档案图像的形式化描述与虚拟化解释

杨 有[1] 尚 晋[2]
(重庆师范大学数学与计算机科学学院 重庆 400047)[1]
(重庆电子职业技术学院计算机二系 重庆 400021)[2]

摘 要 档案是知识的载体,它不仅具有实态档案的具体属性,而且具有虚拟档案的概念属性,档案图像的形式化描述是档案图像这两重属性的一种体现,从档案的定义、档案的数码化过程和档案图像的处理三个方面,都可以对档案图像的形式化描述进行档案虚拟化解释。揭示档案图像形式化描述和虚拟化解释之间的潜在规律,对推进电子档案的检索利用具有重要意义。

关键词 档案图像,形式化描述,虚拟化

Formal Descriptions and Virtual Explanations of Document Image

YANG You[1] SHANG Jin[2]
(College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047)[1]
(The Second Department of Computer, Chongqing Electronic Profession College, Chongqing 400021)[2]

Abstract Document is the media of knowledge. It not only has the concrete attribute of entitative document, but also has the conceptual attribute of virtual document. The formal descriptions of document image are representations of these two attributes. From the perspectives of the definition, the digitalization and the processing of document, we can explain the formal descriptions by document virtualization. To discover the relationships between the formal descriptions and virtual explain, is benefit for the advancement of electronic document using.

Keywords Document image, Formal description, Virtualization

1 前言

计算机的普及以及网络的应用,使得档案利用者越来越希望档案部门能够提供全天候、全文、异地远程优质的信息服务,档案数字化在新的形势下凸显重要。数字化后的档案,即档案图像,如何利用数学工具进行形式化描述,如何利用数字图像处理技术进行加工,进而实现档案图像的高效检索利用成为摆在我们面前的课题。

在档案界,档案图像被认为是虚态档案[1],是档案的"数字"或"电子"形式。而在计算机界,档案图像被认为是包含了文字、图形、图片等区域并具有一定版面规范的图像,而图像是用各种观测系统以不同形式和手段观测客观世界而获得的,可以直接或间接作用于人眼并进而产生视知觉的实体[2]。从数学上看,我们可以利用数学工具对档案图像进行形式化描述,从形态上看,我们可以对档案数码化过程进行虚拟解释,从而将档案界的认识和计算机界的技术更好地结合,推动档案数码化工作的进程。基于此,本文探究了档案图像的形式化描述和虚拟化解释,以及它们之间的关系,分析并总结了档案图像处理的逻辑含义与物理意义。

2 档案图像的形式化描述

文[3]指出,档案图像包括档案概念结构和档案具体结构两部分。档案的概念结构由其概念空间决定,即

$$\omega=(\Sigma,\mathcal{F},\mathcal{O}) \tag{1}$$

其中 ω 代表档案的概念结构。Σ 表示非空的、有限集的字母表,其元素为符号或字母;$\mathcal{F}$ 代表字词的有限集合,其元素就是 Σ 上的一个有限符号序列;$\mathcal{O}$ 是算子的有限集合,其定义必须具有实际意义和语法意义,且满足 $\mathcal{F}\times\mathcal{O}\times\mathcal{F}\in\omega$。档案的具体结构由一个五元组来描述

$$\Omega=\begin{pmatrix}\Im\\ \Phi\\ \delta\\ \alpha\\ \beta\end{pmatrix}=\begin{pmatrix}\{\Theta^1,\Theta^2,\cdots,\Theta^n\}\\ \{\varphi_l,\varphi_r\}\\ \{\alpha^1,\alpha^2,\cdots,\alpha^p\}\\ \{\beta^1,\beta^2,\cdots,\beta^q\}\\ \Im\times\Phi\to 2^{\Im}\end{pmatrix} \tag{2}$$

其中,Ω 代表档案的具体结构;$\Im$ 是档案对象 Θ^i($i=1,2,\cdots,m$)的有限集合,且 $\Theta^i=\{\Theta_j\}^*$,$\{\Theta_j\}^*$ 表示对象的再划分,即一个对象可以再划分为几个更小的对象;Φ 表示连接因子的有限集合,用 φ_l,φ_r 分别表示左连接和右连接。δ 表示逻辑连接函数的有限集合,它指明档案对象的逻辑连接关系;α 表示标题对象的有限集合,$\alpha\subseteq\Im$;β 表示结尾对象的有限集合,$\beta\subseteq\Im$。

对于档案的具体结构 Ω,可进一步分解为几何结构和逻辑结构。其几何结构由档案具体结构 Ω 中的元素 $\Im$,以及作用于 $\Im$ 上的算子集合 β_U 两方面来描述,其中

$$\begin{aligned}&\Im=\{\Im_B,\Im_C\}\\ &\beta_U=\{\cup,\cap\}\\ &\forall i\neq j((\Im_i\cup\Im_j)\subseteq\Omega)\\ &\forall i\neq j((\Im_i\cap\Im_j)=\phi)\end{aligned} \tag{3}$$

杨 有 博士研究生,讲师,主要研究方向是数字图像处理、档案图像压缩。尚 晋 硕士研究生,副教授,主要研究方向是数字图像处理、软件测试。

$\Im_B$ 表示基本对象：$\Im_B=\{\Theta_j^i|\Theta^i\in\Im_C\}$，$\Im_C$ 表示复合对象：$\Im_C=\{\Theta^1,\Theta^2,\cdots,\Theta^m\}$。逻辑结构是人类对档案内容的理解，可由公式(2)中的$(\Phi,\delta,\alpha,\beta)^T$来描述。

另外，可以定义档案的结构强度[4]。假设一页文档可分为 n 个对象，对应着 n 个变量，用 H_i 表示第 i 个变量的偏熵(partial entropy)，H 表示整个文档的熵，则档案的结构强度定义为：

$$S_S=\sum_{i=1}^{n}H_i-H \tag{4}$$

比如，某档案包含 4 个复合对象，对应的变量为 x_1，x_2，x_3，x_4。则该档案的结构强度为

$$\begin{aligned}S_S&=\sum_{i=1}^{n}H_i-H\\&=\sum_{i=1}^{4}\Big[-\sum_{j=1}^{n}p_j(x)\log\,p_j(x)\Big]-\Big\{-\Big[\sum_{j=1}^{n}p_j(x_1,x_2,x_3,x_4)\log\,p_j(x_1,x_2,x_3,x_4)\Big]\Big\}\\&=-\sum_{i=1}^{4}\sum_{j=1}^{n}p_j(x_i)\log\,p_j(x_i)+\sum_{j=1}^{n}p_j(x_1,x_2,x_3,x_4)\log\,p_j(x_1,x_2,x_3,x_4)\end{aligned}$$

3 档案图像的虚拟化解释

3.1 档案定义的解释

档案是知识的载体，它表达了人类的思想，这种思想广泛涉及政治、经济、历史、文化、教育、艺术、科学、工程等[3]。其中“思想”指出了档案的概念性含义，对应于公式(1)定义的概念属性，而“载体”暗示了档案是一具体的二维图像，对应于公式(2)的具体属性。这种定义与文[5]的定义不谋而和，即此处的“思想”就是指档案的历史联系，此处的“载体”就是指文件实体集合，并且“思想”与“载体”相辅相成，二者不能独立构成档案，它们的有机结合是现代档案学的逻辑起点。当人们将自己的思想以符号形式书写时，概念结构就经过人脑编码成了具体结构，即 $\omega\to\Omega$；当人们对档案图像进行处理时，具体结构就解码成了概念结构，即 $\Omega\to\omega$。

按照公式(1)定义的概念档案，具有语言不变性的重要性质，即$\frac{\partial\omega}{\partial\sum}=0$。它意味着当档案从一种语言转换为另外一种语言时，其概念结构保持不变。比如，当档案从英文翻译为中文时，档案的思想是相同的。

3.2 数码化过程的解释

档案数码化过程是档案具体结构 Ω 形成的过程，它可视为档案的一种数字化虚拟。数字化虚拟来源于虚拟的本义——“假定”、“设想”。“数字化虚拟”不是假设了档案，而是假设了数码符号并把它法定化。由这种法定化符号表达出来的档案信息就与假设无关了。

在档案的虚拟化过程中，由于信息载体发生了变化，引起信息与载体的关系也发生变化，即信息对载体的依附性减弱，信息的流动性和相对独立性加强。人们随之把注意的重心从物理载体逐步移向载体上的信息，从物理管理为主转化为逻辑管理为主，强调信息在不同空间位置和不同物理载体之间的永恒的流动性和信息的相对独立性。因此，档案虚拟的本质意义是档案信息的独立与自由。

档案虚拟包括广义的虚拟和狭义的虚拟。广义的虚拟是指规则文明或符号文明，是人类对各种规则的合成、选择及其演化。狭义的虚拟是指当代的数字化的表达方式和构成方式、实践方式和创造方式。

3.3 档案图像处理的解释

档案处理分为两个阶段：档案分析和档案理解。从原始档案中提取几何结构的过程对应档案分析，将几何结构映射为逻辑结构的过程称为档案理解。因此，档案处理就是构建等式(2)和(3)表示的五元组的过程。档案分析就是提取等式(3)中的元素 $\Im$、Θ^i 和 Θ_j^i 的过程，即 Ω 的几何结构。档案理解就是发现等式(3)中 Φ、δ、α 和 β 的过程，即 Ω 的逻辑结构。逻辑结构是人类对档案内容的理解，一旦得到档案的逻辑结构，就意味着档案可以采用人工智能技术或其它技术进行解码。

对于等式(4)定义的档案结构强度 S_S，它代表各对象熵之和与整幅档案图像熵的差，即各对象所包含信息量之和与整幅档案图像信息量之差，因此，S_S 的值越大，表明档案中各对象之间包含的信息量重叠较多，即对象之间的“思想”相关度较高，则档案图像的结构强度就越高；反之，S_S 的值越小，表明档案中各对象之间的独立性越强，即对象之间的“思想”相关度较低，则档案图像的结构强度就越低。

结论 档案图像是实态档案的虚拟化，它提供的信息具有极大的空间流动性，信息能以极高的速度、极大的容量、跨越广阔的地域自由流动，为全社会共享，这是传统档案绝对做不到的。从档案信息与载体的“二位一体”到信息的独立与自由，是档案虚拟化的最具有本质意义的变化。从信息利用的角度来看，档案图像还提供了利用的共享性、利用的复用性、利用的交互性和利用的灵活性。

参 考 文 献

1 丁海斌. 档案虚拟论[J]. 档案学通讯，2004(2)：25～28

2 章毓晋. 中国图像工程. 中国图像图形学报[J]，1995,1(1)：78～83

3 Tang Y Y, Cheriet C D Y, et al. Automatic Analysis and Understanding of Documents. Handbook of Pattern Recognition and Computer Vision. World Scientific. Singapor, 1993. 625～654

4 Watanabe S. Pattern Recognition, Human and Mechanical[M]. Wiley Interscience, New York, 1985

5 冯湘君，刘新安. 现代档案学理论的逻辑起点[J]. 浙江档案，2005(7)：7～9

(上接第 192 页)

5 余建祖，苏楠. 混沌时序的噪声降低技术研究. 航空学报，1999，20(6)：498～502

6 黄显高，徐健学，何岱海，等. 利用小波多尺度分解算法实现混沌系统的噪声缩减. 物理学报，1999，48(10)：1810～1816

7 马丽萍，石炎福，余华瑞. 含噪声混沌信号的小波去噪方法研究. 信号处理，2002，18(1)：83～87

8 Lorenz E N. Deterministic Nonperiodic Flow. J. Atmos. Sci, 1963,20:130～141

9 Mallat S. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Trans on PAMI, 1989, 11(7):674～693

10 Mallat S, Hwang WenLian. Singularity detection and processing with wavelets. IEEE Transaction on Information Theory, 1992, 38(2): 617～643

11 Daubechies I. Orthonormal Bases of Compactly Supported Wavelets. Commun. on Pure and Appl. Math, 1988, XLI:909～996