第23卷 第5期
2004年10月

情 报 学 报

Vol.23, No 5
October, 2004

信息可视化系统的RDV模型研究[1)]

周 宁 杨 峰

(武汉大学信息资源研究中心，武汉 430072)

摘要 信息可视化是信息管理和信息系统的热点研究问题。本文是在应用研究的基础上，从可视化问题的共性出发，建立了一个基于代数结构原理的信息可视化系统RDV模型。文章分析了RDV模型的三级结构，并对他们及其相互关系作了严格的数学描述，最后指出RDV模型的特点及适用性。

关键词 信息可视化 模型 RDV

Research on RDV Model of Information Visualization System

Zhou Ning and Yang Feng

(*Research Center of Information Resources, Wuhan University, Wuhan* 430072)

Abstract Information visualization is a hot research problem in the information management field. In this paper, a information visualization model RDV was presented which is based on the basic algebra principle after many application researches. The paper analyzed the structure of RDV, described the relations of RDV components mathematically, at last, analyzed the characters of the model.

Keywords information visualization, model, RDV.

20世纪90年代中期国际上提出了信息可视化问题。所谓信息可视化，就是利用计算机支撑的、交互的、对抽象数据的可视表示，来增强人们对这些非物理抽象信息的认知[1]。它是研究人、计算机表示的信息以及他们相互影响的技术；是人和信息之间的一种可视化界面，是人机交互技术的重要组成部分[2,3]。

可视化功能越来越成为信息系统的一个不可缺少的组成部分[4,5]。本文通过对可视化功能的一般意义的分析，建立一个可视化系统的一般性体系结构，并给出该体系的数学描述。该模型对建立可视化系统具有一般的指导意义。数学模型是运用数学的语言和工具，对部分现实世界的信息(现象、数据……)加以翻译、归纳的产物[6]。

1 可视化系统的体系结构

信息可视化按照信息系统处理过程包括信息描述与组织的可视化、过程可视化和结果可视化；按照信息本身的特征划分为7类：一维信息可视化、二维信息可视化、三维信息可视化、多维信息可视化、时间序列信息可视化、层次信息可视化和网络信息可视化[7]。

可视化可以看作是从数据到可视化形式再到人的对应过程。从原始数据到人，中间经历一系列数据变换。从这个一般的定性描述出发，我们把信息系统中完成可视化功能的部分称为可视化系统，即一个信息系统中完成可视化功能所涉及的数据及功

收稿日期：2004年1月2日

作者简介：**周宁**，1943年生，教授，博士生导师。**杨峰**，1968年生，副教授，博士生。

1) 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”(项目号：02JAZJD870004)成果之一。

能的全体。

原则上,能完成可视化要求的系统结构可以有多种不同的方式。但其基本功能都必须包括将信息系统中的数据经过一系列的转换,最后以可视化的形式表现出来。我们从系统的角度出发,可以认为可视化系统是由若干组成部分构成,并包括这些组成部分之间的相互关系。因此,基于这种观点可以给出一种可视化系统的体系模型——RDV 模型,其图示如下:

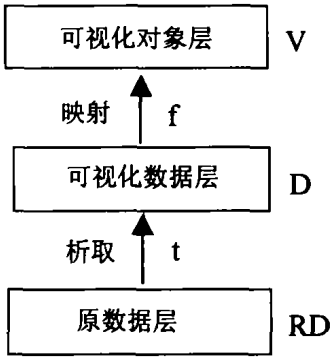

图 1 可视化系统示意图

RDV 模型分为三级结构、一个映射过程和一个数据析取过程。析取过程(t)把原始数据(raw data)中需要可视化的数据选取、转换到可视化数据层(data will be visualed),映射过程(f)把可视化数据对应到相应的可视化对象(visul obiects)完成全部可视化工作。

2 原数据空间

赋予了某些数学结构的集合称为空间或代数结构[8]。一个信息系统是一个复杂的系统,其可视化功能涉及到系统运行的全过程,即可视化系统是信息系统的一个子系统,信息系统是可视化系统的外部环境。因此,研究可视化问题必须从较完整的角度考虑,为此我们给出如下概念。

原数据,在未可视化之前,一个信息系统运行过程中可能出现的所有数据及其数据结构,记作 RD。如初始数据、结果、控制信息、用户交互信息等。

原数据空间,将信息系统中的各个程序单元看成是定义在原始数据集合上的运算,他们构成集合 P,则因为 RD、P 一般非空,称(RD,P)为原始数据空间。

在一定的技术条件或实际需求下,并不是所有的数据都适合或需要可视化的,或者有些是很难实现可视化的[9]。因此,需要对原数据空间的数据进行选择、转换。将那些原数据空间中需要可视化;且可以可视化的数据及数据间的关系转换出来,这是可视化后续工作的重要准备。

3 可视化数据空间

经过上述的准备工作,在一定条件下需要可视化的信息都已经十分清楚,我们称它为可视化数据。即可视化数据是一个信息系统中从原数据空间中选取、转换出来准备进行可视化表示的全部数据及数据间关系。

全部可视化数据构成一个集合,加上其上的操作,构成可视化数据空间。即设 D 是一个可视化数据的非空集合,Ω^D 是定义在 D 上运算的非空集合,称(D,Ω^D)为一个可视化数据空间。其中 D 称为(D,Ω^D)的定义域,|D|称为(D,Ω^D)的阶。显然(D,Ω^D)也是一有限空间。指定 nil 是 D 中的一个元素,当从原数据空间中没有选取到任何值时,D 中就只有该元素。

析取过程是连接原数据空间与可视化数据空间的关键,记做 t。

实现该过程的软件必须了解原始数据及其结构 RD,设计合理的可视化数据集合 D,D 的设计还要考虑到与其上层可视化对象的配合问题。对数值型数据来说,t 的常见功能有:直接选择、对总体特征进行统计、高维数据的降维转换、交互式信息的跟踪存储等。总之 t 是可视化系统完成数据预处理的软件包。

可视化数据 D 是一个数据集合,其数据之间存在相互关系,构成 D 的数据结构。需要注意的是 D 中的数据结构不等于经过选择的 RD 的数据结构。经过选择的 RD 的数据及数据结构在 D 中都以数据形式出现。他们是即将被可视化的,可视化的信息既有 RD 中具体意义的数据,也有抽象的数据间关系。他们在 RD 中是抽象的,但在进入 D 后必须是具体的,否则是不能进行可视化的。所以 t 所采用的方法必须是多样的,才能解决各种复杂的情况。针对多样的 RD,研究通用的可视化系统数据析取方法是一个有意义的工作。

4 可视化对象空间

每一个可视化数据必须有一个可视化的表示,

我们用可视化对象来表示他们,即可视化对象是准确描述可视化数据的图符(图像、图形、分析图和相关图等)。全体可视化对象构成一个集合。

所谓可视化对象空间可定义为,设V是一个关于可视化对象的非空集合,Ω^V是定义在V上运算的非空集合,称(V,Ω^V)为一个可视化对象空间。其中V称为(V,Ω^V)的定义域,$|V|$称为(V,Ω^V)的阶。显然$\Omega^V=\{\omega1,\omega2,\cdots,\omega m\}$为有限集合,即$|V|<\infty$,所以可视化对象空间是一有限空间,也可记为$(V,\omega1,\omega2,\cdots,\omega m)$。

考虑V上的一个特殊对象,在最一般的情况下,任何一个可视化区域都有一个底色,即不显示任何信息时可视化区域的图符状态,将其做为V的一个元素,记作e。

V中元素的实现有多种方法。如图像方法,一个元素是一个存储实体:图形方法,多个元素对应一个存储实体,存储实体中包含图符特征及绘图程序。

作为一个空间,需要定义在V上常用运算。这里,我们给出两种常用运算:

(1)对象合并,定义在V上的一个二元运算,记作*。如果有v1,v2,v∈V,且v的视觉效果与v1,v2的视觉效果叠加相同,则称v是v1,v2的对象合并。记做v=v1*v2。显然有v=e*v,v=v*e(v∈V)。

*运算的实现表现为系统所采用的图形、图像处理算法。研究*运算的目的是为寻找最少可视化对象种类。

(2)对象缩放,定义在V上的一个一元运算,记作I。一般地,可分为x方向的缩放I_x和y方向的缩放I_y。

如果有v1,v2∈V,且$v2=I_x(v1)$。称v2是v1沿x生成的。如果$I_x>1$,则称I_x是x放大;$I_x<1$,称I_x是x缩小;$I_x=1$,称I_x是x平凡缩放。

如果有v1,v2∈V,且$v2=I_y(v1)$。称v2是v1沿y生成的。类似地,可定义y放大、缩小、平凡缩放。

如果有v1,v2∈V,且$v2=I_x(I_y(v1))$。称v2是v1生成的。

一般地,如果v是图像形式存储,则从严格的图像意义上,随着I_x,I_y采用的具体算法不同,有时$I_x(I_y(v))\neq I_y(I_x(v))$。但不相同的只是很少的像素,从可视化对图符的要求精度上可以认为$I_x(I_y(v))=I_y(I_x(v))$。如果v是图形形式存储,则一般有$I_x(I_y(v))=I_y(I_x(v))$。

等比缩放:对于$I_x(I_y(v))$,如果有$I_x=I_y\neq1$,则称为等比缩放,记做$I_{xy}(v)$。如果有$I_x=I_y=1$,则称为平凡缩放,记做I(v)。

由于e的特殊性,所以规定$I_x(e)=I_y(e)=I_{xy}(e)=I(e)=e$。

研究缩放运算的目的是为选择最佳的时空效率。寻找在精度许可条件下同类可视化对象的最小集合。

对于运算*,I有以下性质:

性质1:*一般不是封闭的,即有v1,v2∈V,有可能v1*v2不属于V。

性质2:*满足交换率,即如果有v1,v2∈V,且v1*v2,v2*v1∈V,则有v1*v2=v2*v1。

性质3:I对*满足分配率,即如果有v1,v2,v1*v2∈V,且$I_{xy}(v1)$,$I_{xy}(v2)$,$I_{xy}(v1*v2)\in V$,则有$I_{xy}(v1*v2)=I_{xy}(v1)*I_{xy}(v2)$

5 可视化映射

可视化数据空间D的一个数据d,指定一个可视化对象空间V中的一个可视化对象v与之对应。记作,f:d→v。称v是d的像,d是v数。

对于任何f,定义e与nil对应,即e=f(nil)。

定理1:在可视化系统S中,从可视化数据空间D到可视化对象空间V上的对应f至少是一个满映射。

证明:①首先f是一个映射,即对每个d都有一个v与之对应。

否则的话,则至少有一个d没有任何v与之对应,即至少有一个d不能被可视化,这与S是可视化系统相矛盾。

②其次对每个v都有一个d与之对应。

否则的话,则至少有一个v没有任何d与之对应,即至少有一个v是从不被使用的,则应从S中去掉。所以,从①②可以得证定理1。

对于v=v1*v2,v1,v2,v∈V;且d1,d2,d∈D;当v1=f(d1),v2=f(d2)时有v=f(d),则称f是可视化数据空间D到可视化对象空间V上可视同构映射。称d是可视化意义下的d1与d2的合并,记做$d=d1\oplus d2$。对于任何的$\oplus$,定义$d=nil\oplus d=d\oplus nil$。这里"同构"一词借用于代数空间,但对其限制进行了弱化,并称之为"可视同构映射"。

定义1:若f是D到V上的一个可视同构映射,则称三元组(D,V,f)是一个可视映射系统。

对于一个可视化系统,由于允许其中D中元素

是不同质的,因此很难只用一个可视映射系统实现其全部映射部分。为此,我们通过分解 D 和 V 为若干个子集建立若干个可视映射系统完成映射部分的实现。

定义 2:对一个可视映射系统(D,V,f),如果 D 可分解为 $D_1 \cap D_2 \cap \cdots \cap D_n$,$D_i \cap D_j$ 不一定为空集;V 可分解为 $V_1 \cap V_2 \cap \cdots \cap V_n$,$V_i \cap V_j$ 不一定为空集;对不同的 D_i,V_i 分别可以实现 f_i,使得每一个(D_i,V_i,f_i)都是可视映射系统,则称(D,V,f)是可分解的。

定理 2:若有 $|D| = |V| = n > 2$,则三元组(D,V,f)总是可分解的。

证明:对于 D、V 来说,

因为总有 $nil \in D, e \in V$。

所以(D,V)至少可分解为 n 个(D_i,V_i),其中每个(D_i,V_i)只包含 $nil, d_i \in D_i; e, v_i = f(d_i) \in V_i$($i = 1, 2, \cdots, n$)。

因此,对任一个(D_i,V_i),直接选取 f 作为 f_i,有 e 对应 nil,v_i 对应 d_i。

则 $v_i = v_i * e, v_i, e \in V; d_i, nil \in D$;有 $e = f_i(nil)$,$v_i = f_i(d_i)$。

所以,任一个(D_i,V_i,f_i)是可视映射系统,即三元组(D,V,f)总是可分解的。

定理 2 的意义在于可以通过构造若干个标准的同构映射系统,完成可视化系统映射部分。为开发通用的可视化系统提供了理论支持。

基于以上分析可以得出一个可视化系统 RDV 模型的详细结构图:

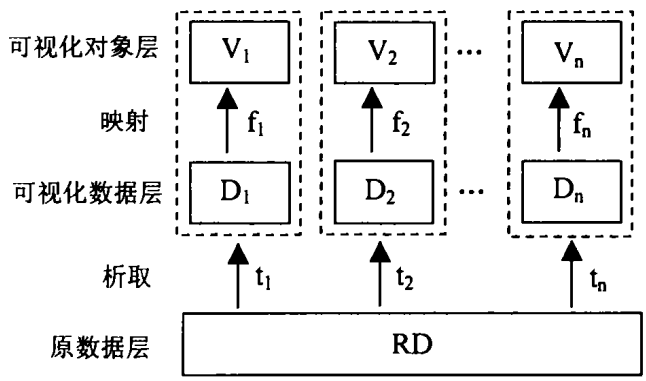

图 2 RDV 模型结构图

RDV 模型系统分为三层结构,即原始数据层、可视化数据层和可视化对象层。可视化数据层和可视化对象层又从纵向分为若干个可视映射系统。对于不同的可视映射系统(D_i,V_i,f_i)有相应的析取变换 t_i 与其对应,完成其可视化所需要的数据预处理工作。对每个可视映射系统(D_i,V_i,f_i)内部构造相应的可视化映射 f_i 完成可视化系统的可视化界面转换功能。

6 一个可视化实例的 RDV 模型描述

中文期刊全文数据库是我国知识基础设施工程之一,由于其语言及特色优势,深受国内广大用户的青睐。中文期刊全文数据库文献描述与组织的可视化,可使对其检索操作不再只有文本描述,而是由主题树结点图符,内容(分析)分布图,文献间的相关联系图等组成。

全文数据库是一种典型的一次文献库,对其可视化方法很多,我们实现了一种通用的可视化方法[10]。本文用其中的两个子系统,即知识概念的可视化(D_1,V_1,f_1)、文献对象内容分析的可视化(D_2,V_2,f_2),作为例子说明 RDV 模型实际构成。

系统的 RD 有:中文期刊全文数据库、《中图法》类目、用户表、查阅记录等。

(1)知识概念的图形化系统(D_1,V_1,f_1)

建立层次型数据结构,形成 D_1,用来储存可视化数据。用数据库技术实现时,对于 D_1 建立表:

编号	父结点编号	类目文字说明	一次文献号

其中:"编号"是层次结构中结点的唯一编号,也是表的关键字。全部结点的层次关系通过"编号"与"父结点编号"实现。对于根结点没有"父结点编号",对于叶子结点的"一次文献号"有具体值。

数据析取过程 t_1 完成转换工作,包括:识别《中图法》类目的层次结构,分析《中图法》类目编号,将类目文字说明填入相应的层次数据结构。

建立专用的可视化图符库(Icon library)数据结构,形成 V_1,用来存储可视化对象。对于 V_1 建立表:

编号	图符	对应类目编号

其中:"编号"是图符的唯一编号,也是关键字。"图符"字段用来存储可视化对象,这里采用压缩图像形式,如图 3。"对应类目编号"存储对应的 D_1 表的"编号",用来实现连接。

f_1 完成对应的维护工作,为了通用性和灵活性,f_1 包括图符的增加、删除以及调整 V_1 与 D_1 的对应关系等功能。图符与 D_1 中结点数目不一定一样,但

一般要多于其数目。其被使用的图符与 D_1 中结点一一对应。

P 天文学　　V 航空、航天

图3 图符与中图法类目对应实例

如果结点数据有了改变(如类目名称变化、类目分类调整等),则只需运行 f_1 完成对应关系的调整,形成新的对应关系;或者是增加了更合适的图符,也需要作调整工作。

如果是分类体系改变,即从 RD 中选取的不是《中图法》类目(如其他分类体系、自由分类、自动提取分类树等)。则无须改变 D_1 的结构,只需重新运行 t_1 完成 D_1 的生成。

(2)文献对象内容分析的可视化(D_2,V_2,f_2)

对于文献对象内容可以从多个角度分析。这里是用关键词对文献的各部分进行词频进行统计与分析,用相关的可视化对象描述内容分析的结果。

储存可视化数据的 D_2 包括表:

关键词表,

一次文献号	关键词号	关键词

关键词分布表

一次文献号	关键词号	部分1	部分2	…	部分M

数据析取工作分为直接从全文库相关字段中抽取关键词。用关键词对全文各部分出现的频率进行统计,生成分布数据。

这里 V_2 的实现不用图像方法,而是将关键词分布表做为图符特征,用绘图程序的形式实现可视化对象,f_2 完成参数传递工作等。系统运行时在屏幕上生成一个二维表格式的直方图阵列表示。关键词频率最高的部分,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长度。每个关键词的直方图为一种颜色。这样各关键词在文献各部分的分布一目了然。

当需要增强(系统扩展)可视化效果时,可重写绘图程序生成三维可视化图,或其他效果更好的可视化形式,而无须对系统的其他部分做任何改变。

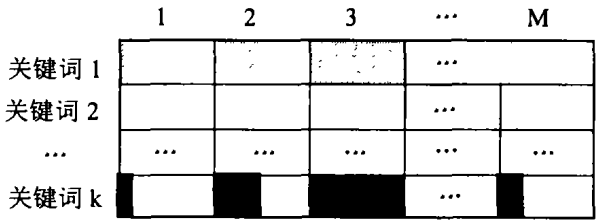

图4 内容分析可视化图

7 RDV模型的特点

对于可视化问题,可以从不同的角度建立不同类型的模型结构。RDV 模型是一种基于代数系统一般原理并结合可视化问题的特殊性建立的一种可视化系统的模型。它有以下一些特点:

(1)相对独立性

相对独立性包括3个方面:

1)(D_i,V_i,f_i)相对与 RD 的独立性。即 RD 的变化不引起(D_i,V_i,f_i)的改动,只需改变相应的析取变换 t_i。这使得(D_i,V_i,f_i)的开发可以独立于系统的其他部分,从而提高系统开发的速度;同时 RDV 的这一特点也为原不具备可视化功能的系统增加可视化功能提出了一个好方法,即不需要为增加可视化功能而重写系统。

2)(D_i,V_i,f_i)之间的相对独立性。不同的(D_i,V_i,f_i)之间完成实现不同要求的可视化功能,相互用不同的 t_i 完成同 RD 的联系。这使系统可以合作开发可视化功能,也使分步骤开发成为可能。例如,可以对数据的可视化意义较易描述的部分先开发,如检索结果可视化、图书库存可视化等;再开发可视化意义较为抽象的部分,如检索过程可视化、高维数据间的关系可视化等。

3)每个(D_i,V_i,f_i)内部 V_i 相对于 D_i 的独立性。即 D_i 的改变不会引起 V_i 的改变,而通过改变它们之间的可视化映射 f_i。这使得在设计 V_i 时能较多地考虑使用的需要。

(2)易扩展性

独立性使扩展性成为可能。首先是可以通过设计析取变换 t 使原来没有可视化功能的系统扩展可视化功能;也可以是在原有 n 个(D_i,V_i,f_i)的可视化系统上扩展第 n+1 个可视映射系统(D_{i+1},V_{i+1},f_{i+1});另一种扩展是扩展 V_i,随着系统的使用 V_i 中的某些图符需要改变、或对开始建立在较弱的软硬

件环境上的系统来说，其 V_i 中的图符精度需要提高、或是几个 D 共用一个 v 的情况需要分开。则 RDV 模型只需要以数据的形式调整 V_i 中的元素即可。

(3)易商品化

独立性和易扩展性是开发商品化软件的必要条件。用 RDV 模型建立的可视化系统是独立于信息系统的其他部分的，且具有良好的扩展性能，因此适合于商品化开发。

参 考 文 献

1 文燕平，周宁，杨峰.浏览界面可视化研究.信息可视化与知识管理——2003 信息化与信息资源管理学术研讨会论文选，2003.62～69

2 Zhou Ning, Wen Yanping, Liu Wei. On the Methods of Information Resources Visualization. In the Proceedings of Digital Library-IT Opportunities and Challenges in the New Millennium, 2002.7

3 Edward Condon, Bruce Golden, Shreevardhan Lele, S. Raghavan, Edward Wasil. A visualization model based on adjacency data. Decision Support Systems 33, (2002)

4 Robert Spence. Information Visualization. Addison Wesley. 2001.5

5 Chalmers, M., Ingram, R. and Pfranger, C. Adding Imagealibity Features to Information Displays. ACM Proceedings of UIST'96, 1996

6 姜启源.数学模型.北京：高等教育出版社，1987.4

7 刘玮，周宁，文燕平.信息可视化的分类研究.信息可视化与知识管理——2003 信息化与信息资源管理学术研讨会论文选，2003.85～92

8 王兵山，李舟军.抽象代数.国防科技大学出版社.2001.5～10

9 刘玮，周宁，赵丹.关于文献信息可视化的几点思考.信息可视化与知识管理——2003 信息化与信息资源管理学术研讨会论文选，2003.191～195

10 周宁，谷宏群，王洪艳.全文数据库可视化方法研究.信息可视化与知识管理——2003 信息化与信息资源管理学术研讨会论文选，2003.16～20

(责任编辑 芮国章)