类、聚类分析方法的数据挖掘(又称为数据库中的知识发现,KDD),应用于图书馆不仅对管理决策方面起调控作用,而且资源也能够得到更充分利用。

(6)参考咨询自动化系统是数字图书馆的重要组成部分,包含演绎和模糊等技术的逻辑数据库的运用,可以大大提高信息的检索功能。

(7)数据网格(Grid)技术的出现,能够实现各种计算资源和数据资源的统一访问,使数字图书馆中海量数据的处理更加有效。

国内外数字图书馆技术与系统的发展,其实际需要及建设中的关键技术重新提出了对数字化过程中的标引与检索、信息抽取中的语义知识和元数据操作、网络环境下信息资源的开发与检索等方面进行深入研究和应用试验。信息管理和数据处理是计算机应用的一个主要领域,而数据库技术始终是这些领域的核心。XML(eXtensible Markup Language)作为一种新的网上数据交换标准正日益流行,对 XML 存储管理技术的探究也因之而成为一大热点,其中利用现有的发展成熟的关系数据库技术来管理 XML 数据更是人们关注的焦点。因此,XML 数据库技术应用与概念信息检索研究是当前图书馆学数据库与信息处理方向的一个重要领域,本文将在 XML MARC 数据库和中文主题概念检索方面,作较为深入的综述和研究。

2 主题信息检索与数据库技术的应用

国际上元数据发展是不平衡的。我国图书馆界在 1996 年制定了部颁标准《中国机读目录格式》,情报界在 1989 年制定了适合于情报信息交换的机读格式标准《中国公共交换格式》,档案界至今还处于实验阶段。国际上对图书、情报和档案学科的集中元数据的理论与实践均未深入研究。

* 本文系南京政治学院上海分院拔尖人才专项课题研究成果之一。

机读目录 MARC 在数据描述、数据交换等方面有其优越性,但也存在数据可读性差,数据的管理、显示和检索依赖于特定的软件平台,不方便直接进行网上发布等局限性,制约了其进一步应用和发展。对元数据描述采用 DTD(Document Type Description,文档类型描述)方案多,缺少 XML Schema 实现的具体研究。我国对数字图书馆、数字档案馆等领域的网上集中信息描述及其实现研究还没有深入或是尚存在空白,迫切需要真正发挥集成、共享的目录信息的文化与知识属性。

信息技术领域的一个热点问题是如何帮助用户高质量地检索获取真正有用的信息。国际上对中文信息的概念检索非常关注,在主题概念检索方面尚有许多研究或应用的空白。首先,主题标引是主题检索的前提,自动分词是中文自动标引的基础课题,加上这一问题本身的难度,很难从根本上解决,所以在很长时间内受到人们的关注。其次,自然语言检索法的特点是检索标识使用不加规范的自然语言,检索标识或者从文献的题名、文摘、正文中自动抽出,或者由标引人员自由标引,标引过程被不同程度地弱化,标引系统的质量乃至检索效果就主要依赖于控制词表的规范控制能力。第三,随着网络技术与智能技术不断创新计算机应用环境,网上关键词不能按需要进行扩检、缩检或作相关检索而影响检索效果,目录型网络信息检索分类处理跟不上信息扩张的速度且类目很少被合理和统一地组织,即使在自由词标引环境下,主题语义关系也通常被作为一种较为有效的控制手段以提高检索系统的性能。基于主题概念知识进一步减少分词的歧义性和缩短标引抽词的时间,通过文本主题分析实现动态分类和词句概念检索以提高信息检索的质量,成为当前开拓研究和能产生实效的新研究课题。

当今信息技术发展迅猛,图书馆界大多数关键数据都是放置于数据库中进行管理的,一来目前数据库技术已经相当成熟,二来其管理功能非常强大。以往的数据库应用,其数据低层结构一般来说都是相对固定的,即应用程序开放性较差。而 XML 作为一种可扩展性标记语言,其自描述性使它非常适用于不同应用间的数据交换,而且这种交换没有预先规定数据结构,因此具备很强的开放性,具有广阔的应用前景。为了使基于 XML 的机读目录数据交换成为可能,就必须实现数据库的 XML 机读目录数据存取,并且将 XML 数据同应用程序集成,进而使之同现有的利用规则相结合。纵观现有的对 XML 和关系数据库的集成技术,主要沿着两个方向进行:一是对现有的 RDBMS 技术进行扩展使之提供对 XML 数据的管理支持;一是将 XML 应用构架于关系数据库之上,直接利用底层的 RDBMS 技术实现对 XML 数据的存储管理。前者主要体现在商业领域,几大主要的关系数据库运营商都提供了专门的商业工具包以支持 XML 管理,如 DB2 的 XML Extender, Oracle 的 XSQL 等。但是,这些研究成果在图书馆界还没有得到很好应用,必须尽快抓住有利条件,开展应用研究。

3 XML MARC 数据库的研究现状

MARC 作为一种元数据格式,在信息描述的详尽性、信息存储的方便性、信息交换的结构化、标准化和信息检索的检准率方面具有不可比拟的优势。经过 30 多年的发展,MARC 在国内外图书馆界得到了广泛应用,它本身也得到进一步完善和发展,目前仍有继续存在和发展的必要。但由于 MARC 自身的局限性,满足不了网络海量信息资源的整序需求。XML 技术与 MARC 标准相结合,形成了 XML MARC 数据库,恰好能够解决这些问题。

王晔等提出了一种新的基于元数据的检索方式,通过 Dublin Core 元数据集到 MARC 的相互转化,保持了与通用的 Z39.50 客户端的兼容性[1]。有文献介绍了加拿大不列颠哥伦比亚省档案联合目录系统,开发建立了基于 Linux 平台的 Web 数据库系统,采用 ISO 2709 的信息交换格式实现了档案 MARC 的 Web 发布,具有目录多级检索功能[2]。瑞典国家档案信息系统 ARKIS II 型,其数据不仅是文件级著录,而且还可以是案卷级、分类别级、类别级、分全宗和全宗级著录,符合 ISAD(G)著录标准[3]。徐周、黄上腾介绍了基于 XML 实现数据库间信息交换的方法[4]。国外有学者通过论述 MARC、Dublin Core、档案 MARC、EAD 的制定与结构等方面的特点和比较,认为元数据是解决网络资源从无序走向有序的方案之一,从著录格式、著录对象、著录主体、著录详简程度、应用范围、产生途径及字段转换等方面论述了元数据中的两种常用格式 Dublin Core 和 MARC 的区别和联系[5~7]。黄伟红、张福炎研究基于 XML/RDF 的 MARC 元数据描述技术,从而使得专用的 MARC 规范格式的书目数据转换成通用的机器可读和机器可理解的元数据成为可能[8]。苏新宁通过对音像资料的特点及其与 CNMARC 的分析比较,设计了音像资料的 MARC 格式[9]。作者所在单位成功开发了《中国档案机读目录计算机处理系统》(上海市科技发展基金项目 00JG05043)和《中国档案置标著录 WEB 检索系统》(国家档案局科技项目 2001-X-16),系统借鉴了国际先进的档案信息管理理念和规范,对基于互联网环境的档案信息共享提供了标准化的应用模式,选用 Microsoft. NET 系统构建和 ASP.NET、XML Schema 等先进技术作为解决方案,采用本地计算机处理系统和基于 Web 的远程 MARC 记录查询子系统的结合作为系统开发方案,以适合单机、局域网、因特网等各个层次用户需求,从而形成最为广泛的用户群。

以上成果或文献表明,国外已有档案 MARC 的研究及其档案信息 Web 发布的系统,国内仅有初步研究成果;国内外均未发现有关 MARC 集中信息描述成果或方法的介绍;国内缺少基于 XML Schema 的应用研究,也没有发现基于 XML 的 MARC 核心元数据实现方案及其比较研究。总之,图书馆界缺乏深入运用 XML 数据库的先进成果。

4 主题分类概念检索的研究现状

作为一门学科的信息检索,从 Granfield 确立标引语言到 Smeaton 等人关于计算机语言学上检索技术开发等等,已对传统信息检索领域产生了重要影响。国内外目前研制的

中文信息检索方法和网上搜索引擎，无论是关键字符的机械式匹配，还是结合布尔逻辑运算提供更为复杂的查询表达方式，都是以关键词匹配为基础的。此方法有两个缺陷：一是检索结果只是在字面上符合用户的要求，实际内容往往偏离用户的实际需要；二是用户输入的查询词稍有偏差，检索系统就无法确定用户的真正需要，因而无法提交正确结果。缺乏主题知识环境而期望检索效果和质量要有大的提高已很困难，用增强关键词检索功能的各种措施并不能消除其本质缺陷。目前已有进行网络信息自动分类方面的研究，但没有发现既基于主题知识，又能脱离种类多、更新慢、编制工作量极大的分类主题词表的相关研究，以及实现的动态分类概念检索。

情报学家 K. Sparck、G. Salton、R. M. Needham、M. E. Lesk、K. S. Jones 等众多学者在这一领域进行了卓有成效的研究工作。有人介绍了信息检索领域在索引模型、文档内容表示、匹配策略等方面取得的研究成果[10]。Pirolli 等针对从文档中抽取关键信息，提出使用中心文档代表文档集合，使用中心词汇表示文档及求取中心文档和中心词汇的算法[11]。国外还有学者指出传统的信息检索系统主要基于布尔模型、向量空间模型和概率模型，认为非实质性语义检索的基于字、词匹配传统检索有局限性[12]。Loh 提出一种新的信息检索代数模型，即隐含语义分析，提出基于概念描述种种属性，概念与子概念密切联系，从而形成概念层次[13]。为了有效地在结构文档中进行查询，Shin 的研究提供了几种不同的索引结构；基于传统的概率模型提出的结构检索，都是基于词的检索[14]。Wolff 和 Jennifer Chu 等人提出，应建立文本中每个字词的索引及其关系，实现基于内容的查询；国际上推出的 TRIP、ZyLAB、Ariadne、Envision 等检索系统，都具有一定的面向内容检索的文献分析能力[15~16]。萧璐等认为在文本和检索请求的表达形式上更接近自然语言，即全文检索必须奠基在语言的最基本构件上，他们对自动标引中的向量空间模型和实现作了深入介绍[17~18]。有学者研究现代扩展布尔检索和检索策略，论述了信息检索的本体论、数字图书馆、虚拟图书馆及多媒体系统，描述了基于 Z39.50 的联机书目检索服务系统[19~22]。有的文献介绍了信息组织和检索的方法与知识库系统，介绍了基于向量空间模型的信息检索系统的设计与实现，研究了文本信息的自动分类技术[23~24]。李蕾等认为国内外目前的搜索引擎几乎没有一个能达到概念信息检索的要求，并研究概念检索接口的问题[25]。朱毅华等学者提出了字面和词汇相似匹配的方法，并在计算中考虑匹配字数、词汇结构和语义方面的因素[26]。

以上成果或文献表明，国内外缺少对领域主题分类与主题概念的一致检索的研究，关于中文语义词素相似度识别仅用于自动标引，而未见到用于提高检索质量方面的报道，也没有发现通过主题范畴模式抽取以实现文献的动态分类和词句的概念检索的解决方法。所以研究、使用主题知识并实现自适应分类的概念检索，是网上信息自动检索软件发展的方向之一。

5 深入或创新性的研究工作及展望

当前主题信息检索应用数据库技术的研究可在两个方面开展。

5.1 XML 的关系数据库应用研究

随着 XML 数据日益成为数据交换的标准，对 XML 的高性能数据管理要求越来越迫切。目前，管理 XML 数据的方式主要有：传统的关系（或者对象）数据库、中间件和 Native XML 数据库。使用关系数据库的方法虽然利用了现有的成熟技术，但是由于 XML 数据和传统的关系数据有着本质上的区别，比如结点的有序性、半结构化性等，其每一步都要做大量的映射转换工作，如 XML 数据转成关系存储、XQuery 转成 SQL 查询、关系查询结果还要转换成 XML 输出，等等。因此，我们要在效率上和准确性上做更多的应用研究。

(1)多 DTD 环境中 XML 的关系数据库存储。目前基于多 DTD 环境的利用 RDBMS 技术存储 XML 数据的方法[27]，采用独特的编码体系对 DTD 和 XML 信息分别进行编码，然后将编码后的节点信息存入数据库对应的表中。因为 DTD 的信息存放在一个单独的表中，所以能比较容易地对多 DTD 以及对应的 XML 文档进行管理；而将 DTD 和 XML 数据分开存放，也给查询带来了便利，利用 DTD 快速定位查询路径中的条件节点、目标节点，避免与 XML 数据相关的多表连接带来的问题。

(2)基于 XML 技术的 Web 信息提取和集成。相对 DTD，XML Schema 具有更强大的描述能力，主要体现在一致性、扩展性、互换性和规范性上，有更好的应用前景。将半结构化的 HTML 网页信息转换为具有 Schema 的 XML 信息，并通过 XML Schema 与关系数据的映射关系，将 XML 文档保存到数据库表中，从而实现半结构化 Web 信息到结构化关系断据的转换和集成。刘世杰等提出了一种新的 Web 信息提取和集成算法[28]。

(3)XML 搜索引擎的查询扩展。研究以简单的查询方式帮助用户进行 XML 查询，扩大搜索引擎的搜索范围以提高查全率(Recall)和查准率(Precision)。有学者提出结构库在 XML 查询和 XML 搜索中的应用主要体现查询扩展、近似搜索和查询建议，研究工作主要包括构造不同粒度的结构库，把结构库应用到 XML 搜索引擎的查询扩展方面，评估它是否能提高查全率和查准率[29]。

(4)XML 文档集公共模式的获取。XML 文档集公共模式获取的一种技术是从元素的重复出现、元素组、选择出现三个方面描述元素序列的规律，通过消除冗余和合并表达式限制中间结果的数量。温俊等人讨论了一种基于规则获取 XML 文档集公共模式的策略，保留了表达能力强的表达式，保证系统在处理规模较大的 XML 文档集时仍可工作[30]。

5.2 XML MARC 的信息描述和主题检索

(1)基于 XML 的 MARC 信息集成描述和实现。对 MARC 的优点及其局限性进行理论分析，研究档案领域的基于 XML MARC 元数据设计和应用研究，研究基于 XML

的MARC集中的信息描述机制，并进行相关元数据之间的比较和分析。实现的研究包括XML MARC的文档类型定义和模式定义设计，基于XML MARC书目内容描述和资源描述框架的Schema设计，基于XML MARC的核心元素集及其XML Schema的实现模式。

(2) 构建基于主题范畴的概念知识库方法。研究并构建合理的主题概念知识网络，自动构建多级主题范畴索引，依据词义标注自动构建主题词族索引，依据主题词表内词素及其关联的标注将信息主题标引和概念检索构成一个有机的整体知识环境，并探讨MARC与XML MARC数据库中信息映射和数据转换的实现。

(3)基于概念知识关系的XML MARC主题分类概念检索。从提高自动标引质量和速度方面，改进有关的自动标引算法和方法，实现XML MARC数据库特征数据的自动提取；研究新的主题分类概念检索，它既基于主题知识，又能脱离种类多、更新慢、编制工作量极大的分类主题词表，以实现主题范畴索引下的自适应自动分类，实现主题知识标引下自动分类概念检索，评测其检索效果和质量。

作者目前的研究工作进展是：首先，提出了基于COM的跨库检索代理模型，并研究使用Web Service技术解决异构数据集成的问题。其次，研究并初步建立了基于XML的XMARC信息集中描述机制，设计了基于领域内容和框架的两种XMARC数据库方案，定义了集中XMARC的核心元素集及其XML Schema的实现模式。第三，初步实现了MARC向XMARC信息的无损映射。第四，构建了K-S-C(Keyword-Subject-Category)主题概念的语义关系，进而运用于XMARC文本的自动标引。第五，提出使用切分位置信息、最小回溯匹配和主题概念词素等新的自动标引方法，改进了现有的MM(Maximum Matching，最大匹配)自动标引算法。第六，研究新的XMARC主题分类与主题概念的一致检索，通过主题范畴模式抽取实现动态分类，通过语义词素相似度识别以提高检索质量，为当前自然语言的检索提出了一种新的思路和解决方法。

文献领域主题信息检索应用数据库技术的系统研究是数字图书馆建设的关键之一，所述工作不仅是国内外MARC研究的热点，而且能够验证、完善相关的标准并推动标准化进程，推动我国图书、情报、档案大学科领域元数据的研究与发展，努力为国内外网上自动分类概念检索探索新路，为增强我国网络信息检索工具实用性和提高信息检索质量做出贡献。

参考文献

1 王晔，王继成，张福炎. 基于元数据的Web信息检索研究. 情报学报，2001(3)
2 British Columbia Archival Union List Background Report. Archives Association of British Columbia. http://aabc.bc.ca/aabc/bcaulbac.html，2001，9
3 Groan Kristiansson. ARKIS II — a Swedish Archival Information System. http://www.knaw.nl/ecpa/sepia/workinggroups/wp5/1.html
4 徐周，黄上腾. 基于XML实现数据库间信息交换的方法. 计算机工程，2001(1)
5 Michael J. Fox., Peter L. Wilkerson, Introduction to Archival Organization and Description. Edited by Suanne R. Warren. Getty Information Institute，1998
6 Steven L. Hensen. Archival Cataloging and the Internet: The Implications and Impact of EAD. Journal of Internet Cataloging, Volume4，No.3/4
7 Weibel，Stuart. The State of the Dublin Core Metadata ini-tiative. D-lib Magazine，April 1999. http://www.dlib/apri199/04weibel
8 黄伟红，张福炎. 基于XML/RDF的MARC元数据描述技术. 情报学报，2000(4)
9 苏新宁. 音像资料的MARC格式研究. 中国图书馆学报，2000(3)
10 Gudivada. Information retrieval on the World Wide Web. IEEE Internet Computing. 1997，1(5)
11 Pirolli P，Schank P et al. Scatter/gather browsing communicates the topic structure of a very large text collection. In: Proc of the ACM SIGCHI Conf on Human Factors in Computing Systems，1996
12 Baeza-yates R，Ribeiro B. Modern Information Retrieval. New York: Addison Wesley，1999
13 Loh S，et al. Concept-based knowledge Discovery in Texts Extracted from the Web. ACM SIGKDD，2000
14 Shin D，et al. BUS: An Effective Indexing and Retrieval Schema in Structured Documents. Proc. Digital Library 1998
15 Wolff J E，et al. Searching and browsing collections of structural information. In: Proc. Of the IEEE Advances in Digital Libraries，2000
16 Jennifer Chu，Bob Carpenter. Vector-based Natural Language Call Routing. Computational Linguistics，1999，25(3)
17 肖璐等. 多DTD环境中XML的关系数据库存储. 计算机科学，2003(10)
18，27，30 温俊，阳国贵. XML文档集公共模式获取技术研究. 计算机科学，2003(10)
19 J. F. Martinez-Trinidad. A Tool To Discover the Main Themes in a Spanish or English Document. Expert System With Applications，2000(19)
20 R. L. Popp，B. P. Maksymiuk，M. R. Poreda. Efficient information retrieval on the World Wide Web using adaptable and mobile Java agents. Proc. IEEE Int. Conf. Systems，Man，and Cybernetics，Oct，11－14，1998
21 Fox E. A Networked digital library of thesis and dissertation. D-Lib Magazine，1998，2
22 Ricardo Baeza-Yates，Berthier Ribeiro-Neto. Modern information retrieval. New York: ACM Press，1999
23 Abecker A，Bernardi A，Hinkelmann K，et al. Techniques for Organizational Memory Information Systems. DFKI Document，1998.2
24 Y. H. LI and A. K. JAIN. Classification of Text Documents，The Computer Journal. Vol. 42，No. 8，1998
25 李蕾等. 基于语义网络的概念检索研究与实现. 情报学报，2000(5)
26 朱毅华等. 计算机识别汉语同义词的两种算法比较和测评. 中国图书馆学报，2002(4)
28 刘世杰等. 基于XML技术的WEB信息提取和集成. 计算机科学，2003(10)
29 W. Qian，H. Qian，L. Wei，Y. Wang and A. Zhou. Structure-Based Query Expansion For XML Search Engine. In Proc. Of the 12th International Conference on New Information Technology，May，2002

王兰成 解放军南京政治学院上海分院信息管理系教授、室主任。通讯地址：上海，邮编200433。 (来稿时间：2004-01-05)