中国科技资源导刊 ISSN 1674－1544
2009年9月 第41卷第5期 36－41
CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
ISSN1674-1544 Vol 41 No 5 36-41,Sep 2009

科技档案异构数据整合及其检索的研究

王兰成
(南京政治学院上海分院信息管理系,上海 200433)

摘 要:实现科技档案异构数据的高效整合与检索需要进一步研究其信息的集成平台和知识环境。信息集成中XML是广泛应用的数据交换语言接口,数据仓库在信息集成中有强大的功能,.NET在分布式应用中表现出优势。平台异构的科技档案信息整合技术,能够较好地运用于数字档案文献的标题或摘要的信息标引,实现质量更高的科技档案主题概念检索。最后,介绍了笔者开发的一个NET框架信息检索实验系统。

关键词:数据整合;信息检索;Microsoft. NET;科技档案;异构数据
中图分类号: G250 **文献标识码**: A DOI: 10.3772/j.issn.1674－1544.2009.05.008

1 引 言

科技档案异构数据的整合是数字档案馆建设的关键。数字档案馆建设是进行档案自动化管理、实现档案资源共享的重要内容，目前国际上对数字档案馆的研究和建设已有很大发展，一些档案馆和技术部门正在加快进行数字档案馆关键技术的研究。与发达国家相比,我国也已经开始研究并取得了一定的成果，但研究与建设工作尚处于起步阶段。对整合的档案数据进行检索,是实现档案资源共享、最大限度利用档案资源的关键工作环节。档案著录过程中实行自动标引是提高档案著录质量、提高档案工作效率的重要途径。利用基于词首最长匹配的词典分词法，并结合基于段句分割符表及停用词表的切分标记分词法，即“正向扫描＋最大匹配＋最小推进”的中文分词,对数字档案文献的标题或摘要进行自动标引，进行检索研究,是提高检准率、检全率和充分利用档案资源的关键。因此,运用当前先进的Web信息技术和开发平台，对异构档案数据的整合与检索技术进行研究,具有很大的现实意义与应用价值。

2 异构数据集成方法及其信息整合技术

数据集成的核心是从数据源抽取信息，按照需求和数据模式转换成公共的数据格式，加载到数据仓库中并提供给目标程序使用。但数据转换过程中还必须实现对数据的转移，因此数据集成是将许多不同的功能综合在一起的一个复杂的系统[1]。当前的几种主流异构数据集成架构为:(1)基于传统数据仓库的集成架构;(2)虚拟数据整合系统;(3)联邦数据库系统(FDBS)。目前联邦数据库系统的实现方法有数据库转换法和模式转换的方法[2-6]。根据模式转换器功能的不同,模式转换的方法可以分成一对一、一对多和多对多映射,其中多对一映射是一种逻辑上异构、物理上同构的集成系统,所有操作都在核心库上进行,其他方法是要转化为对系统中某一局部数据库的操作，所

作者简介:王兰成(1962－),男,博士,教授,主要研究方向是数据库与信息处理。

收稿日期:2009年7月14日。

以模式转换的方法不会产生完整性和安全性方面的问题。

2.1 XML与信息集成

由于XML具有扩展性好、数据表达能力强以及具有自描述性、设备和平台无关的特点，已经在信息转换和集成领域得到了广泛应用。数据集成中的核心问题是信息描述的标准化，主要解决信息可理解性的问题，包括人和机器对信息的理解。更重要的是机器对信息的识别，并能根据数据进行自动处理。因而XML可以充当数据交换的中间语言。

在科技档案信息集成中对XML的支持可分为两个方面。一是信息集成支持用SQL直接访问XML文件中的数据，其机制是动态地将XML的层次型结构映像为一张或多张关系型的表来实现，数据的层次关系转化为虚拟表之间的主外键关系，这样对本地或远程XML的访问就转化为对虚拟表的访问，在应用XML作为标准信息接口规范的同时，大大简化了访问的复杂性；二是信息集成可以提供一组用于生成XML的函数，直接将数据库中的内容动态地转换为XML格式的输出，并在此基础上提供了XML元素分类及排序等功能。XML技术和信息集成技术的结合能够更有效地发挥XML的价值，满足不同层面的集成需求。

2.2 数据仓库与数据集成

数据集成是建立数据仓库的基本要求。从数据仓库的建立过程来看，由于数据仓库是面向主题的，所以首先应该根据具体的主题进行建模，然后根据数据模型和需要从多个数据源加载数据。由于不同的数据源的数据结构可能不同，因而在加载数据之前要进行数据清洗和数据集成，使得加载的数据统一到需要的数据模型下。

近年来，由于信息技术的高速发展，对数据仓库的要求也越来越高，这些要求涉及数据的时效性和可扩展性，其目的在于使用户在需要时可以得到当前的、远程的或非结构化的数据。传统的不断将新的数据源数据加载到数据仓库的方法成本很高，而且有些数据由于用法、大小或格式不适合于数据仓库或用户查询，因而不能或不需要保存在数据仓库中。通过信息集成对数据仓库进行扩展是数据仓库技术逻辑发展的必然结果。科技档案信息集成的最终目的是屏蔽数据源的复杂性，为用户提供单一的数据视图，而数据源可以分布在不同的地方，存储语义、格式不同，访问方法也不相同。用户通过SQL或XML、标准网络服务等对数据进行访问。目前，IBM DB2、Microsoft SQL Server、Oracle等主流数据库管理系统都支持数据仓库。

2.3 Web Services新兴分布式应用技术

Web Services技术是一项新兴的Web应用技术，是建立可互操作的分布式应用程序的新平台。当把应用扩展到广域网时，传统的DCOM模型就不能完全满足分布式应用的要求，一是DCOM在进行网间数据传递一般采用Socket套接字，要求开放特定的端口，这就给有防火墙的网络带来安全隐患；二是DCOM进行远程对象调用使用的协议是远程过程调用RPC，这使得基于DCOM的构件无法与其他组件模型的构件进行相互调用，比如CORBA使用的是IIOP协议，J2EE使用的是远程方法RMI。新出现的Web Services技术的特点是跨平台调用和接口可机器识别，使用简单对象访问协议SOAP作为服务调用协议。SOAP是在XML基础上定义的，完全继承了XML的开放性和描述可扩展性；使用基于TCP/IP的应用层协议(如HTTP，SMTP等)，可以很好地解决穿越防火墙的问题；更重要的是各种组件模型都可以将数据包装成SOAP，通过SOAP进行相互调用[3]。SOAP可以消除组件平台之间的差异。因此，Web Services是目前较好的一种分布式应用的层间交换技术。

实现科技档案异构数据集成的方案较多，一些专家学者分别从不同的角度论述了公共图书馆异构检索与用户统一管理平台[7]、异构数据库信息资源整合系统的实现[8-10]、Web数字文献统一检索系统[11]、数字资源整合的模式与解决方案等[12]。考察上述的技术水平和主流平台，可以认为XML是目前得到广泛应用的数据交换语言接口，数据仓库在信息的组织和信息提供方面具有强大的功能，.NET在分布式应用中表现出优异的特性。

2.4 平台异构的科技档案信息整合技术

现有的数据整合系统在解决数据异构性问题时，大多是从其数据库系统异构性出发(例如SQL SERVER的数据转换服务)，并没有考虑到平台异

构性，而现实情况中很多档案馆的数据库服务器可能并非都基于 Windows 平台，而且在数据库结构方面，由于档案管理缺乏标准化的协议，各办公自动化系统的数据源和数据结构都与归档系统不一致。这样在进行异构数据整合的时候，我们就必须重点考虑其平台的异构等特性。

本文利用 .NET Framework 平台实现异构数据库转换模型，该模型的中间层（逻辑层）的具体实现依赖于若干 .NET Framework 提供的相对独立的类，这些类并不依赖于所在的操作系统，它们在应用逻辑和功能上通过相互的协作来实现更为复杂的应用逻辑和功能，直至实现整个应用系统。同时设计了一个数据转换模块，功能是将每个数据源各自的数据接口转换到统一的接口，并协调数据整合的过程，来满足系统应用的跨平台性和可扩展性。

3 基于知识环境的信息自动标引与检索技术

科技档案信息的自动标引研究目前已有词典标引法、切分标引法、单汉字标引法、词句分析及神经网络模型理解分析法等，这些方法各有利弊。真正在各领域通用、能够保证标引质量的自动标引系统还没有实现。如果将基于词首最长匹配的词典分词法的研究，结合基于段句分割符表及停用词表的切分标记分词法，即“正向扫描＋最大匹配＋最小推进”的方法，运用于档案文献的标题或摘要的自动标引，则能改进由于缺少知识环境而使目前的信息整合检索系统满足不了用户需求的状况。

3.1 领域知识环境及切分词表设计

切分标志词表分为段句分割符表和停用词表的设计，其中段句分割符表存有各种标点符号，用来分割句子和段落。停用词表包括句子切分符号，例如《、》、(、)、—等，还有停用字（词），如“的、了、是、在、通知、命令、国务院、中央军委”等。停用词表有两个字段 unuse(停用词) 和 wzdw(文种、单位词标记)。对已切分的摘要句子正向扫描，遇到停用词表中的词剔除。经停用词表处理后标题被分为 n 块，分别存放在 source_1，source_2，…，source_n 中。

主表设计是基于档案领域的电子主题词表，并对该词表的各关系项进行了合理安排。主表结构为4个字段：主题词（T）、范畴号（N）、用词（Y）、指针（P），其它语义参照词（C、D、S、F、Z）和注释项在主表关系表中保留，并在主表的 P 字段中指向关系表中相应的记录；Y 表示正式主题词，匹配成功后，查看该词 Y 字段，若不为空则用其替换该词再登记截词结果。另外，N 皆应记录在截词结果中。对主表的访问是通过建立的索引表进行的，以提高检索速度。当系统提交需定位的汉字时，先将该汉字的国标码利用一定算法进行处理，得到在索引表中的地址；再访问索引表，根据索引表指针到主表中以该汉字开头的记录集合中；最后执行匹配操作。

3.2 词首最大匹配最小推进标引方法

对已经初步分好的各块 source_1 ~ source_n 进向正向扫描，与主题词表进行最大匹配。指针移位按照“最小推进原则”，然后继续按这种方法处理后续文字，输出分词结果。词首最大匹配最小推进分词算法（图 1）。

对各 source 正向扫描，进行最大匹配。最大匹

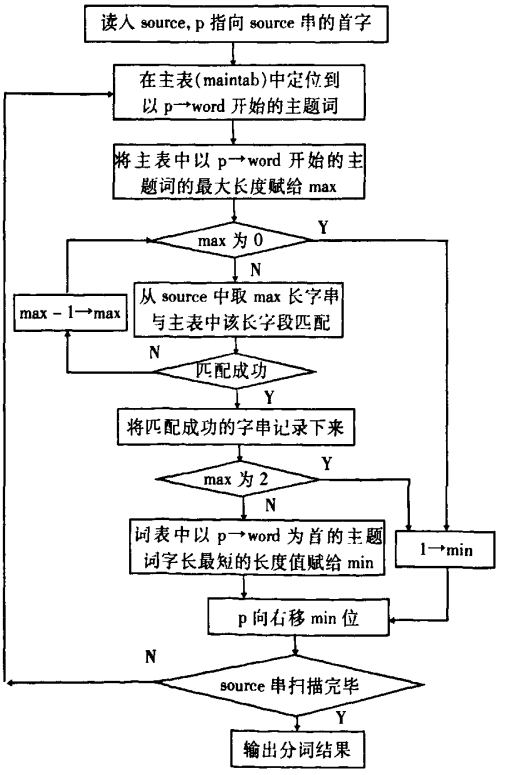

图 1 正向最大匹配及最小推进的中文信息标引

配是汉语截词常用方法，准确性相对较高，但完全依据最长的词来截词，会使标引结果忽略了主题概念组配的情况。因此，补充了“最小推进”的原则，即当在词表中按最大匹配的原则匹配到长度为 max 的词后，在以 p -> word 为首字几条记录中找到最小匹配的词，并按最小匹配词的长度移动指针 p。若没有最小匹配的词或 max 的长度为 2，则指针 p 向右移一位。如，在题名为《建设部关于进一步加强产品质量管理的决定》中，“建设部”、“关于”、“的”、“决定”被停用词表剔除。经分词处理，“进一步加强”无法匹配则被剔除。当 p→word 指向“产”时，最大匹配得到“产品质量”，而主表中有“产品”一词，所以 p 指向“质”，最大匹配后得到“质量管理”，而在主表中有“质量”一词，p 再指向“管”，得到“管理”。分词结果是产品质量、质量管理、管理。若分词结果为 0，则被视为不理想的结果。对各 source 块从后向前进行二次扫描，并在开头和结尾的段落中进行词匹配。若 source 中某词在段落中出现，则赋予此词一个权值。处理完毕后将 5 个以内权值最高的词作为参考结果。

目前，我们在自动标引技术中引入了“首字匹配”的概念，即先将主题词的第一个字符与要标引的文本进行对照，过滤掉无用信息，在保留信息中再进行二次匹配，以提高检索速度。但中文自动标引一直是难点，该方法还有一定的局限性，接入应用系统还需要继续完善。

4 一个 .NET 框架的信息检索实验系统

运用上述的研究成果，笔者开发了一个档案信息自动检索系统[13]。其功能包括：(1)数据处理，有数据导入和数据加工两个子功能。数据导入是将 Visual Foxpro 数据库、Microsoft Access 数据库、Excel 电子表格等分布式异构档案数据导入到 Microsoft Sqlserver 数据库中，再通过源文档与目标文档的字段映射来实现的；数据加工是对在档案数据库中的主题数据项内容进行自动标引，再通过档案主题词表扫描匹配档案题名抽词入库来实现的。(2)信息查询，有简单查询、组合查询和主题查询 3 个子功能。简单查询提供档案依 MARC 标准的目录信息必选项的单项检索；组合查询提供各单项档案信息的“与”检索；主题查询提供基于档案主题词表的主题词入口信息检索，及其上位检索、下位检索和同位检索。(3)系统维护，对系统的维护性进行设置，包括对登陆用户的账号和密码的修改。修改登陆口令是对具体操作员所能进行的数据处理权限进行设置，保证数据的安全性。(4)其他的系统帮助功能。

系统是建立在 .NET 框架上的网络应用集成系统。微软 .NET 平台为连接计算机、设备和服务群组，提供更广泛更丰富的解决方案，系统本身可以作为一种服务，并能够集成在高度分布式的应用服务框架中。系统采用 ASP.NET 技术，在开发中大量用到 .NET 平台的类库、数据访问和内存管理，使用 Visual Basic 作为 CODE BEHIND，安全性和效率更好。

系统开发环境为：(1) 硬件环境服务器。CPU 主频 Pentium Ⅲ 1G，内存 256MB，硬盘 20GB，网卡 10M/100M；客户端 CPU 主频赛扬 900MHz，内存 256MB，硬盘 10GB。(2)软件环境。Windows NT 或 Windows2000 Server 版网络操作系统，.NET 框架；数据库管理系统 Microsoft SQL Server 2000；软件工具 Visual Studio 7.0。系统开发在保证多个模块同步性上采用了 Microsoft Visual Sourcesafe 6.0，开发期间建立临时服务主机，数据库存储过程存于其上，确保了系统接口和样式的统一。

主要数据库表的设计：(1) 档案信息表 Marc：DH(档号，如 J009 - 01 - 00226 - 0048)，TM(题名，如关于设在上海市的外资银行、财务公司等金融机构税收优惠问题的批复)，ZRSM(文号，如沪价涉(90)第 257 号)，ZLDWXCSJ(著录时间)，ZTTM(主题词 A，如银行)，JLLY(责任者，如中华人民共和国财政部)，ZTZN(主题词 B，如文物出境)，ZTC(主题词)。(2)主题词表 ZT：ZTZN(主题词)，C1(参考词 1，用于同位词查询)，C2(参考词 2)，C3(参考词 3)，S1(上位词 1，用于上位词查询)，F1(下位词 1，用于下位词查询)，F2(下位词 2)，F3(下位词 3)，F4(下位词 4)，F5(下位词 5)。(3) 用户密码 UserTable：username(用户名，存储用户名，如 admin)，password(用户密码)。

为实现对异构档案数据的整合，与档案信息

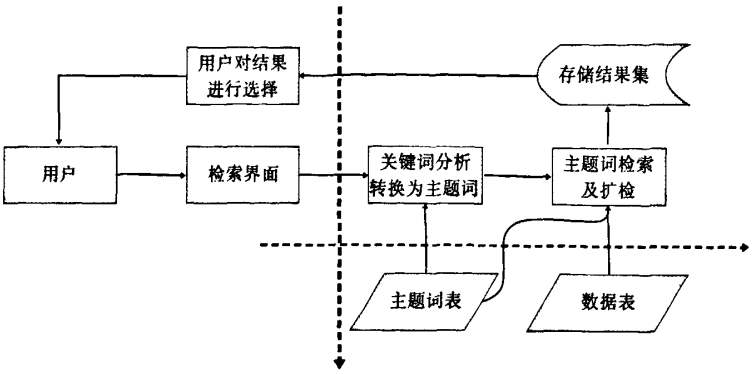

图2 系统检索工作流程

的自动标引和主题词查询的需要，该软件可实现基于异构档案数据的整合，档案基本信息的单项查询和组合查询，基于主题词和参考词（下位词、上位词）的查询，以及对档案标题信息的自动标引和手动标引。在自动标引技术中引入了“首字匹配”的概念，即先将主题词的第一个字符与要标引的文本进行对照，过滤掉无用信息，在保留信息中再进行二次匹配，以提高检索速度。基于档案主题词表，能够实现质量更高的主题概念检索，并用正向扫描最大匹配的方法实现档案主题的信息自动标引（图2）。

系统工作流程显示用户首先通过检索界面输入关键词，业务层通过主题词表对关键词进行查询分析，并转换为规范化的主题词，而后再根据主题词表当中上位词、下位词和参考词字段在数据表中进行扩展检索，输出后供用户选择相应的查询结果。例如，用户输入“文档”一词，查询分析器先在主题词表中查找该关键词，然后将该关键词转换为标准主题词“档案”，再根据其下位词“电子档案”、“科技档案”、参考词“资讯”来进行扩展检索，得到多个结果集，用户可以选择输出。另外，在实际使用过程中，由于各种新词汇的不断出现，经常导致用户输入的词汇无法在词表中检索到，这就需要一个及时的用户反馈机制。针对这种现象，系统中设计了一个自由词表，专门来存储用户输入的新词汇，并进行词频统计，最后专家通过系统将这些新词加入到主题词表当中或者直接上升为主题词，保证了系统的实时更新。

参考文献

[1] Zachary G Ives. Efficient Query Processing for Data Integration. University of Washington,2002.

[2] Busse S, Kutsche R, Leser U. Federated Information Systems: Concepts, Terminology and Architectures. Brelin: Techniques University, 1999,9.

[3] 高建强，李伟，秦克明．异构数据源间数据转换技术的研究与实现[J]．计算机工程,2005(18):93－95.

[4] 黄晓斌，夏明春．数字资源整合方式的比较与选择[J]．情报科学,2005(5):690－695.

[5] 张岩，周明全，焦翠花，网络科技资源中异构数据库访问技术的研究[J]．计算机系统应用，2008(11):87－90.

[6] 彭泽华．数字资源整合技术在数字图书馆建设中的应用[J]．高校图书馆工作,2007(5):9－12.

[7] 钟新革．公共图书馆异构检索与用户统一管理平台的实现[J]．图书馆杂志,2006(8):52－57.

[8] 李晓莹．图书馆异构数据库检索系统功能分析[J]．情报杂志,2007(2):132－134.

[9] 张燕萍．高校图书馆信息资源整合系统实现方法[J]．情报探索,2007(4):31－33.

[10] 王凤华，董玉英．图书馆电子信息资源集成管理研究现状和实践进展[J]．图书馆学研究,2006(9):34－38.

[11] 曹方，施韶亭．基于Web过程模拟的异构数字文献统一检索系统设计与实现[J]．情报学报,2006(5):575－579.

[12] 白海燕．数字资源整合的模式与解决方案[J]．图书情报工作,2005(10):54－59.

[13] 敖毅，王兰成．基于Web Services的异构数据集成技术研究[J]．上海交通大学学报,2004(37):76－80.

Research on Technology of Technique Archives Information and Retrieval Based on . NET Framework

Wang Lancheng

(Department of Information Management, Shanghai Political College PLA, Shanghai 200433)

Abstract: The knowledge environment of information integration and retrieval must be studied for digital archives. In the information integration, XML is extensively applied for the data exchange, the data warehouse has strong function in the information integration, . NET expresses advantage in the distributed system application. Now the information integration knowledge environment can't already satisfy a higher request. The method of positive scanner, the longest word match and the shortest word push has been realized

Keywords: data integration, information retrieval, Microsoft. NET, technique archives, isomeric data