HDF 5格式特点及其对遥感数据格式标准化的几点启示

王永韬,刘良明
(武汉大学遥感信息工程学院,武汉 430079)

摘要:阐述了HDF 5数据格式的特点及其对遥感数据标准化的启示和借鉴作用。HDF 5格式层次式的逻辑结构、B树的物理存储方式,面向对象的特性、数据类型的广泛支持、自我描述以及内容与表达的分离等特点,使得它在记录和存储科学数据时具有强大的优势。遥感数据标准化如何设计一种结构简单且扩展性很好,同时支持不同平台的标准格式,可以从HDF 5的实现方法上得到借鉴。

关键词:HDF 5;层次结构;B树存储;跨平台;遥感数据标准化
中图分类号:TP 311.12 **文献标识码**:A **文章编号**:1001-070X(2005)03-0039-05

0 引言

遥感科学自20世纪60年代兴起以来,经过50余年的发展,在应用领域、观测技术和理论算法上均取得了巨大的进步[1]。可以看到,在GIS领域,数据格式标准化已经付诸实践,OpenGIS、SDTS和DLG/F等方案不断被提出[2],同时,遥感数据具有数据量大、元数据信息丰富、现存数据格式多样化等特点,其数据标准化工程具有较大的挑战性,因此,需要结合其它学科的经验推出相应的解决方案。

HDF(Hierarchical Data Format层次式文件格式)是美国国家计算中心推出的一种新型数据格式,其目的用于记录科学数据。美国国家宇航局(NASA)在HDF的基础上提出了HDF—EOS子集,用于记录MODIS传感器数据。随着MODIS数据在国内遥感研究领域的大量应用,HDF格式也逐渐广为人知。作者通过对HDF文件结构的深入研究,认为其在文件结构和设计上的优点,对于遥感数据格式标准化工程有很好的几点启示作用。

1 HDF 5格式特点及优势

总体上讲,HDF格式是一种具有自我描述性、可扩展性、自我组织性的可用于绝大多数科学研究的数据存储格式[3]。同时,对比通常的文件头—数据体这种物理结构的文件格式,HDF的物理结构更为复杂,HDF 4采用了类似TIFF的tag方式分块建立文件内容的"索引",而HDF 5则是采用了二叉树的方式建立文件内容的"索引",通过"索引"可以方便快捷的访问数据内容。

因为HDF 4以及更早一些的HDF文件格式版本相对HDF 5的结构存在一定的不足,因此这里主要介绍HDF 5格式的特点。

HDF的全称为层次式文件格式,层次式是HDF逻辑结构的核心思想。如同关系数据库利用二维表和记录来建立实体之间的逻辑关系一样,HDF通过层次式的方式,有效地建立了文件内对象之间的逻辑包含关系和组织方式,这种方式有些类似XML的逻辑表达,如图1所示。

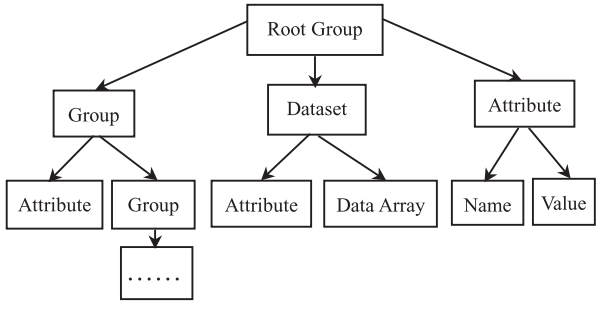

图1 HDF 5文件逻辑结构

HDF 5有两种基本对象——组(Group)和数据集(Dataset),同时有其它的辅助对象类型即数据类型(Datatype)、数据空间(Dataspace)和属性(Attribute)。

收稿日期:2004-12-29;**修订日期**:2005-05-16

其中,组可以看作一个容器,包含任意数量的其它组和数据集。在 HDF 5 的逻辑结构里,整个文件作为一个 Root 组存在,所有内容均为 Root 组的成员。HDF 5 的组织结构参考了 UNIX 的文件结构组成,其中组(group)对应于 UNIX 文件系统中的目录,根目录("\")为最基层目录,"."目录表示当前目录。组(group)中存在软连接(soft link)和硬连接(hard link),分别代表建立对其它对象引用的不同方式。上述概念与 Windows 中的文件结构也有些类似,可以把组作为一个文件夹,数据集作为一个数据文件,文件夹里面可以包含数据文件,也可以包含其它的组对象,这样层层嵌套下去,从而形成一个复杂的数据对象,并最终形成一个根组(root group)即 HDF 文件。数据集是一个可以存储任意类型数据的多维数组,它包含有两个属性,数据类型定义数据集所包含数据的类型,数据空间定义数据集的维数信息。

HDF 5 与文本方式表达数据信息的 XML 格式非常相似,二者均只是提供了一个层次式结构框架。XML 通过层次式的文件结构构造出丰富的数据类型,通过自我描述的方式实现了跨平台。

1.1 HDF 5 格式的特点

1.1.1 层次式表达

HDF 通过层次式逻辑结构来表现文件中不同数据元素之间的逻辑关系,如包含、被包含、并列等等,对其中的内容进行了有效的逻辑组织。

在层次式的框架下,数据的内容不加限制,例如,利用 HDF 存储一幅全国归一化植被指数(NDVI)数据时,可以采用如图 2 所示的结构来表达。

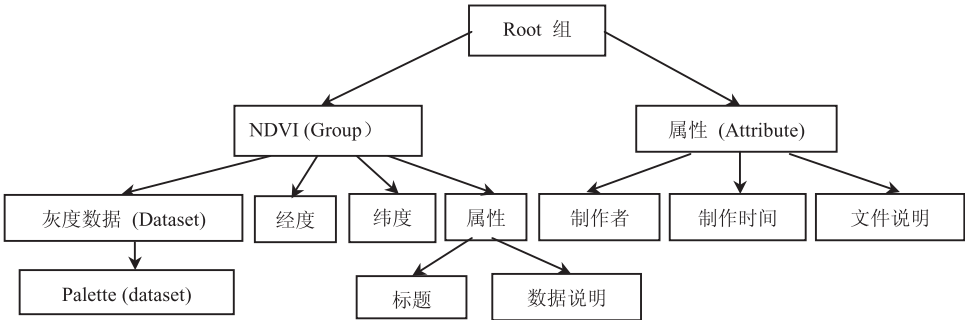

图 2 NDVI 数据结构

美国宇航局(NASA)在 EOS—HDF 应用中所采用的格式如图 3 所示。

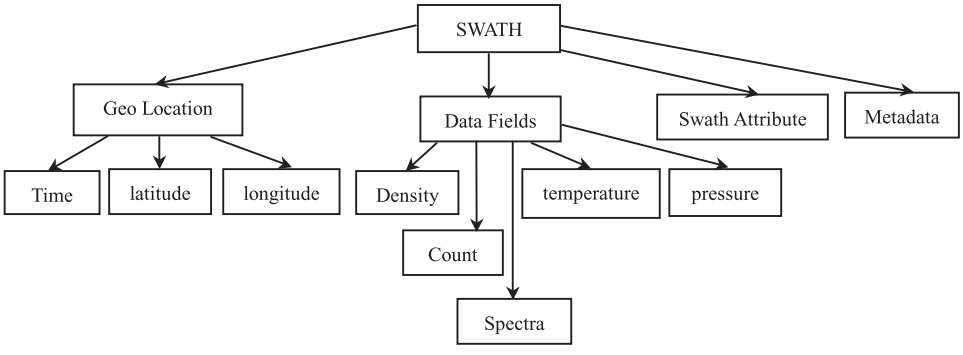

图 3 HDF—EOS 数据结构

可以看到,HDF 只是提供了一种机制,实际应用时可根据需要放置数据。层次式为用户表达提供了最大的灵活性,这是 HDF 5 和传统遥感数据格式如 TIFF、BMP、JPEG 最大的不同之处。

1.1.2 B 树存储

HDF 5 文件的物理结构采用了 B 树的方式,这是相对于 HDF 4 及以前版本最大的改动。其文件的物理结构如图 4[4] 所示。

HDF 5 文件由一个超级块(super block)、众多的 B 树节点(B - tree node)和空闲区(free space)组成。其中超级块在文件的头部,内容包括格式判别符、版本号、B 树叶节点容量以及 B 树根节点地址;而每个 B 树节点存储一个 HDF 5 对象,内容包括对象的类型,如组、数据集、数据类型、数据空间等,对象的属性和内容,以及相应的子节点地址。通过 B 树的层次性将逻辑上对象之间的层次关系表达出来。例如图 4 中,根节点存储了根组(root group)对象的信息,同时,在成员列表(group table entires)中存储根组

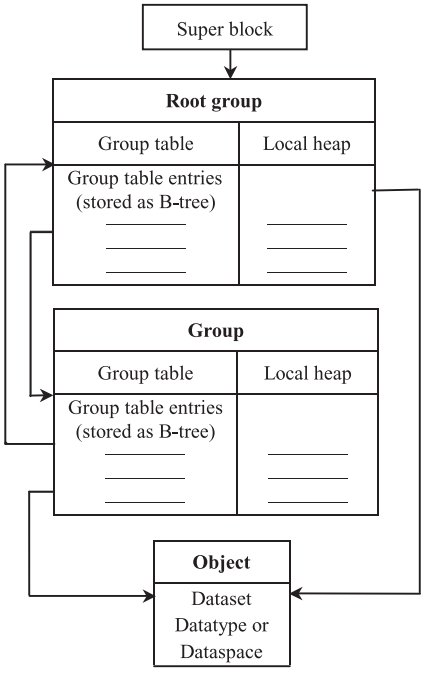

图 4　HDF 5 文件物理结构

(root group)中所包含成员节点的地址,而 group 节点中在 group table 中存储了其子节点的地址,可以是数据集、数据类型或者数据空间对象。

文件中的基本元素 Group\dataset\datatype\dataspace 作为对象(Object),以 B 树节点的方式存储在文件中,通过 B 树存储的方式,借用关系数据库中成熟的信息检索思想,允许对象分散灵活存储而又能高速访问。

1.1.3　面向对象

HDF 5 在逻辑结构上主要对象包括组和数据集,其它辅助对象包括数据类型、数据空间和属性,HDF 5 采用面向对象的优点在于对象的重用。例如,对于某一景 MODIS 数据,空间分辨率为 250 m 的两个波段可以采用同样的数据类型和数据空间,对于 500 m 的 5 个波段同样如此。HDF 5 文件中的对象在需要的时候均可以被重用,利用 soft link 和 hard link 来实现。二者类似编程语言中的指针,通过记录对象的 id 来建立对对象的引用。

重用机制或者说面向对象的特性使得 HDF 5 能够有效地降低数据的冗余度。同时,通过基本的对象元素可以构造结构复杂的数据。这是 HDF 5 相对传统影像文件例如 JPEG、TIFF、GIF、BMP 等的进一步完善。

1.2　HDF 5 格式的优势

1.2.1　数据类型丰富

前面已经提到数据类型是 HDF 5 中的一个对象,通过定义这个对象,用户在 HDF 中除了可以使用基本的数据类型(例如整型、浮点、双字节等)来保存数据以外,也可以使用复合类型(类似结构体),以及自定义基本数据类型来保存数据内容,以适应某些复杂的应用。

例如,对于一个复杂的气象学标量元素,其中可能包括三维坐标(浮点型),其对应的大气压强(整型)以及温度(浮点型),对这样的数据元素,普通的数据类型是难以表达的,这时就可以按照上述的顺序自定义一种复合类型(compound type)。对于复合类型,其成员本身也可以是复合类型的,由此构成任意复杂结构的层次式数据类型。同时,对于常见的数据类型例如整型、浮点型等等也可以加以定制,例如,当常见的 16 位整型不能满足需求时,可以自定义 20 位的整型类型以存储更大范围的整型数值。

TIFF/BMP/JPEG 等图像数据格式,在数据类型的兼容上比较单一,对于 BMP/JPEG 只对灰度影像(整型数值范围: 0 ~255)支持,TIFF 同时兼容了浮点型,对于复杂的遥感数据,上述数据类型仍是不全面的。

1.2.2　自我描述

在 HDF 5 中,每一个数据集对象都有相应的数据类型和数据空间属性,通过这两个对象,数据集在文件中记录了自己的数据类型和数据维数信息。因此,即便 HDF 5 文件由不同的平台不同的软件创建和修改,数据内容的统一性依然得到保持,通过这种机制 HDF 5 实现了数据的自我描述。

自我描述的特性使得 HDF 5 跨平台使用成为可能,同时保证了 HDF 5 文件传输中的独立性和归档保存的方便性。这一特性与 XML 类似,这也是 XML 现在被广泛使用的原因之一。

1.2.3　数据内容与表现的分离

遥感数据包罗万象,其格式也是非简单的彩色或黑白影像可以完全表达出来的,传统的 GeoTIFF/TIFF/BMP 等文件格式主要是将数据用影像的方式表现出来,但是,随着遥感技术的发展,数据不仅仅是辐射反射信息,同时兼有复杂的属性信息,这些都不是传统的影像文件所能保存和表达的。

HDF 5 将内容和表现形式分离。HDF 5 专注于记录数据内容,而将数据的表现交由用户自身处理,因为不同的用户需求不同。配合调色版的影像表达只是 HDF 5 的一种方式。当遥感数据并不仅仅局限于二维时,HDF 5 仍然可以高效的对这些内容进行存储管理。

对简单影像,HDF 5 使用数据集记录对应的像

素灰度值,同时记录对应的调色板(palette),对于复杂的数据,例如高光谱影像、雷达影像等,用户根据自身的应用需求对内容进行表达。

2 遥感数据标准化

目前,由于数据来源和软件平台不同,存在大量互不兼容的遥感数据格式,由此造成了大量的数据转换工作,其中很多转换属于有损转换,即经过转换原始数据中部分信息丢失。同时,数据共享的需求也日益突出,而共享的前提是格式统一。因此,遥感数据标准化显得非常重要。

通过上述对 HDF 5 介绍,对于遥感数据标准化有以下启示。

2.1 结构简单扩展性较强

作为标准,必须涵盖整个遥感学科,从陆地遥感、大气遥感到海洋遥感等,包括了大量不同层次不同应用的用户,如果标准结构太过复杂,用户使用的门槛过高,从而降低了用户使用的兴趣。同时,格式扩展性要强,因为作为标准格式,必须具有一定的生命力才有存在的意义。随着学科的发展,应用会越来越复杂,对数据格式的表达能力要求越来越高,如果标准格式的扩展性不够强,那么面临着随时被淘汰的可能性,也就失去了作为标准的价值。

参考 HDF 5 的格式,逻辑上采用层次式结构,虽然结构简单但是扩展性很强。用户在符合层次结构的总前提下,可以充分地构建满足自身应用需求的局部结构。传统的影像格式文件,文件内部元素大部分属于线性关系,不利于扩展。HDF 5 的层次式结构是一个很好的参考。

2.2 跨平台支持

跨平台支持是遥感数据标准化的首要作用。这里的跨平台包括不同软件平台和不同硬件平台,其中软件平台包括操作系统和应用软件。由于不同的操作系统(windows、linux 和 unix 等)在文件编码和数据存储方面存在一定的差异,因此要求标准的遥感数据格式能够支持目前存在的所有操作系统。

HDF 5 目前可以很好地在主流与非主流操作系统上应用,它所支持的操作系统包括 windows、linux、unix 等主流操作系统。在数据类型上,通过构建一个中间过渡类型的数据格式,实现了各种数据格式在不同平台的统一,实现了完全的跨平台支持。

2.3 技术支持

作为 HDF 文件格式的发源地美国国家超级计算应用中心 NCSA 在推出 HDF 格式的同时,向用户和软件厂商免费提供了 HDF 的 API(应用程序接口)库和针对不同级别用户的文档,结合 API 库和开发文档,开发者可以方便快捷地对 HDF 文件进行读写,获取文件内容、创建新文件、修改文件内容等,开发基于 HDF 的软件产品; 普通用户借助用户手册可以充分了解 HDF 5 性能,借助成熟的相关软件在自己的研究中使用。正是有了权威的支持,加之优越的性能,HDF 5 才会在多项世界级的项目中被应用,其中包括美国 NASA 的地球观测系统(Earth Observing System, EOS) MODIS 计划,美国数字图书馆技术(Digital Library Technology, DLT)项目,GLOBUS 网格技术[5]。

标准化不仅仅是推出一个数据格式,格式背后不同级别用户的支持至关重要,因为这关系到标准化的推广和使用,孤立的标准格式没有任何价值。同时需涵盖不同级别的用户,从高级开发者到普通用户,都能得到所需要的技术支持。

3 展望

遥感数据标准化是一个系统工程,并不仅仅是推出一个统一的数据格式这么简单。HDF 5 是美国国家超级计算应用中心 NCSA 为存储和传输科学数据而专门设计的数据格式,其中对跨平台和复杂数据类型的良好支持,使其被美国 NASA 作为 EOS 系统的格式,通过研究可以看出,它的几个特性对于遥感数据标准化具有很好地启示。同时,必须意识到遥感学科必须紧跟其它学科的步伐,意识到标准化的重要意义,同时加快标准化的进程。

参考文献

[1] 陈述彭. 遥感应用与数字地球[A]. 路甬祥. 中国科学进展[C]. 北京:科学出版社,2003.

[2] 钟耳顺,王康弘,宋关福,等. GIS 多源数据集成模式评述[A]. 99'中国 GIS 年会论文集[C]. 深圳:1999.

[3] 刘玉洁,杨忠东. MODIS 遥感信息处理原理与算法[M]. 北京:科学出版社,2001.

[4] HDF 5 File Format Specification[EB/OL]. http://hdf.ncsa.uiuc.edu/HDF5/doc/H5.format.html.

[5] HDF 5 User Guide[EB/OL]. http://hdf.ncsa.uiuc.edu.

CHARACTERISTICS OF HDF 5 FORMAT AND THEIR REFERENCE VALUE TO THE STANDARDIZATION OF REMOTE SENSING DATA

WANG Yong – tao, LIU Liang – ming
(*School of Information Engineering on Remote sensing*, *Wuhan University*, *Wuhan* 430079, *China*)

Abstract: This paper describes in brief the capabilities and design of HDF 5 (Hierarchy Data Format), such as its structure which is hierarchical in logic and B – Tree in physics, its advantage in straightforward implementation and self – description means of sharing science data among people, projects and types of computers, and its object – oriented specificity. The standardization of remote sensing data can benefit from these characteristics.

Key words: HDF 5; Hierarchy; B – tree; Cross – platform; Standardization of remote sensing data

第一作者简介: 王永韬(1978 –),男,硕士,研究方向: 遥感数据格式及其标准化,网络地理信息系统,遥感图像处理系统实现。

(责任编辑: 周树英)

(上接第 22 页)

从以上两种试验结果可以得出,Givens 算法在一至五次多项式模型中都很稳定,并且纠正精度也很好; Householder 方法运算不稳定,一至三次误差小,但四、五次误差较大; Gauss 和 PCI 方法运算稳定,但是 *RMS* 误差较大,导致纠正精度较差。所以,采用 Givens 方法能有效提高图像纠正精度。

参考文献

[1] 赵英时,等. 遥感应用分析原理与方法[M]. 北京:科学出版社,2003.
[2] 徐 涛. 数值计算方法[M]. 吉林:吉林科学技术出版社,1998.
[3] 徐树方. 矩阵计算的理论与方法[M]. 北京:北京大学出版社,1994.

A COMPARISON OF REMOTE SENSING IMAGE RECTIFICATION EFFECTS BASED ON SEVERAL MATRIX ALGORITHMS

HUANG Shi – cun[1], ZHANG Wen – yi[1], HE Guo – jin[1], ZHENG Wan – qin[1], WU Hai – ping[2]
(1. *China Remote Sensing Satellite Ground Station*, *Beijing* 100086, *China*; 2. *China Land Surveying and Planning Institute*, *Beijing* 100035, *China*)

Abstract: This paper describes and tests three kinds of transform algorithms for remote sensing image rectification. A comparison with results from PCI commercial remote sensing software shows that the Givens transform algorithm is on the whole superior to other algorithms in precision in 1 ~ 5 order multinomial rectification.

Key words: Multinomial rectification model; Geometry rectification; Linear least – squares algorithm; RMS error

第一作者简介: 黄世存(1977 –),男,中科院中国遥感卫星地面站硕士研究生,地图学与地理信息系统专业,研究方向为遥感图像信息处理。

(责任编辑: 刁淑娟)