

●周宁 杨峰 刘玮

数字图书馆可视化接口方法探讨*

摘 要 数字图书馆系统中,利用可视化人机接口,可以向普通用户显现与提示图符、动画、图像等可视化图元,使用户获取信息与知识。可视化人机接口包括三方面内容:信息和知识资源对象描述的可视化,人机接口操作的可视化,操作结果提供的可视化。图6。参考文献4。

关键词 数字图书馆 信息资源 知识资源 可视化接口

分类号 G250.76

ABSTRACT In a system of digital library, we can use visual human-machine interfaces to provide users with visual primitives, such as icons, animations and images, and facilitate them to get information and knowledge. The authors think that visual human-machine interfaces includes three aspects, i. e. the visualization of the objects of information and knowledge resources, the visualization of the operation of human-machine interfaces and the visualization of the display of operation results. 6 figs. 4 refs.

KEY WORDS Digital library. Information resources. Knowledge resources. Visual interface.

CLASS NUMBER G250.76

信息资源主要包括三类:文本信息、语音信息和视频信息。给信息以形象和智能,这无疑会增大存储空间。对可视化信息的描述进行有效压缩是可视化存储的重要内容,它是降低成本、优化系统的有效方法。

1 数字图书馆信息资源的可视化

数字图书馆是知识经济社会的知识、信息中心。在因特网环境下,大量的信息、知识被数字化,存储在各式各样的数据库(文献数据库、事实数据库、数值数据库、多媒体数据库)、知识库、数据仓库和知识仓库中。用户要获取信息与知识,都要经过客户机的人机接口操作来实现。过去的计算机应用系统,人机接口通常用操作命令来实现。对用户培训,要求普通用户知道系统结构并掌握烦琐的操作命令格式是件不容易的事。它是系统推广应用的瓶颈。数字图书馆是面向全社会服务的,设计可视化接口是必然的选择。

信息可视化(Information Visualization)首先来自科学计算可视化(Visualization in Scientific Computing),它是运用计算机图形学和图像处理技术,将数据转换为图形或图像在屏幕上显示出来,并

进行交互处理的理论、方法和技术。它涉及到计算机图形学、图像处理、计算机辅助设计、计算机视觉及人机交互技术等多个领域。

数据可视化技术的主要特点是:交互性、多维性和可视性。信息可以用图像、曲线、二维图形、三维体和动画来显示,并可对其模式和相互关系进行可视化分析。数据可视化可以大大加快数据的处理速度,使时刻都在产生的海量数据得到有效利用;可以在人与数据、人与人之间实现图像通信。近年来,国际上对信息可视化专题进行了广泛而深入的研究,取得了一些重要进展。这些进展和研究成果的展示与交流体现在一些国际研讨会上及其论文集中。比较有影响的国际会议是IEEE所组织的两个系列国际研讨会:一个系列的国际研讨会是从1997年开始,在英国伦敦每年7月举办的“International Conference on Information Visualization”(如IV'97、IV'98、IV'99、IV2000、IV2001、2002和IV2003)。另一个系列的国际研讨会是从1995年开始,每年10月在美国的一些城市召开的“IEEE Symposium on Information Visualization”。这两个系列的国际研讨会集中体现了当代该领域的研究水平。

1.1 文本信息的可视化

* 本文系教育部人文社科重点研究基地重大项目“信息可视化与知识检索”研究成果之一。

文本信息大量存在于各类文献中。文献资源有它的外部特征(资源的创立者、资源名称、资源创立日期、资源类型等),还都有它的外部特征(资源内容、主题、分类、摘要等)。文献信息的可视化主要讨论文献内容的可视化。“一幅图胜过一千句话”,形象地说明了信息可视化(InfoVis)是描述信息资源的一种高效方法。

1.1.1 图符法

图符法是文本信息对象知识组织体系分类可视化的有效方法。我国的大多数人对文本信息资源的分类都采用《中国图书馆分类法》(CLC)进行分类。CLC有5个部分、22个大类、224个二级类目。它为文本信息资源的可视化勾画了一个蓝本和可操作平台^[1]。

对照《中图法》所述的分类(主题)领域,用相关图符表示其含义是一种简洁的可视化方法。大类均用特色鲜明的图符或动画表示,一目了然。二级类目既可以用相应的图符标注,也可以在大类图符下加分类代码标注。这使22个大类、224个二级类目都有个性鲜明的图符表示。

1.1.2 主题信息的可视化

文献主题的可视化用高位描述法实现,以《中国分类主题词表》(1994年版)为基础对中文文献主题内容进行可视化处理。

首先,选择相关的主题词(关键词或短语)对文献内容进行规范描述。一般地,用3~5个词对其标引。

然后,用这一组词对文本对象的各个部分(章节或自然段)进行词频统计。

第三,该组词的个数(m)和每个词在文本对象各个(k)部分的词频统计结果数据是一个 $k \cdot m$ 矩阵。该矩阵的参数,具体表述了高维空间模型。

中国
数字化可视人体
可视化人体
断面解剖
计算机三维重建

图1 高维空间描述法实例

第四,用彩条图表示高维空间描述的结果^[2](如图1所示)。一是用不同的颜色(红、橙、黄、绿、蓝、靛、紫)彩条表示不同的词。二是用不同的色彩深度(0~255)来表示词频的多少。

1.1.3 其他可视化方法

汉字文本信息的可视化还有一些方法:如模型图、映射图等;由统计情况而绘制的直方图、圆饼图、折线图等都是些有效的方法。

1.2 语音信息的可视化

语音信息资源是另一类大量的信息资源。汉语语音信息的可视化充分利用了文本信息可视化成果。其可视化过程分6个步骤实现:(1)面向具体汉语语音资源采样,用基本的语音样本建立相应的语音音素库。不同的用户有不同的音素库。(2)播放语音信息资源,实施语音音素分割。汉语语音信息资源来自于广播、电视、会议、谈话等各种活动而产生的录音带、录像带、光盘、唱片或现场交流等。(3)对照已建立的语音音素库进行语音识别。汉语语音的识别结果是形成相应的汉字序列(汉字字符串),同时在屏幕上显示。(4)将汉字文本(汉字编码)信息存储,生成汉字文本文件。(5)对已转化为文本的语音信息进行内容特征的抽取,按照《中国分类主题词表》进行规范。(6)内容特征的可视化(当然,它需要相应的图符库、词库和相关平台的支持)。

1.3 视频信息索引与可视化方案

视频信息数字化产生了系统数据流、视频数据流、音频数据流、同步数据流等,多种数据流合成而产生了视频文件。其压缩是对多种数据流分别压缩之后整合而成。对视频信息的索引是一项十分复杂的工作。

利用MPEG-7和相关的知识库可以抽取视频信息的特征值建立特征库。视频库中存放视频的数字化信息。根据MPEG-7中的多媒体描述方案MDS(Multimedia Description Schemes)分别对其中的视频、音频、文本信息特征进行描述。

例如,视频特征有颜色、形状、纹理、运动、位置特征,以及栅格分布、时间系列、多视点、空间坐标等。一个视频对象可以细分为镜头、场景,一段场景由关键帧、预测帧等系列组成。这些单位都可以作为抽取相关特征值的地址。

声频的特征包括:音频描述框架、声音效果描述工具、音质描述工具、语音内容描述、统一静态段和旋律描述符等。

文本信息特征有内容特征(知识元)、样式(字体、字号、颜色、字型位置关系)、排列位置等。

图、文、声的信息如何显现,时间轴同步信息如

何控制,数据编码、解码方法等都靠特征值描述。它与帧地址组合可生成相应的索引。

特征库的建立及其索引文件的生成成为视频文件的可视化提供了方便。由于视频文件本身就是可视形式,而且数据量巨大,直接用镜头中的关键帧作索引也不能准确地将其内容可视化。而将特征值和索引项聚类进行抽象化、概念化、知识化,用知识结构体系对其进行可视化较为方便。有了内容的知识化特征,可用类似于前面所述的文本信息、语音信息那样,对视频信息对象进行综合的可视化描述与存储。

1.4 信息资源可视化模型与压缩措施

综上所述,我国的信息资源可视化主要包括汉字文本信息的可视化、汉语语音信息的可视化和视频信息的知识化、可视化(其可视化的过程如图2所示)。图中指出了三个分支汇集一处。一个分支是直接对文本信息资源可视化,而第二个分支是将汉语语音信息转换为汉字文本信息来研究的,第三个分支是先在视频信息资源特征库和相关知识库的基础上建立索引,然后对其可视化。

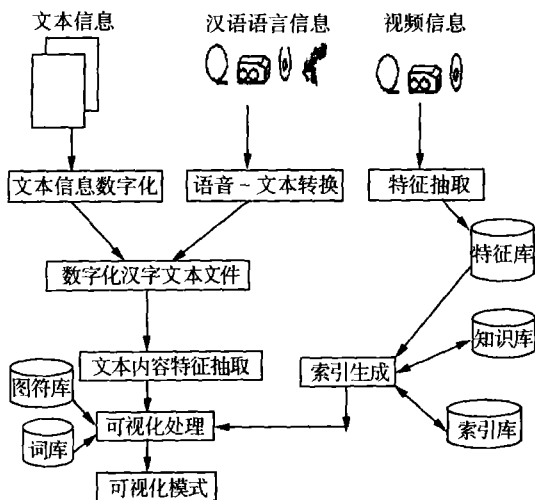


图2 中文信息资源可视化的一般过程

信息资源可视化模型由反映资源内容特征的图符、高维空间描述彩条图、特征库、知识组织体系和相应的数据压缩格式构成。给信息以形象,让数据有智能。

各类信息资源的可视化需要图符库、词库、特征库、索引库的支持。这样一来,又会增加一定的空间开销。但在实际应用中,这一部分空间开销不会成为额外负担。虽然图符库、词库、特征库和索引库占

据了一定的存储空间,但由于信息资源描述与存储采用了多种压缩方法(逻辑编码压缩, Lempel-Ziv-Waltch 信源编码压缩等),使原本描述所需要的信息空间大大减少。这样一来,它不仅可以抵消这一“额外”开销,而且可以使存储总量大幅度下降。因而它是一个优化的可视化模型方向。

2 数字图书馆可视化接口设计

数字图书馆的可视化接口由统一的界面操作。而系统内部有多个模块支持,主要有数据库资源利用可视化模块和 WWW 信息资源利用可视化模块。

2.1 数据库操作的可视化

数据库资源是数字图书馆利用的重要内容,无论是本地数据库还是远程的 Web 数据库都属于数据库检索的范围。为了叙述方便,我们以全文数据库为例来讨论数据库检索的基本情况。全文数据库文献描述与组织的可视化借助于图符库、词库(关键词库和自由词库)、索引库和可视化接口软件实现(如图3所示)。

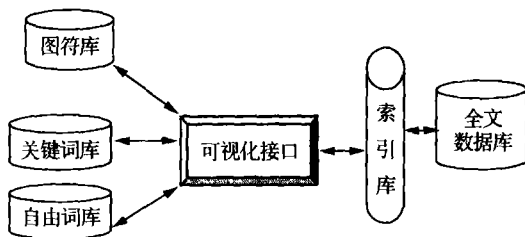


图3 全文数据库可视化系统构成

全文数据库采用《中图法》分类,关键词以中文为主加少量外文或(和)缩略语构成。对文献主题的概念化、知识化以《中图法》的22个大类和224个二级类目为准,设计图符或动画,建立专用图符库(Icon library),使之与分类号一一对应^[3]。

词库分两类:关键词库和自由词库。关键词直接从全文库相关字段中抽取,而自由词先由原文截词抽取,每个文献象抽出几百个不同的自由词,去掉非用词(无实际内容含义的虚词、冷词)进行自由词频率统计,选取该文献对象的频率高的20个自由词送入自由词库。

可视化界面接口软件用 Java 书写,设计的可视化方案的框架如图4所示。一篇文献的可视化框架分四个区:A,B,C,D。其中,A为《中图法》分类号所对应的图符。

B为用关键词对全文各部分出现频率的统计,然后生成分布图,用一个二维表格式中给出不同高度的直方图阵列表示。该关键词在哪部分频率最高,用满格直方图表示,其余部分根据所在部分频率与最高频率部分之比决定该部分直方图的相应高度。每个关键词的直方图作为一种颜色。这样各关键词在文献各部分的分布一目了然,内容分析结果显示图告知用户一个强烈的信号。

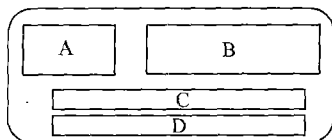


图4 可视化方案框架

C为关键词与相关文献链接区。根据该文献所标的关键词,查阅关键词库中最相关(该词出现频率最高)的10篇文献联结在一起供用户进一步链接查询用。

D区为自由词相关联络图区。自由词表是专用软件经过对原文的抽词统计,将其频率最高的20个自由词送来而构成的,记录格式如下:

Key	N	RNO
-----	---	-----

Key为自由词,N为词频,RNO为该篇文献的记录号。全文数据库中各篇文献的20个高频词送入自由词库后用(Key,N,RNO)组合字段排序,相同的自由词不归并。

有了图符库、关键词库和自由词库之后就可以用专用可视化软件对其进行可视化处理,按照图5所示的布局进行组装。其过程如下:

(1)读取全文数据库中一条记录。

(2)抽取分类号,根据分类索引在图符库中找到与之对应的惟一一个图符显示在A区。

	1	2	3	...	M
关键词1	N_{11}	N_{12}	N_{13}	...	N_{1m}
关键词2	N_{21}	N_{22}	N_{23}	...	N_{2m}
...	...	...	...	...	...
关键词k	N_{k1}	N_{k2}	N_{k3}	...	N_{km}

图5 内容分析图

(3)抽取关键字段,进行关键词分割,关键词逐个在原文中各部分分别进行词频统计,可显示该文献各部分的词频数,然后用二维表的分布图显示内容分析结果。假设文献中(抽取)有K个关键词,M

个部分,根据词频统计显示其K行M列的分析图。在实际操作中,我们发现个别关键词在原文中并不出现,不仅在正文中各部分均未出现,而且在文摘和标题中也未出现。为弥补这一缺憾,并扩大文献间的关联性,因而设计了自由词库。自由词库中的记录格式与关键词库中的格式完全一样,只是它的来源直接从原文中截词抽取。

(4)将关键词及其最关联的文献指针插入C区。

(5)首先,将该文献的10个高频自由词填入D区,然后搜索自由词库。对于D区中每个自由词,找出最相关的文献号指针链接上,一旦点击一个自由词,系统立即跳转链接上全文数据库中10篇文献的可视化界面,根据实际需要可以获得相关文献。

全文数据库检索的可视化只需填入相关的检索码值就可以了。在可视化检索接口中,对图符(或动画)、分析图和关联图进行交互,它仅用可视化界面的可视化元素(图符、动画、图表和超链接标识)点击实现,完全去掉了烦琐的检索命令和复杂格式。

2.2 WWW信息资源利用的可视化接口

因特网信息资源极为丰富,加之其资源的描述与组织不像数据库那样规范,因而其信息检索可视化接口就更加复杂。为了方便讨论,我们通过一个实例介绍来讨论WWW信息资源利用的可视化方法[4]。

VisWeb是WWW环境下对Web页面及其链接的可视化检索系统。它通过抽取页面的重要属性,如URL、网页摘要、超链接等利用形状、颜色、线条等视觉属性,来实现页面的图形化描述。用户可以就显示的形状进行动态调整。它实现了对国内使用较为普遍的天网、百度、搜狐三个独立搜索引擎检索命中结果及与命中结果有链接关系的页面的可视化显示。

VisWeb基于WebLogic服务器采用Java开发,界面简单直观。它改变了传统的搜索引擎的线性文字列表的结果显示方式;借鉴元搜索引擎的思路,能调用多个独立的搜索引擎。

图6显示的是以检索式XML作为检索需求由搜狐返回的检索结果。线性列表中的每个命中页面对应着可视化显示中的一个图形区域,并用颜色的深浅来区分。颜色深的代表命中结果,颜色浅的代表非命中文档。每个图形区域最上面的文字代表页面所在的最上一级网站的URL,用户可以点击图形

区域进入相应页面。当鼠标移向某页面对应的图形区域时,该页面的摘要文字内容会从窗口弹出,供用户阅读以便决定是否进入该页面进一步获取相关信息。

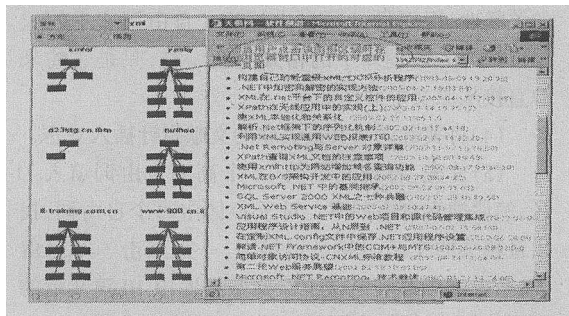


图6 点击图形区域打开的浏览器窗口

为了保证用户能获得较为全面的搜索结果,我们在传统的搜索引擎只显示命中结果的基础上将与命中页面有链接关系的相关页面也揭示出来,通过连线关联起来。与点击图形进入命中结果页面一样,点击相关页面所在图形也可以进入该页面。这样就可保证因用户检索需求的模糊性所导致的相关页面也可让用户访问到。这是传统的搜索引擎的线性列表方式所无法实现的。

系统提供了两种图形描述方式:长方形与椭圆形。用户可根据自己的兴趣,选择不同的显示方式。

在 VisWeb 可视化系统中,用可视化操作方法废除了命令方式,用户的输入部分简单易用,方便用户进行检索需求的表达。系统提供了三种搜索引擎供用户选择:天网、百度和搜狐。用户可根据不同的检索需求选择合适的搜索引擎。一般说来,天网适用于检索学术资料,百度与搜狐则适合搜索生活、娱乐等广泛的信息资源。系统提供相关词检索,可以帮助用户在搜索一个不太了解领域的资料时也能准确表达检索需求,得到满意的检索结果。

在这些主题中,所属的媒体有视频、文本、音频信息对象。抽取其特征值所建立相应的索引特征就可以实现信息检索可视化。这种基于内容的检索或知识检索不是一次操作能完成的,而是多次反复的渐进过程。经过多次反复检索,最后得到用户满意的检索结果。

3 结束语

信息可视化系统内部有多个并列的模块组成,它们分别处理数据库资源和 WWW 信息资源。在数

字图书馆可视化接口设计中,系统分别由处理文本信息、音频信息和视频信息的模块构成。该系统配备有图符库、词库、特征索引库及专用硬件、软件资源的支撑,利用关联图和信息结构图等辅助工具实现了多种信息资源提供的可视化。

参考文献

- 1 Zhou Ning, Wen Yanping, Liu Wei. On the Method of Information Resources Visualization. The Proceedings of Library — IT Opportunities and Challenges in the New Millennium. Beijing Library Press, 2002. 7
- 2 Hearst, M. User Interfaces and Visualization, in Modern Information Retrieval, R. Baeza-Yates and B. Ribeiro-Neto(eds), Harlow, England, Addison-Wesley, 1999
- 3 Robert Spence. Information Visualization. Addison Wesley, 2001. 5
- 4 周宁,文燕平,刘玮.文献信息可视化研究.情报学报 2003,22(4)

周宁 武汉大学信息资源研究中心教授、博士生导师。

通讯地址:武汉。邮编 430072。

杨峰 刘玮 武汉大学信息资源研究中心博士研究生。通讯地址同上。

(来稿时间:2004-01-05)

Elsevier(爱思唯尔)出版集团北京代表处揭牌

Elsevier(爱思唯尔)出版集团北京代表处于5月26日在北京举行挂牌典礼,并正式开始开展业务。Elsevier是提供科技和医学信息出版服务的国际著名出版集团。其北京代表处包括来自 Reed Elsevier 的科技部、医学部、法律和商务部等不同部门的人员,在中国开展市场销售及客户服务。此前十年,该公司就与国内出版社及清华大学等有着密切业务关系,为多所高校提供电子期刊全文数据库的服务支持。Elsevier 北京代表处的成立,将进一步扩大公司在中国的业务,更好地为中国的科学教育发展服务。