海量视频数据标引平台的设计和开发

张 博,张 勇,朱 义,邢春晓
(清华大学 信息技术研究院, 北京 100084)

摘 要:就海量视频数据进行标引的方法进行了阐述,对元数据、Dublin Core、OA IS进行了分析.通过研究这些技术在视频挖掘中所具有的优势,提出了一种海量视频数据标引平台的体系结构并实现了相关的功能模块,并对视频数据的搜索提出了一个基于标引的解决方法.实验结果证明,该平台可为互联网视频搜索的发展提供更加快捷、方便、准确的标引和检索模式,有效降低了用户获取相关视频数据的时间.

关键词:数据挖掘;标引;都柏林核心元数据集;元数据;开放归档信息系统
中图分类号: TP31 **文献标识码**: A **文章编号**: 1673-4785(2009)02-0107-05

Research and development of a massive video data indexing

ZHANG Bo, ZHANG Yong, ZHU Yi, XING Chun-xiao
(Research Institute of Information Technology, Tsinghua University, Beijing 100084, China)

Abstract: This paper describes indexing methods for massive video data. It analyzes metadata, Dublin Core, and the open archival information system (OAIS) in detail. To apply these technological advances to video mining, we suggested a platform for massive video data indexing and the relevant functional modules were established. Furthermore, a solution for video searches based on indexing was proposed. It was proven by experiments that this platform, with its more detailed and convenient indexing approach, would be a great help to the development of Internet video searches by effectively saving users' time and energy in the search for valuable data.

Keywords: data mining; indexing; Dublin core; metadata; open archival information system (OAIS)

世界已经进入一个信息化、高速化的阶段,流媒体已经越来越广泛地在日常生活中得到应用,互联网上视频类文件呈现直线上升态势. 2009年 1月,中国互联网络信息中心(CNNIC)发布的《第 21次中国互联网络发展状况统计报告》显示:网络视频用户相比 2007年底净增 4 000多万用户,达到 2.02亿[1].随着网民数量的不断增加,更多的用户喜好在互联网上收看视频类文件.但随即出现了需要考虑的问题,视频文件不同于文本文件,可以直接搜索查找,对于视频文件,搜索起来是相当困难的.

为了解决这个问题,对这个问题进行了详细的研究.首先,对于视频海量数据,如果是 MPEG7标准的,那么在文件的头部可以获得一些关于视频内容的描述性信息,通过元数据抽取,然后针对视频元数据进行详细标引.如果是非 MPEG7标准的,可以直接对相关的元数据进行标引.这样,将基本的视频文件元数据信息储存至数据库.

1 相关技术和知识

1.1 数据挖掘

数据挖掘(data mining),又称为数据库中的知识发现(knowledge discovery in database, KDD),就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的、人们事先不知道的,但又是潜在有用的信息和知识的过程[2].

收稿日期: 2008-12-16.
基金项目:国家"863"计划资助项目(2009AA01Z143);铁道部—清华大学科研计划资助项目(J2008X009).
通信作者:张 博. E-mail: hackfly@163.com.

数据挖掘可以在任何类型的数据上进行,既可以来自社会科学,又可以来自自然科学产生的数据,还可以是卫星观测得到的数据.数据形式和结构也各不相同,可以是传统的关系数据库、面向对象的高级数据库系统,也可以是面向特殊应用的数据库,如空间数据库、时序数据库、文本数据库和多媒体数据库等,还可以是 Web数据信息[3].

1.2 元数据

元数据 (metadata) 是从数据发展而来,同时作为数据的一种功能,这就是为什么将其称为"数据的数据"或者"信息的信息".在实际使用中,元数据以标签或标记的形式存在,用于标识所有类型的信息.一条元数据记录由一组属性或元素组成,这些属性或元素对于描述被查询的资源是必需的.元数据有助于查找和描述信息资源以促进和改善对资源的检索、管理和利用.在那些需要制作或管理大量文件的环境下,元数据有多种不同的用途.在图书馆里,使用标准工具生成的元数据被广泛用于资源描述,提高了检索的效率和可靠性.在网络环境下,元数据被用于尽可能地挑选出大量的可用信息,从而改进万维网上可用信息的可获取性.除了捕获和检索科研语境中的结构化信息之外,元数据还可以帮助组织电子资源,促进其互用性,验证其标识,以及确保对它们的长期保存.通过元数据可以检索、访问数据库,有效利用计算机的系统资源,以及对数据进行加工处理和二次开发等[4].

1.3 DC (Dublin core)元数据

1995年 3月,由 OCLC (online computer library center)和 NCSAC (national center for supercomputing)联合在美国俄亥俄州的都柏林镇召开的第一届元数据研讨会上,产生了一个精简的元数据集——都柏林核心元素集 (Dublin core element set,DC)[5].

DC是国际通用的适用于网络资源描述著录的格式.它的结构简单,数据元素的含义清晰易懂,即使是非图书馆编目人员也能掌握.有德语、日语、葡萄牙语、西班牙语等 10多种语种的版本,可扩性好,可以与其他元数据连接使用.DC由 15个数据元素组成:题名、著者、主题及关键词、说明、出版者、其他责任者、出版日期、类型、格式、标识、来源、语言、相关资源、覆盖范围、版权.这 15个元素依据其描述的内容类型和范围可分为 3组:对资源内容的描述、对知识产权的描述、对外部属性的描述.在15个元素中,DC概括了电子信息的主要特征,如重要检索点、辅助检索点和关联检索.DC修饰词是对 15个元素的语义进行限定和修饰的词.它的制定遵循著名的 Dumb Down (向下兼容)原则,即修饰词的语义包含于未修饰词中.在范围上,对未修饰词的语义进行限定,在深度上对未修饰词的语义进行延伸.

DC在网络信息组织方面具有如下作用:1) DC可以直接处理网络数据.DC提供了全新的元数据定义,既是 DC的交换格式,也是元数据的内部处理格式,给数据处理带来极大的便利;2) DC是为网络资源的著录而制定的,适用于众多领域,同样很好地解决了数据变长、可重复问题.结构简单、易懂,自学就可以掌握.它的 15项核心定义可根据需要扩展,弹性好,又实用;3)它适用于世界上通用的软件成果,便于系统与时俱进,便于网络资源编目的自动化;4) DC著录格式简单,大大减轻了编目人员的劳动强度.在发展网络环境下的数字化信息系统中有广阔的应用前景;5) DC元数据是结构化的数据格式,它支持字段查询[6].

1.4 OAIS

1993年 12月,澳大利亚成立"面向 2001年保护澳大利亚数字信息调研组",其目标是制定数字信息存取和保护的指南[7]. 1994年 3月,欧洲保护与存取委员会 (European Commission Preservationand Access, ECPA)在荷兰首都阿姆斯特丹成立,其目的是发展与扶持欧洲各国图书馆、档案馆及相关组织间的协作,以确保各种格式的出版物和文档的长期保存,并促进人们对文化遗产的存取.同年 12月,ECPA与研究图书馆组 (the research libraries group)联合创立了数字归档特别工作组 (the task force on digital archiving),目的是"确保对未来以数字格式存储的文件的存取".1995年初,国际标准化组织 (ISO)为了开发其领域内的归档标准,授权空间数据系统咨询委员会 (The Consultative Committee for Space Data Systems, CCSDS)开发其领域内的归档标准,以支持空间领域数字信息的长期保存.CCSDS接受任务后,积极发动其会员机构着手制定空间领域数字信息长期保存的归档标准,并逐渐将该标准扩大到为政府、私企和学术界等组织的资源服务.经过 CCSDS各成员的不懈努力,2003年 2月 24日,国际标准 ISO14721:2003《空间数据和信息归档系

统——开放档案信息系统——参考模型》(space data and information transfer systems—open archival information system—reference model)终于诞生了. OAIS[8] (open archival information system)就是一个开放的档案馆,是由人和系统组成的有机体,其职责是为指定的社会群体保存信息并使之可以利用,具体包括 6方面的内容:1)与生产者谈判并接收恰当的信息; 2)对需要长期保存的信息取得充分的控制权;3)由自己或联合其他团体决定哪些群体应该成为指定用户,并且这些用户应该能够懂得 OAIS所提供的信息; 4)确保提供的信息对指定用户而言是可以独立理解的,也即是说,在没有信息创建人员的帮助之下,指定用户群能够理解信息; 5)遵循已制定的政策和程序,确保信息的保存不发生任何意外事故,并确保传播的信息是已授权的原作品的拷贝或可追溯到原作品;6)确保指定用户可以利用到保存的信息.

OAIS中的术语"Open"指的是这一参考模型以及将来相关的标准将在开放式论坛中不断地发展,而不是指档案的存取不受限制. 数字信息是 OAIS中信息的基本格式,但 OAIS不仅支持数字信息,同样也支持非数字信息[9].

2　体系架构设计

整个系统架构如图 1所示,可以看出,该结构层次清晰,而且削弱了模块之间的耦合度,更符合代码复用的规范.

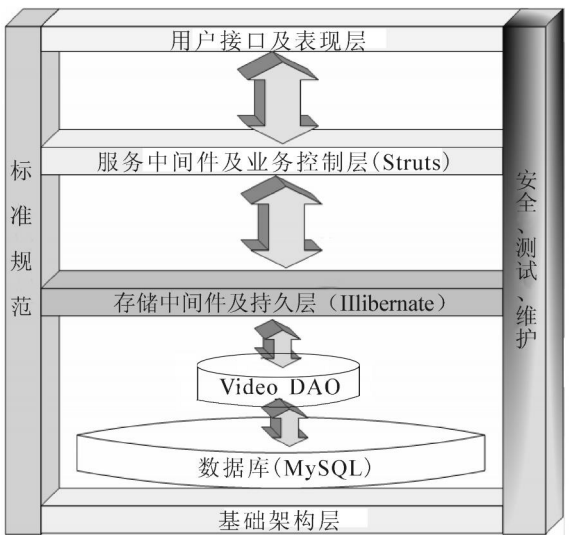


图 1　系统架构图

Fig 1　System architecture

从图 1可以看出,系统架构主要有 4层:基础架构层、存储中间件及持久层、服务中间件及业务控制层、用户接口及表现层.

基础架构层的上面是存储数据的数据库,它通过 DAO和存储中间件及持久层进行通信. 存储中间件及持久层通过 Hibernate和 Spring进行控制管理,形成一个完整的业务逻辑. 最上层是用户接口及表现层,用来将整个系统体现给用户并进行使用.

3　功能模块设计

3.1　功能模块设计

如图 2所示,系统用户可以根据标引模块对数据进行详细标引,提交标引存储至数据库,然后普通用户可以针对数据库进行搜索查询寻找自己有价值的数据,系统根据用户查询条件将有价值的数据返回给用户.

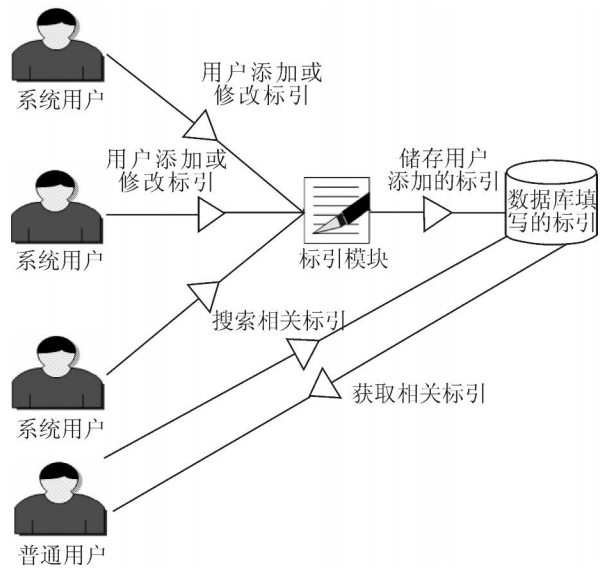


图 2　系统功能图示

Fig 2　System function

系统主要功能如下:

对于管理员来说,需要对视频数据进行详细标引并存入数据库,提供修改、删除功能,方便针对标引不完整或者标引出错的数据进行方便修改,对于垃圾数据进行删除以防止影响搜索结果.

从科研角度来讲由于本系统为自动标引加人工标引,而并不是单纯地人工标引,所以在很大程度上减轻了人工标引的工作量;并且更重要的是对于视频数据也更大程度上增加了其准确度,更方便用户在更短的时间内搜索到最有价值的数据.

3.2 具体功能模块

1)添加标引模块

如图 3所示,本模块中用户可以根据系统定义好的 DC对文件进行标引,其中 15项为 DC核心元数据,其余为扩展的元数据,按照数据分类为 22～33项不等.确认无误提交后存储到数据库.

图 3 添加标引页面

Fig 3 Add indexing

考虑到一般性和通用性,本系统中,添加标引的类型只有 3种:下拉菜单、指定路径和输入框.添加过程中自动定义日期以及时间格式是本模块的一大亮点,也是一大难点.

2)修改标引模块

本模块实现用户对于数据文件标引出现失误导致错误的情况下进行修正的功能.

3)查找标引模块

本模块实现用户对于已标引或者为标引的数据进行查找功能,此模块包含简单搜索以及高级搜索,简单搜索只针对于文件名以及文件内容所包含的数据进行搜索,高级搜索可以依据一些核心的元数据进行搜索.

4)删除标引模块

本模块实现针对一些重复数据以及个别错误数据进行删除操作.

5)自动标引模块

如图 4所示,本模块实现系统用户指定目标目录,针对指定目录中的视频文件进行批量扫描并提取部分元数据,存储所提取元数据至数据库.

图 4 自动提取视频文件元数据

Fig 4 Automatic extraction of video files metadata

4 标引的实现和应用

本系统使用 Java语言,采用 Eclipse、Mysql、Tomcat等工具开发.该系统采用 Jsp + Struts + Hibernate + Mysql的架构.

使用本系统,可以使用户将数据进行更详细的标引,供给用户填写和提交,并存入数据库,方便对于已标引的数据进行修改和删除,在搜索引擎方面可以更方便快捷地获得有价值的数据.系统的应用场景有很多,比如,在图书馆中,用户需要获得一本书,而这本书的书名有很多作者写过,而用户需要固定作者、固定出版日期的书.此时,使用本系统就可以很轻松地标引完书目,使得用户可以在最短的时间内找到所找书籍的位置.

系统界面的截图参见图 5.

图 5 查看标引页面

Fig 5 See indexing

5 结束语

随着互联网带宽的不断增大,视频文件越来越多地应用在互联网上,并且视频网站也是越来越多.而像这么多的视频数据想要搜索到自己有价值的数

据犹如大海捞针,因为视频文件不像文本文件,可以直接进行文本搜索就能得到.在视频领域基本很多都是自动提取元数据进行标引,但是这样带来的问题是:由于视频文件的种类很多,在自动标引的过程中很容易出现错误而导致用户搜索出的数据没有价值.所以采用自动提取加手动标引的方法,扩展 DC元素,对每个视频文件都进行更详细的标引,方便了更多用户使用更多元素在更短时间内搜索获得有价值的数据.本系统可以作为任何有视频标引需求的系统的子系统,能够为用户提供更方便快捷的服务.同时,系统中也存在一些有待进一步改进和增加的功能,例如,在自动提取中增加更多视频文件格式,对所标引的数据进行更详细分类等等.作者正逐步完善这些功能.

参考文献:

[1]中国互联网络发展状况统计报告[EB/OL].[2009-01-12]. http://www.cnnic.net.cn/index/0E/00/11/index.htm.

[2]李　晶.视频数据挖掘技术研究[J].今日湖北:理论版,2007,1(4):168-169.
LI Jing. Research of video data mining technologies[J]. Today Hubei: Theory Edition, 2007,1(4):168-169.

[3]林淑玲.浅析数据挖掘技术[J].科技信息:学术研究,2008(1):329,331.
LIN Shuling. Introduction of data mining technologies[J]. Technical Information, 2008(1):329,331.

[4]李文峰,刘雪涛,贾月琴.基于元数据标准的标准资源库建设研究[J].中国标准化,2007,(1):37-39.
LI Wenfeng, LIU Xuetao, JIA Yueqin. Research of the standard resource database construction based on metadata standard[J]. China Standardization, 2007(1):37-39.

[5]赵慧勤.网络信息资源组织——Dublin Core元数据[J].情报科学,2001,19(4):439-442.
ZHAO Huiqin. Organization of network information resource: Dublin core metadata[J]. Information Science, 2001,19(4):439-442.

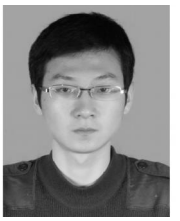
[6]何志兰.网络信息资源组织——Dublin Core[J].现代情报,2005(1):83-84.
HE Zhilan. Organization of network information resource: Dublin core[J]. Modern Information, 2005(1):83-84.

[7]颜晓栋.电子文件的长期保存研究[D].武汉:武汉大学,2004.
YAN Xiaodong. Long-term preservation research of electronic documents[D]. Wuhan: Wuhan University, 2004.

[8]Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS)[S]. BLUE BOOK, 2002.

[9]李明娟.OAIS参考模型与数字信息长期保存[J].图书情报知识,2007,119:65-69.
LI Mingjuan. OAIS reference model with the long-term preservation of digital information[J]. Document, Information & Knowledge, 2007,119:65-69.

作者简介:



张　博,男,1985年生,工程师,主要研究方向为海量数字资源存储和管理.

张　勇,男,1973年生,副教授,副研究员,主要研究方向为海量数字资源管理和服务、大规模并发事务处理等,发表学术论文 20余篇,其中被 EI检索 9篇,SCI检索 6篇.

朱　义,男,1972年生,高级工程师,主要研究方向为海量数字媒体管理、多媒体应用等,发表学术论文 2篇.

邢春晓,男,1967年生,教授,中国数字图书馆工程、全国文化共享工程专家组成员,中国计算机学会咨询工委副主任、软件工程和电子政务专业委员会委员,IEEE会员,主要研究方向为海量数字媒体管理、数字图书馆等.曾获软件著作权 2项,申请发明专利 1项,教育部科技成果 1项.发表的学术论文 40余篇被 SCI、EI、ISTP检索.