

● 图书情报学

数字图书馆中基于内容的多媒体 数据查询和检索技术

林 慧 勤

(深圳职业技术学院 图书馆, 广东 深圳 518055)

[作者简介] 林慧勤(1964-),女,湖北武汉人,深圳职业技术学院图书馆馆员,主要从事图书馆自动化研究。

[摘 要] 数字图书馆是一个开放式的硬件和软件的集成平台,不再是传统意义上的图书馆。基于内容的多媒体数据查询和检索技术是数字图书馆中至关重要的一项技术。目前,基于内容的检索研究已经取得了很多成果,然而还有许多问题需要去探索。

[关键词] 数字图书馆;基于内容的检索;多媒体

[中图分类号] G 250.76 [文献标识码] A [文章编号] 1000-5374(2000)06-0841-04

“数字图书馆”一词是由英文“Digital Library”翻译而来。“数字”(Digital)是指计算机可识别的方式,即“0”和“1”。在数字图书馆中,可以利用当今先进的数字化技术,将图书馆馆藏信息数字化,通过诸如 Internet 等计算机互联网络上网服务,供用户随时随地查询,使处在不同地理位置的用户能够方便地利用大量的、分散在不同存储处的电子物品的全部内容。这些电子物品包括网络化文本、地图、图表、声频、视频、商品目录,以及科学、企业、政府的数据集,还包括超媒体和多媒体等。

数字图书馆是一个开放式的硬件和软件的集成平台,不再是传统意义上的图书馆。数字图书馆的任务虽然仍为传统图书馆所具备的信息采集、存储、传播及版权控制,但通过对现有技术和正在研究的技术的应用和集成,其各项任务的内容、采集的手段以及服务的方法均发生了根本的变化。在其中所用到的相关关键技术包括:

1. 数字化信息的生成。

数字图书馆的存储对象是数字化信息,一方面需要把大量现已存在的、以不同形式和载体存储的信息资料,如文件、图片、声像资料等,转化为能够为

计算机处理的数字化信息。直接用于该任务的技术包括图像扫描与处理,文字、图像和语音的识别以及对数字化初始信息的各种加工技术。另一方面则是提供直接生成数字化资料方法,以便为计算机直接使用。为便于信息的存储,降低信息传送成本,还有相应的数据压缩与转化技术。

2. 数字化信息的存储。

与传统图书馆不同,数字图书馆的信息资源是存储在计算机、磁带或光盘中,信息存储量是以存储多少 GB 字节来计算的。对于数字图书馆来说,其存储量应该是几百个 GB,几千个 GB 或者更大,即海量信息。随着信息量的剧增,存储规模也越来越大,已用万亿(或太)字节来表示。存储这些海量信息主要涉及存储设备的容量、硬件随机读取的速度以及数据集中与分布存储管理方法等问题。目前,其研究热点首先集中在存储管理方式问题上。

3. 数字化信息的访问和查询。

更有效的文本数据查询和多媒体资料的查询策略是数字图书馆的重要技术。

由于信息的多样性,对检索方法也必然提出新的需求。原有的文本类型的文件检索,传统的基于关键字索引以及通过布尔表达式的查询方式已无法满足用户的查询要求,现在正向全文检索方向发展。随

着 Internet 上 WWW 服务器的发展,网上已出现了快速全文检索软件。

为多媒体信息建立索引,便于用户快速和简便地对声音、图像等信息内容进行筛选和确定,这是当前研究的热点问题之一。

4. 数字化信息的传播。

数字化信息传播的主要问题在于网络,图像、音频和视频信息的传输对于现有网络来说负担过于沉重。单频和视频信息的实时传送,不仅对网络的速度有很高的要求,同时对提供信息的相关服务器的共享性也有很高的要求。在这一方面,除正在研究与改善网络协议,研制新的网络互联设备外,随着文件的多样性,各种压缩技术也提高了传播速度。

5. 数字化信息的权限管理。

由于数字图书馆的主要结构原则是对不同类型开放环境的操作能力,为用户共享信息资源,获得信息服务提供了极大的方便。因此,对数字图书馆来说,安全性显得非常重要。数字图书馆具有一般计算机网络系统的管理功能,旨在对访问权的控制,该问题在利用原有计算机读取管理技术以及域名管理技术的基础上已得到了较好的解决。更重要的是知识产权的保护,通过采用技术手段来防止未经授权使用版权人的资源,用来保护信息拥有者和最终用户的利益,如标记数字化图像的技术,加密技术(像“水印”技术),还有数据的检测、记账等。

二

对于多媒体数据来说,每一种媒体数据都具有难以用符号化的方法描述的信息线索,例如图像中某对象的形状和颜色、视频中的运动、声音的音调等。当用户具有这样的线索,并且要利用这些线索去对数据进行检索时,目前他不得不首先将其转化为某种符号的形式。但是,这种转换带有一定的主观性,并且是极其费时的。现在的计算机所能处理的信息媒体范畴迅速扩大,不仅要求数据库和其它信息系统能对图像、视频和声音等媒体进行存储和基于关键字的检索,而且要对多媒体数据的内容进行语义分析,以达到更深的检索层次。

所谓基于内容的检索(CBR, Content-Based Retrieval),是指根据媒体对象的语义和上下文联系进行检索。它有两个方面的特点,即从媒体内容中提取信息线索和大型数据库(集)的快速检索。

人一般对于特征比较敏感,能迅速分辨出目标的轮廓、音乐的旋律以及它们的含义等。但对于大量的对象,一方面难以记住这些特征,另一方面人工从

大量数据中查找目标非常费时,又是一种重复性工作,而这些正是计算机的长处。

因此,在实施 CBR 系统时,要充分发挥人和计算机各自的长处。基于内容的检索要利用图像处理、模式识别、计算机视觉、图像理解等学科中的一些方法作为部分基础技术。CBR 不仅仅是基于内容,而且是一门信息检索技术。它从认知科学、用户模型、图像处理、模式识别、知识库系统、计算机图形学、数据库管理系统以及信息检索等领域中获得启发,导致了新的媒体数据的表示和数据模型、有效和可靠的查询处理算法、智能查询接口以及与领域无关的检索技术和系统结构的产生。CBR 与模式识别、图像理解、计算机视觉等学科的重要区别是,CBR 是一种信息检索技术,要能够从大型分布数据库(集)中以用户可以接受的响应时间查询到要求的信息,它不需要去理解或识别媒体中的目标,关注的是基于内容快速发现信息。

1. 基于内容查询和检索的系统结构。

作为一个信息检索的核心技术,CBR 应接入或嵌入到其它多媒体系统中去,如超媒体(浏览器)系统、多媒体信息系统、关系数据库系统等,提供基于多媒体数据内容的信息查询和检索。因此,把 CBR 设计为多媒体数据库(信息)进行存取的检索引擎结构。CBR 系统一般由二个子系统构成:数据库生成子系统和数据库查询子系统。每个子系统由相应的功能模块和部件组成,包括:(1)目标标识,即为用户提供一种工具,以全自动或半自动(需用户部分干预)的方式标识在静态图像、视频镜头的代表帧等媒体中用户感兴趣的区域(静态目标),以及视频序列中的动态目标,以便针对目标来进行特征提取并查询。(2)特征提取,即对视频、图像等多媒体数据进行自动或半自动的特征提取,提取用户感兴趣的、适合检索要求的特征。(3)数据库,生成的数据库由媒体库、特征库和知识库组成。媒体库包含多媒体数据,如图像、视频、音频、文本等。(4)用户查询和浏览接口,主要以示例查询(QBE)和模糊描述等可视查询向用户提供查询接口。(5)检索(匹配)引擎,检索是利用特征之间的距离函数来进行相似性检索。对于不同的媒体数据类型,具有各自不同的相似性测度算法。(6)索引/过滤器,检索引擎通过索引/过滤模块来达到快速搜索的目的,从而可以应用到数据库中的大型多媒体数据集中。过滤器作用于全部数据,过滤出的数据集合再用高维特征匹配来检索。

2. 基于内容的查询和检索过程。

基于内容的查询和检索是一个逐步求精的过

程,基于内容的检索存在一个特征调整、重新匹配的循环过程。其过程如下:(1)初始查询说明。用户查找一个对象时,最初可以用QBE或查询语言来形成一个查询。(2)相似性匹配。将查询特征与特征库的特征按照一定的匹配算法进行相似匹配。(3)满足一定相似性的一组候选结果按似度大小排列返回给用户。(4)特征调整。对系统返回一组满足初始特征的查询结果,用户可以通过遍历(浏览)来挑选到满意的结果,或者从候选结果中选择一个示例,经特征调整,然后形成一个新的查询。(5)如此逐步缩小查询范围,直到用户满意查询的结果为此。

3. 快速存取方法。

当数据库中包含成千上万的图像、视频、文本等媒体数据时,为了有效地查询,合适的存取访问结构和检索方法就变得非常重要了。在常规数据库中采用像R树的结构索引,提供有效的访问机制,即向用户返回满足查询条件的记录,而不需要逐个检索数据库中的每个记录。但是,对于多媒体数据库,传统的索引方法可能不再适用,因此需要研究新的适合于多媒体内容特征匹配的快速访问结构。对于大容量媒体库的快速搜索,可以用以下两种技术:(1)索引。索引是对特征库的快速访问。对于数据库中每个数据项(特征项),索引项包含关键属性值,以及可以直接访问该数据项的指针。索引构成树结构,索引树中的中间节点是它们的子节点的抽象。一个索引树既可以自底向上通过抽象来构造,也可以自顶向下方向通过分类来构造。

对于多媒体数据,不仅仅用一个关键字属性来产生一个索引树,而是利用一种抽象数据类型,它可以是特征(特征矢量)、多维矩阵(多维特征、图像数据、图像序列等)、或指向数据结构的指针。在索引树的不同的级别上,允许所用到的关键属性可以不同。例如,在索引树的顶层,关键属性可以是面部轮廓特征,在下一级可以是头发特征。另外,分级的准则和函数不仅是基本的逻辑表示式,例如, $01 \leq a_i \leq 02$,而且主要是特征矢量的相似性度量,或用集束(cluster)的方法来归类相似的视频镜头,还可以利用特征的空间、时间约束来分类等。(2)过滤。用快速计算的过滤器扫描数据库中的所有特征数据,只有通过过滤器的项才完整地计算相似度,这样可以加快检索过程。例如,对于颜色查询,首先用非常快的三维颜色空间算法对10000个元素的数据库进行一次过滤,产生1000个初步的匹配集合,然后这些已过滤的元素顺序进入较慢的256维颜色直方图(矩阵矢量相乘)的计算和匹配操作,最后获得最佳

的20幅匹配图像显示给用户。对于每种查询类型,都有相应的过滤器。

4. 视频检索和浏览。

现有的多媒体信息系统,包括能够存储多媒体数据的关系数据库系统和VOD系统,对视频媒体仅仅局限于存储和关键字的检索。从内容的观点,这样的系统所管理的数据资源是非结构化的,即没有对视频资源做任何分析,因此不能索引,不能进行基于结构特征的存取。因此,要对视频进行真正意义上的查询和检索,就要抽取并描述视频的结构化特征。

视频数据可用幕、场景、镜头、帧等描述。帧(frame)是一幅静态的图像,是组成视频的最小单位。镜头(shot)是由一系列帧组成的一段视频,它描绘同一场景,表示的是一个摄像机操作,一个事件或连续的动作。场景(scene)包含有多个镜头,针对同一批对象,但拍摄的角度不同,表达的含义不同。幕是由一系列相关的场景组成,包含一个完整的事件或故事情节。

检索和浏览首先需要视频数据能够被有效地索引,以前的索引是基于关键字的,而基于内容的检索需要利用可视特征来建立索引。具体来说,常用的特征是:(1)镜头和代表帧。任何视频都是由一个个镜头衔接起来的,因此镜头是视频检索的基本单元。因此对视频中的镜头分割是视频分析中最基本的内容。视频分割即将视频数据分割为一个个镜头的过程,其核心处理是识别镜头的切换。视频分割成镜头后,要从每个镜头中抽取代表帧,简称R帧,又称关键帧。代表帧是用于描述一个镜头的关键图像,它反映镜头的主要内容。对代表帧的特征提取和一般静态图像的特征提取是一样的,包括颜色、纹理和形状等特征。(2)运动特征。代表帧是视频的静态特征,但视频还具有动态的特征。视频的运动特征主要有:摄像机操作和目标运动。目标的运动可以用运动方向和运动幅度来描述。事实上,许多目标的运动也与摄像机操作有关,通过对视频的研究发展,当目标运动时,在视频上表现为背景在迅速地变化,运动目标实际上相对镜头没有太大的运动,而是摄像机在操作,这样就可以利用摄像机操作的特征进行检索。

一旦建立了视频内容的表示,就可以在这些内容之上进行基于内容的视频检索和浏览了。即利用相似性测度,基于关键帧特征,或镜头的时间(运动)特征,或二者的结合来查询。查询过程仍然可以迭代,以系统可以接受的反馈重新形成搜索。

目前,基于内容的检索研究已经取得了许多成果,推出了一些原型系统,某些技术已试用于国内外

一些数字图书馆之中。然而,还有许多问题需要去探索和研究。

三

随着计算机应用的普及,Internet 的迅速发展,数字图书馆的网络化、互联化、世界化的程度将进一步加深,而且各个数字图书馆也不会再孤立地存在,而是和其它许多图书馆紧密协作,形成数字图书馆群,实现资源共享,共同为用户提供服务。甚至最终有可能全世界的所有数字图书馆都实现协作与资源共享,形成一个规模空前的“世界图书馆”,成为涵盖人类几乎所有知识的超级信息资源库。就目前来说,也正在朝这一趋势发展。一部分已建立的数字图书馆都通过 Internet 进行了互联。反过来,数字图书馆的发展也将进一步推动 Internet 的发展。数字图书馆提供的文字、图像信息需较大的网络带宽,必然使 Internet 的主干网向宽带、高速、综合服务的发展方向,数字图书馆也将吸引更多人成为 Internet 的使用者。

另一方面,数字图书馆的发展也将促进出版物的电子化、格式的标准化、统一化。目前,出版物的电子化是一种趋势,许多刊物都以光盘或者直接在 Internet 上发行。出版物的电子化给数字化图书馆的发展带来了极大的便利。因此,免去了扫描、图文转换等劳动。但由于各种出版物的格式不统一,也给数字图书馆带来了新的问题,很多时候不得不进行格

式间转换,加大了处理难度。因而,需要制定出出版物的标准格式。在标准格式中,不仅对出版物的各个构成元素有标准的格式要求,还可能给各个出版物以统一的索引代号,这样,便于使出版社的电子刊物无需过多处理便可直接进入数字图书馆,降低了数字图书馆的技术复杂性,提高了可靠性和易维护性。

再一方面,数字图书馆的发展必然推动其技术的发展,例如图像处理(包括图像抽取、合成压缩)技术、多媒体技术、数据库技术、网络技术等。这些技术在数字图书馆中有重要的应用,数字图书馆的发展必然推动这些技术的相应发展。

[参 考 文 献]

- [1] 徐文伯. 建设中国数字图书馆工程[N]. 光明日报, 1999-06-07.
- [2] 张丽虹. 数字图书馆及其相关问题和技术研究[J]. 图书杂志, 1997, (1): 32-34.
- [3] 汪冰. 数字图书馆: 定义、影响和相关问题[J]. 中国图书馆学报, 1998, (6): 9-17.
- [4] 胡晓峰, 李国辉. 多媒体系统[M]. 北京: 人民邮电出版社, 1997.
- [5] KLEMPERER K. Digital libraries: a selected resources guide [J]. Information Technology & use, 1997, 16(3): 26-131.

(责任编辑 叶娟丽)

Content-based Query and Retrieval Technology of Multimedia Data of Digital Libraries

LIN Hui-qin

(Shenzhen Polytechnic College, Shenzhen 518055, Guangdong, China)

Biography: LIN Hui-qin (1964-), female, Librarian, Shenzhen Polytechnic College, majoring in the library automation.

Abstract: Digital Libraries is an open and integrated surface that based on the hardware and software. Digital Libraries is a new concept which different from the Conditional Libraries. Now, the content-based query and retrieval technology of multimedia data is becoming the key technology of Digital Libraries. We have taken in content-based retrieval, but there are a lot of projects to be studied.

Key words: Digital Libraries; content-based retrieval; multimedia