数字图书馆信息检索技术及其应用

王 预
(安徽财经大学 信息工程学院,安徽 蚌埠 233041)

摘 要:从数字图书馆的发展现状展开研究,对数字图书馆与传统图书馆进行了比较,分析了信息检索概念和技术。介绍了CBR的特点,重点阐述了数字图书馆多媒体检索技术分类及各自特点,明确指出了数字图书馆建设的意义、应用中存在的问题,展望了数字图书馆的前景。

关键词:数字图书馆;信息检索;基于内容的检索;多媒体检索技术
中图分类号:TP391.3 **文献标识码**:A **文章编号**:1673-629X(2006)10-0226-04

Information Retrieval Technique of Digital Library and Its Application

WANG Yu
(College of Information Engineering, Anhui University of Finance and Economics, Bengbu 233041, China)

Abstract: This text launches research from the current situation of the development of the digital library, compares the digital library and traditional library, analyses information retrieval concept and technology, introduces the characteristic of CBR, explains multimedia classifying and each characteristic of retrieval technique of digital library especially, question existing while pointing out the meaning of construction of digital library, has looked forward to the prospect of the digital library.

Key words: digital library; information retrieval; content-based retrieval; retrieval technique of multimedia

1 数字图书馆的发展现状

1.1 数字图书馆与传统图书馆的比较

数字图书馆与现代图书馆既有联系又有区别。从组织机构角度看,数字图书馆不同于拥有物理空间的图书馆,数字图书馆没有太大的空间,信息资源也不以占有空间的多少作为图书馆规模大小的衡量标准;从资源建设角度看,现有图书馆拥有记载在多种媒体上的信息,这些资源是可视、可听、可触摸的,而数字图书馆的信息资源是电子化、数字化信息,只有经过还原才可以为人们所感知[1]。

数字图书馆建设在工作原理上与传统图书馆有许多相通之处,同样需要对信息进行收集、加工、整理和保存,只是在具体操作上与传统图书馆不同。数字图书馆是由建立在现代通信技术基础上的电子计算机技术、通信网络技术、信息处理技术共同构成的。以图书馆的概念来分析数字图书馆的工作原理和工作理念,可以概括为:

①数字图书馆仍然具有图书馆收集、加工、整理、保存信息和提供信息服务的基本功能;

②数字图书馆以计算机可处理的数字化形式存贮信息,与传统图书馆的多载体文献形式完全不同;

③数字图书馆的信息收藏在内容的广泛性和深入性上远远超过传统图书馆,它不仅收集本馆馆藏,还将全球网络上的信息资源经过筛选、处理集中在一起,其信息加工不局限于信息整体,而是深入到信息内容;

④数字图书馆提供更加广泛、迅速、便利和多形式的信息服务,它依托Internet,利用先进的信息处理技术和计算机终端设备为全球用户提供远程服务;

⑤数字图书馆与传统图书馆相互补充,相互完整[2]。

1.2 国内数字图书馆建设的现状

国内数字图书馆理论研究与建设早在1996年开始起步,1998年,由国家图书馆与北京曙光天演信息技术有限公司合作完成了国家"863"攻关项目——知识网络:数字图书馆系统工程项目。在IBM数字图书馆系统基础上,由东北大学阿尔派软件公司系统集成和二次开发了辽宁省图书馆的数字化图书馆项目。由清华、北大、上海交大等单位承担了包括数字化图书馆的结构和检索机制及应用标准和规范的研究的攻关计划。国家图书馆信息网络中心,先后实现了与CHINANET(中国公用计算机网)、CERNET(中国教育科研网)、CSTNET(中国科技网)、CNCNET(中国网通公用互联网)等国家骨干网络互联。已实现通过中国广电总局网络中心使国家图书馆与其他省市共用1000兆宽带网沟通,加速了数字图书馆建设的步伐。经过几年时间,图书情报界对数字图书馆的认识逐步深入,数字图书馆建设所需的技术条件逐步成熟,数字图书馆的理论研究也取得较大的进展,与国外的差距正在缩小,还出现了一些数字化产品等。

收稿日期:2006-01-11
基金项目:安徽省教育厅自然科学基金资助项目(2006KJ052B)
作者简介:王 预(1965-),女,天津人,副教授,研究方向为情报学、信息管理。

1.3　国外数字图书馆建设的发展状况

数字图书馆在美国被作为"信息基础技术应用"中挑战性课题进行部署,研制了很多数字化产品。1993 年美国自然科学基金会开始立项支持数字图书馆研究。1994 年 9 月,美国自然科学基金会等组织联合资助了数字图书馆研究工程 dli,六所大学利用先进计算技术和网络技术实现大规模分布式电子内容访问、互操作和应用开展研究开发工作,1998 年开始,dli 的二期工程在更大范围内展开研究。2001 年 2 月;美国总统信息技术咨询委员会 pitac 向布什总统提交的报告中就有《数字图书馆:对人类知识的普遍访问》,提出:"数字图书馆能够支持本委员会 1999 年 2 月的报告《信息技术研究:投资未来》中提出的所有"国家挑战性变革",这些挑战性变革是所有公民能够融入信息时代并从中受益的基本先决条件。数字图书馆将在这些变革中扮演核心角色,每一种变革都会利用或需要数字图书馆才能成为现实。"在报告中还提出了调查结论和建议:支持包括元数据及其应用、缩放性、互操作性、档案存储与保存、知识产权、隐私和安全、易用等数字图书馆技术研究;建立大规模数字图书馆测试床[3];联邦政府应提供必要的资源,使得所有联邦公共材料以数字形式在互联网上永久使用等等。继美国之后,加拿大、英国、法国、意大利、荷兰等许多国家也投资研究建造自己的数字图书馆。加拿大政府在 1996 年 5 月公布的《建设信息社会:使加拿大进入 21 世纪》行动计划中,很重视数字图书馆内容建设,1997 年遗产和工业部联合成立"数字化工作小组",统筹全国数字化工作,同时还计划修改版权法,解决多媒体、因特网发展带来的知识产权问题。

2　数字图书馆信息检索理论基础

2.1　信息检索概念及技术

信息检索是指对文献或记录的信息集合进行查询以检索出能够满足个人或团体信息需求或感兴趣的信息内容的过程。信息检索技术是应用于提问与文献表示的匹配比较的技术,依检索文献集合及其所用的标引方法的特性,可分为准确匹配技术与局部匹配技术两大类。准确匹配要求文献(标识) 中包含的需求模式必须与所表达的模式完全匹配,才能作为命中文献。目前绝大部分检索系统采用的布尔逻辑检索、原文检索和字符串检索技术均属于此类;局部匹配只要求文献(标识)中包括的需求模式与提问表达的模式部分匹配,即为命中[4]。信息检索技术从开始时基于关键词的检索发展到 20 世纪 80 年代基于概念的检索,再到如今基于内容的检索(CBR, content - based retrieval)。这一演化过程反映了对某一文献的检索由对内容知识的检索代替了关键词、概念知识的检索。目前绝大多数中文检索仍停留在关键词检索阶段,甚至是"字"索引阶段,运用的是关键字匹配算法,效率低且检索精度差。

现在有人指出:"未来的信息系统应当是概念匹配,又称语义检索"。但是笔者认为今后数字图书馆的信息检索应当是基于内容的检索,因为数字图书馆所处理的对象是数字化的信息资源,既包括数字化的文本信息、图形与图像信息,又包括数字化的音频与视频信息。对于这些结构化信息,赋词标引方法为主的目录或摘要二次文献,或以词检索为主(包括概念检索)的全文检索均不能满足数字图书馆信息检索的内在需求,应该采用新的信息检索技术——基于内容的检索。

2.2　数字图书馆信息检索方式的特点

数字图书馆完全将信息实体虚拟化,在网络环境下,以各类文献为载体的知识信息,都可以方便地转化为数字形式,在全球范围内传输,读者可以在网上浏览、下载。就信息载体而言,数字图书馆提供各种各样的电子化文档,供读者在网上浏览、下载,电子文档是用字节的形式贮存在存储介质上的,因此管理是完全数字化、无纸化的,归类、制作、提供都需要通过计算机来进行。同时数字图书馆只要求用很少的人员来管理庞大的数据资源,其管理也将是完全自动化的。这就要求数字图书馆在建设时必须能够满足用户对信息管理的数字化、自动化管理的要求,例如系统应该能自动处理用户请求,判断其权限,并自动提供相应的服务;对读者信息自动分类识别;工作人员信息的自动化管理等。

2.3　CBR 与传统信息检索相比的特点

(1)从媒体内容中提取信息线索。CBR 突破了传统的基于表达式检索的局限,它直接对文本、图像、视频、音频进行分析,抽取特征,利用这些内容特征建立索引并进行检索;

(2)CBR 是一种近似匹配(或称局部匹配)。传统的信息检索都是采用准确匹配技术,CBR 是采用相似性匹配的方法逐步求精获得查询的结果;

(3)大型数据库(集)的快速检索。CBR 拥有数量巨大、种类繁多的多媒体数据库,能够实现对多媒体信息的快速检索;

(4)能满足用户多层次的检索需求。CBR 数据库系统通常由媒体库、特征库和知识库组成,媒体库包含多媒体数据,如图像、视频、文本等;特征库包含用户输入的特征和预处理自动提取的内容特征;知识库包含领域知识和通用知识,其中的知识表达可以更换,以适应各种不同领域的应用要求;

(5)CBR 是一种交互式的、形象化的检索。通常按照与用户输入的查询信息的相似程度排列检索结果,并能为用户提供人机交互和形象化的操作示例与浏览界面[5]。

3　数字图书馆多媒体检索技术分类及各自特点

3.1　文本信息检索

文本信息检索包括两方面核心技术:如何建立和维护检索索引库和如何提供快速有效的检索机制。基于内容的文本信息检索是涉及文档内容查寻的检索技术,检索模型的构造是基于内容文本信息检索的核心技术,检索模型

包含3方面内容:文档与用户查寻的表示;查寻匹配策略;匹配结果的相关度表示。

常用检索模型有布尔模型、向量空间模型、概率模型。

(1)布尔模型是一种严格匹配模型,定义了一个二值变量集合表示文档,这些变量对应于文档中的特征项,若词条对文档内容贡献则赋予True,否则置为False。检索时,根据用户提交的检索条件是否满足文档表示中的逻辑关系将检索文档分为两个集合:匹配集和非匹配集。P范数模型是对布尔模型的扩展,它克服了简单布尔模型匹配函数过于严格而导致漏检率高的缺陷。

(2)向量空间模型(V,SM)将文档作为是由相互独立的词条组($T_1, T_2, \cdots, T_n$)构成,对于每一词条T_i,根据在文档中的重要程度赋以一定的权值W_i,并将$T_1, T_2, \cdots, T_n$看成一个n维坐标系中的坐标轴,$W_1, W_2, \cdots, W_n$为对应坐标值。由($T_1, T_2, \cdots, T_n$)分解而得的正交词条矢量组形成一个文档向量空间,文档则映射成为空间中的一个点。对于所有文档与用户查寻都可映射到此文档向量空间,用词条矢量($T_1, W_1; T_2, W_2; \cdots; T_n, W_n$)表示,从而将文档信息的匹配问题转化为向量空间中的矢量匹配问题处理。

(3)布尔模型和向量空间模型都将文档表示词条视为相互独立项,忽略了表示词条间的关联性,概率模型考虑到词条、文档间内在联系,利用词条间和词条与文档间的概率相依性进行信息检索。其中二值独立检索模型(B,IR)是实现简单且效果较好的检索模型[6]。概率推理网络是一种新型检索模型,它模拟人脑的推理思维模式,将文档内容与用户查寻匹配的过程转化为一个从文档到查寻的推理过程。

3.2 图像信息检索

基于内容的图像信息检索技术是把图像的可视特征(如颜色、纹理结构、轮廓、位置关系等)作为图像内容进行匹配、查找。图像特征的抽取是图像信息检索基础,图像特征主要包括图像的颜色和纹理结构,图像颜色通常以色彩直方图表示,图像纹理结构是指图像的像素灰度级或颜色的某种规律性变化,其变化是与空间统计相关的。纹理结构反映了图像本身的属性,不同的物体有明显不同的空间特征[7]。

基于色彩直方图图像信息检索的核心思想是在一定的色彩空间中对图像各种颜色出现的频数进行统计。色彩直方图能较好地反映图像中各颜色的频率分布,但没有保留像素的空间位置信息,为此可以采用基于图像分割的直方图检索方法。一种方法是对图像中所包含的对象的边界进行提取,然后对每个对象所包含的颜色进行直方图统计,以减少图像中不相关信息的干扰,采用snake边界提取算法、手工交互结合、小波变换等方法对图像对象的轮廓进行提取,对象区域中像素的统计采用计算机图形学中的跨距扫描线算法,采用距离法对待查图像和数据库中的图像进行相似匹配。另一种方法是将一幅图像划分为$n \times n$个子图像,然后对应位置的子图像进行比较。采用图像分割的直方图检索方法比整幅图像直方图检索方法的检索精度有较大幅度的提高。基于图像颜色的另一种查寻方式是将图像的颜色主色调作为图像的颜色特征进行相似性匹配,以查找图像库中具有类似主色调的图像。一幅图像的主色调能够反映该图像的基本概貌,因而可作为查寻的主要特征。基于知识的图像信息表示及检索技术是图像检索的另一重要方法,图像本身是一定数量的带颜色的像素点集合,人类能够识别出像素点集合的含义是人类以自身的知识不断赋予图像意义并将其提升的过程[7]。计算机可以为图像检索提供知识库,基于知识的图像检索流程如下:针对一个查寻要求,检索引擎依次调入每幅图像的事实,结合知识库中的知识以查寻要求为目标进行推演,若查寻目标得到满足,则这幅图像是符合要求的,否则不符合要求。

3.3 音频信息检索

常见的声音媒体是语音的音乐,对声音进行数字化处理得到的结果称为音频。音频是一种正弦波,检索前需进行预处理或媒体转换,以提取音频特征或文本描述。对于基于内容的音频信息检索,应提取数据的音频特征,然后对音频特征进行匹配,从而进行音频数据的分类和检索。在音频数据中提取特征有两种方法:一是提取感性特征,如音高、响度;二是计算非感性属性或称物理属性,如线性预测系数、对数倒频谱系数等。特征提取通常在频域进行,先对音频数据进行加窗处理,加窗大小在10~30ms左右,然后对加窗后的音频数据即每一帧作离散傅里叶变换(DFT),常用快速傅里叶变换(FFT),最后应用不同的算法计算相应的特征[8]。音频检索的基础是建立数据库,对音频数据进行特征提取,将音频数据装入数据库的原始音频库部分,通过特征对音频数据聚类,将聚类信息装入聚类参数库部分。数据库建立后就可以进行音频信息检索。音频信息进行检索主要有3种方式[9]:

(1)基本属性检索:通过查找文件名、文件大小、生成时间等一般属性及取样率等音频属性来检索音频信息;

(2)特征值检索:通过查找声强、能量、带宽等特征值进行音频信息检索;

(3)示例检索:通过查找与给定音频相似的音频数据来检索音频。

3.4 视频信息检索

视频是一个时间坐标上的图像帧序列,视频内容可以用全局和局部特征来表示。视频的名字、产生日期、制作人和一些描述性文字等属于全局特征,它们表明了视频的整体属性。另外,可以将视频看成一个个连续镜头的集合。视频中镜头之间的相互关系,即视频的结构,体现了制作人拍摄视频时所运用的手段,这种风格也属于全局特征。视频具有不同的时间跨度,不同时刻的镜头描述的内容不一样。因此,这些不同镜头所描述的内容体现了视频的局部特征,同时也具有空间上的局部性。视频的

内容可以用切分点检索算法、视频结构化算法有效地表示。视频检索过程如下:

(1)根据运动和视觉信息将一段视频分为视频序列,在此基础上构造高层的语义结构,如场景等,同时在镜头内找到若干个关键帧来代表镜头的视觉内容;

(2)在视频结构化的基础上,提取各关键帧的观察特征以及运行参数和像机参数,并存入视觉数据库中;

(3)由用户构造查寻,系统基于数据库中的特征处理查寻并将结果反馈给用户[9]。目前常用的视频查寻方式是关键字查寻和示例查寻,前者是指用户输入若干个查寻主题,如导演、影片名等,要求找出相关视频,处理这种查寻是通过查找全局数据库中的注释来实现的;后者是根据用户提交的视频例子,在视频特征库支持下,定义一个相似度模型,然后计算特征向量距离来实现视频检索。

3.5 媒体检索技术

多媒体检索的对象涵盖了文本、图像、视频、音频等信息。一般来说,检索文本信息采用的是比较先进的全文检索技术。在检索过程中,中文分词问题是全文检索的关键技术,把按词典进行的最大匹配、逆向最大词组匹配、最佳匹配法、基于神经元网络和专家系统的分词方法、基于统计和频度等方法运用于全文检索,效果不甚理想。智能化检索将进一步推动全文检索技术的发展。对图像信息可以按颜色、形状、纹理及在图像中的位置查找对象。

4 数字图书馆建设的意义、存在问题及展望

数字图书馆是一个应用前景非常广阔的研究领域,特别是对于教育,数字图书馆将成为非常重要的教育设施,由国家出面组织建设数字图书馆的意义很大。首先,数字图书馆是一个国家的数字文化平台,包括图书馆、博物馆、档案馆、大学、政府部门提供的各种文化资源。第二,数字图书馆还应该是一个国家数字教育平台,成为网上业余教育中心、在职教育中心,甚至趣味教育中心等。第三,数字图书馆也是一个国家数字资源中心,包括卫星、遥感、地理、地质、测绘、气象、海洋等科学技术数据和人口、经济等统计数据。

基于内容的多媒体检索是一个新兴的研究领域,在国内外仍处于研究、探索阶段,因此在基于内容的检索领域中仍然存在许多问题。这些问题主要包括多媒体特征的描述和特征的自动提取、多媒体的同步技术、匹配和结构的选择问题,以及按多相似性特征为基础的索引、查询和检索等。作为一个新兴的研究领域,同时由于其检索对象和范围的多样性,基于内容的多媒体检索还要解决多种检索手段相结合的问题,以提高检索效率。此外,更好地理解检索内容以及使检索性能更接近人类视觉的特征,也是未来研究中需要解决的问题。

中国目前正加大力度投资 3 亿元建立首都数字图书馆和建设中国数字图书馆。数字方舟信息公司已同中国数字图书馆签订了合作协议,并与首都图书馆商讨进一步合作的意向。建立数字图书馆需要大量的资金投入,也需要有较高的信息管理素质的专业人才。因此,在中国建立数字图书馆还需要一个漫长的过程,但随着网络信息的发展和经济实力的不断增强,数据图书馆技术的应用在中国将会有特别广泛的市场。

参考文献:

[1] 王彩霞,万　君.数字图书馆的发展及其相关技术[J].信息技术,2002(9):23－25.

[2] 孙　坦.论数字图书馆与传统图书馆的关系[J].大学图书馆学报,2001(2):27－30.

[3] 汤珊红.数字图书馆的发展动态及相关问题研究[J].图书情报知识,2000(1):13－15.

[4] 盛小平.数字图书馆的信息检索技术.业务知识讲座[J].图书馆理论与实践,2001(3):22－25.

[5] 邹　涛.文本信息检索技术[J].计算机科学,1999(9):31－34.

[6] 齐向华.文本信息检索模型[J].晋图学刊,1998(3):16－19.

[7] 李国辉.几种典型的基于内容检索系统[N].计算机世界,1998－05－18.

[8] 薛　锋.基于内容的音乐检索[J].大学图书馆学报,1999(4):25－27.

[9] 周立柱,邢春晓.数字图书馆技术研究与应用[J].中国计算机用户,1999(49):31－33.

(上接第 225 页)

第三步::拉开 D。$F(D)=1$。故可以拉开 D,计算机拉开 D 并将[StatusCode]的值由"1"改为"0"。经上述模拟操作,验证该操作票是正确的。

4 结　论

本系统和 SCADA 系统相连接,应用于 DTS 仿真培训功能中,形成操作票,进行调度员考核,具有重大意义和实际效果。该系统于 2004 年底在河南南阳供电局投运,大大提高了仿真培训的效率和操作票生成的准确性。目前,该系统运行正常,满足培训需求。

参考文献:

[1] 彭云建,申群太.于面向对象编程技术的调度操作票专家系统[J].中南工业大学学报,2002,33(3):313－316.

[2] 胡海涛,孙宏斌,张伯明,等.变电站操作票专家系统的研究与应用[J].电力自动化设备,2002,22(8):42－45.

[3] Giarratano J,Riley G.专家系统原理与编程[M].北京:机械工业出版社,2000.

[4] 李晓明,戴承伟,王　平,等.无人值守变电站操作票专家系统[J].中国电力,2001,24(增刊):60－62.

[5] 刘　蔚,杨宛辉.操作逻辑函数在操作票专家系统中的应用[J].电力系统及其自动化学报,1999,11(4):39－43.