

DOI :10.3876/j.issn.1000-1980.2009.05.002

大尺度流域河网二叉树编码方法

王 皓 ,李铁键 ,高 洁 ,傅旭东 ,王光谦

(清华大学水沙科学与水利水电工程国家重点实验室 ,北京 100084)

摘要 :数字流域模型(DWM)中基于二叉树理论的河网编码方法能够实现任意河段的直接定位和高效的拓扑运算 ,然而由于该模型针对大尺度流域河网存在编码能力不足的隐患 ,降低了该模型的通用性和有效性 .为此 ,研究了河网二叉树编码的通用方法 ,并开发了全自动可视化平台 ,可以实现任意尺度流域河网的二叉树编码 ,为进行后续流域水沙运动模拟奠定了基础 .将该方法应用于中国西藏境内约 3.2 万 km² 的拉萨河流域 ,实现了 59 197 条河段的快速编码 ,并按河网自然拓扑形态和流域天然分界将全流域划分为 3 级 14 区 ,验证了该方法改进后的有效性 .

关键词 :大尺度流域 ;分布式水文模型 ;河网编码 ;拉萨河

中图分类号 :TV212.4 文献标识码 :A 文章编号 :1000-1980(2009)05-0499-06

王光谦等^[1-4]阐述了黄河数字流域模型原理 ,提出了一个具有多层次结构的数字流域模型框架 ,以及一种基于二叉树理论 ,以二元形式表示的河网编码方法(二叉树编码)^[5-6] .该方法能够实现任意河段的直接定位和高效的拓扑运算 ,以及同一河流产流、产沙过程和计算演进的统一 ,同时该方法使得分布式水文模型很容易进行子河网的自动拆分 ,实现全流域水沙运动的并行计算^[7-8] .然而在实际应用中发现 ,数字流域模型现有编码机制对大尺度流域的处理能力明显不足 ,存在编码溢出、河段缺编的潜在隐患 ,会导致后续计算中断执行 .虽然通过人工手段可以解决这一问题^[6] ,但是其可操作性低 ,适用性较差 ,很难进一步推广 ,限制了二叉树编码方法在大尺度流域中的应用 .

针对以上问题 ,本文研究了大尺度流域河网二叉树编码的一般方法和实现流程 ,并开发出全自动可视化平台 ,可快速实现大尺度流域河网的二叉树编码及天然子流域的自动分区 ,为后续流域进行水沙运动的并行模拟奠定了基础 .

1 河网二叉树编码原理及实现方法

1.1 数字化河网提取

对流域进行河网编码首先要获取全流域数字化河网及相应拓扑信息 .现有多种软件平台可以使用 ,常用的有 RSI 公司开发的 River Tools^[9] ,ESRI 公司提供的 GRID 模块^[10] ,美国农业部农业研究中心开发的 TOPAZ 模块^[11] 等 .其中 TOPAZ 模块能够提取详尽的流域河网拓扑属性信息 ,在国内外众多水文模型中得到了广泛使用 .本文即采用 TOPAZ 模块 ,通过数字高程模型(digital elevation model ,DEM)数据进行数字化河网提取 .

河网提取结果如图 1 所示 ,每个基本产流单元(元流域)由源坡面、左坡面、右坡面、沟道 4 部分组成 .通过 TOPAZ 模块可以得到各个元流域的拓扑属性 ,如面积、坡度、长度等信息 ,并以“记录”的形式存储于数据库表中 ,见表 1 .其中只有 Stralher 级别为 1 的单元具有源坡面 ,其他级别只具有左

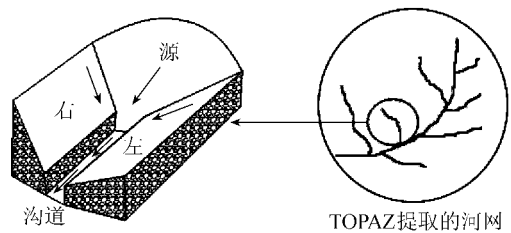


图 1 由沟道和坡面系统组成的基本产流单元
Fig. 1 Basic runoff unit composed of a channel and slopes

坡面和右坡面.在元流域属性中用4个值表征相邻元流域间的连接关系,即元流域入口和出口的横纵坐标.坐标值以出入口在DEM输入文件中所对应的行列值来表示,见表1.从表1可以发现,元流域1的出口坐标与元流域2的入口坐标完全相同,因此元流域1在元流域2的紧邻上游,这样就可以通过该4条属性,将所有河段组织成完整的流域河网.为叙述方便,本文提到的“河段”和“元流域”代表相同意义,对河段进行的编码即是对元流域的编码.

表1 TOPAZ提取的基本产流单元属性信息
Table 1 Attribute information of basic runoff unit extracted by TOPAZ

元流域	元流域编号 (唯一)	Stralher 分级数	元流域坐标				坡面面积/m ²			沟道 长度/m	沟道 坡度/(°)
			入口		出口		源坡面	左坡面	右坡面		
			行	列	行	列					
元流域 1	1625	1	26	1560	53	1559	308 289	1 180 681	1 338 105	2 899	0.003 1
元流域 2	1624	2	53	1559	70	1538		1 675 911	1 833 335	3 014	0.002 3

1.2 流域河网的二叉树化

对流域河网进行二叉树编码首先要将其转化为标准的二叉树结构.二叉树有2个明显特征:(a)任一节点有且仅有1个父节点(根节点除外);(b)任一节点有且仅有2个子节点(最末端节点除外).具有同一父节点的2个节点,以下称为兄妹节点.

如图2(b)所示,这是一个标准的二叉树结构,其中每个节点对应于流域中的一条河段(图2(a)).根节点A是全流域的出口河段;D和E是兄妹节点,作为汇入河段B的子节点,是B的紧邻上游河段;A作为B的父节点,是B的紧邻下游河段. TOPAZ提取的河网在很多情况下并不是一个标准的二叉树结构,会出现某节点只有单个子节点或多于2个子节点的现象,其他很多软件也有类似问题.前者比较容易解决,只需在编码时将不存在的子节点用空(NULL)来代替,即可保证子节点数目的完整性;对于后者,本文采用的方法是在任意2个节点之间插入一个“虚节点”,以维持河网的二叉树特征.以3个节点为例,如图2(a)所示,河段H,I,J同时汇入河段E,即节点E存在3个子节点,此时在节点I和J之间插入“虚节点”K,令节点I和J汇入节点K,再指定节点K和H汇入节点E,这样E和K刚好都具有2个子节点,符合二叉树的基本特征.“虚节点”具有零面积、零坡度、零长度特性,因此在流域水沙模拟过程中只占据编码位置,并不影响实际计算.

1.3 集水面积的遍历更新

二叉树编码方法以集水面积来判定兄妹节点的身份.集水面积大的为兄节点(占据左侧位置),集水面积小的为妹节点(占据右侧位置).这样可以很方便地区分干支流.如图3所示,河段B的集水面积比河段C的大,则河段B作为兄节点隶属于干流,C属于支流.因此对全流域河网进行二叉树编码前,首先要计算出所有节点的集水面积,以判定兄妹节点间的编码次序,该过程是很自然的二叉树后序遍历^[12]过程.如图3所示,图中圆圈里数字代表河网集水面积的遍历更新顺序,与二叉树后序遍历次序完全吻合;圆圈上面的英文字母为河段标识,与图2中的河段标识一一对应.以节点D为例,设节点D的集水面积为 S_0 ,其源坡面、左坡面、右坡面的集水面积分别为 S_s, S_l, S_r ,2个子节点F,G的集水面积分别为 S_F, S_G ,则

$$S_0 = (S_s + S_l + S_r) + (S_F + S_G)$$

(1)

式中: S_s, S_l, S_r 为已知量,由TOPAZ模块提供,这样按照图3标出的从上游到下游的二叉树后序遍历顺序,

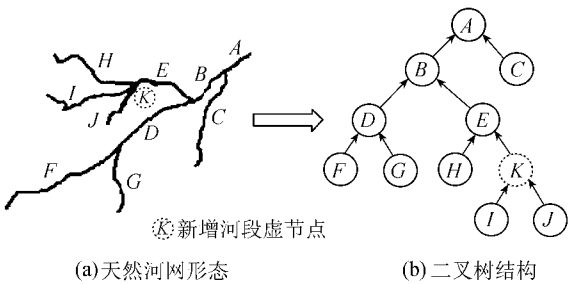


图2 流域河网的二叉树表示
Fig. 2 Binary-tree structure for drainage network of a basin

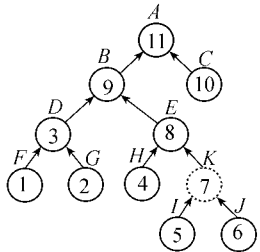


图3 河网集水面积的遍历更新
Fig. 3 Traversal updating for catchment area of drainage network

执行式 1) 操作 ,可以得到所有二叉树节点的集水面积值.

1.4 现有二叉树编码方法及存在的问题

现有二叉树编码以(BsLength ,BsValue)二元形式表示 ,如图 4(a)所示 ,其中 BsLength 表示河段节点在河网二叉树中所位于的层数 ,最小值为 1 ,从上到下依次递增.该值越大 ,说明河段距离流域出口越远. BsValue 表示河段节点距离干流的距离 ,在同一层内 ,从 0 开始 ,从左到右依次递增.该值越大 ,说明该河段距离干流越远.二叉树编码满足以下关系 :

$$\begin{cases} B_L = G_L = L_L + 1 \\ B_V = 2L_V \\ G_V = 2L_V + 1 \end{cases} \quad (2)$$

式中(L_L, L_V)为当前节点的二叉树编码 (B_L, B_V) (G_L, G_V)为其子节点的二叉树编码.则根据式(2),可以得到图 2 中流域河网的二叉树编码结果(图 4(a)). 同样 ,根据当前节点的二叉树编码也很容易反推出其父节点的编码值(P_L, P_V) ,其关系满足 :

$$P_L = L_L - 1 \quad P_V = \left[\frac{L_V}{2} \right] \quad (3)$$

式(3)是式(2)的逆运算 ,它在流域水文模拟中有很实际的应用 ,因为流域水文过程计算要求按照从上游到下游的顺序依次进行 ,上游河段的计算输出是下游河段的计算输入.而根据式(3)可由上游河段的二叉树编码直接计算出其紧邻下游河段的编码值 ,进而快速定位该河段以进行信息传递 ,避免了传统河网编码方法低效率的查找和搜索的弊端 ,大大提高了计算效率 ,保证了水流运动与计算模拟的一体化和同步性.

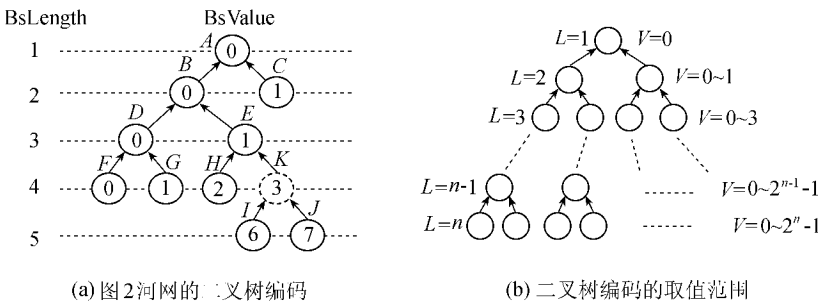


图 4 河网二叉树编码及取值范围

Fig. 4 Binary-tree coding example of drainage network and corresponding value range

虽然二叉树编码方法本身有着很多优良的性质 ,然而在实际运用中发现 ,现有二元编码机制还不能完全解决大尺度流域河网编码问题 ,存在着编码溢出的潜在危险.图 4(b)标出了流域河网二叉树中每层 BsLength 的取值及相应 BsValue 的取值范围(图中用 L, V 表示) ,根据式(2)可以很容易计算得出.从图 4(b)可以看到 , BsValue 是以 2^n 量级迅猛增长的 ,而现有计算机可存储的最大正整数为 2^{64} ,也就是说 ,如果流域中存在某条较大支流 ,其纵深发育层数超过 64 层 ,就会存在河段无法用某正整数来标识 BsValue 值的问题 ,出现编码溢出的情况.

1.5 大尺度流域的自动分区及耦合

如上所述 ,为解决现有二元编码机制在大尺度流域上的应用问题 ,本文引入第三元关系 :子流域标识码 RegionIndex ,用于进行大尺度流域河网支流的分区标识.其基本思想是 :如果流域规模过大 ,支流纵深发育很长 ,不能用二元编码来标识所有河段 ,那么可以对流域先进行分区.分区方法是从代表全流域的巨型二叉树中裁减出大小规模合适(BsLength 值小于或等于 64)的子二叉树 ,每个子二叉树就代表 1 个分区并被赋予 1 个唯一的 RegionIndex 值 ,然后对各分区分别进行二叉树编码 ,由于分区二叉树结构的 BsLength 最大可能值为 64 ,从图 4(b)可以知道 ,分区 BsValue 的最大可能值为 $2^{64} - 1$,这就不会出现编码溢出的问题 ,最后再通过某种方式将编码后的各分区进行耦合 ,即可得到经过二叉树编码后的全流域一体化河网.由此可知 ,改进后的二叉树编码可用三元关系(RegionIndex ,BsLength ,BsValue)来表示 ,RegionIndex 被设定为如下关系 :

$$L_R = 100P_R + i \quad (1 \leq i \leq 99) \quad (4)$$

式中 : P_R ——流域分区前的 RegionIndex 值 ; L_R ——新分区 RegionIndex 值 ; i ——新分区的 RegionIndex 预留的

两位编号.如全流域干流为一级河流,RegionIndex 设为 0,那么由式(4)可知,与干流紧邻的二级河流的 RegionIndex 取值范围为 1~99,即在式(4)意义下,最多有 99 个二级支流汇入一级支流,以此类推.对于大自然中的实际河网,如果有多于 99 条二级支流的情况,式(4)也很容易进行再拓展.如可将公式中的 100 变为 1000, i 范围变为 1~999.不管如何变化,式(4)均可保证子流域标识码 RegionIndex 的全局唯一性.

流域具体分区过程见图 5(a).最开始流域所有河段的 RegionIndex 设为 0,并根据式(2)对河网进行二叉树编码.如果在编码过程中发现某节点 $(0, L_1, NaN)$ 的 BsValue 超出允许阈值(阈值可事先设定,最大为 2^{64} .图中 NaN 表示 BsValue 超出允许阈值),表明该节点所在二级支流纵深太长,规模过大,需要将其所在支流进行裁减分区.于是进行二叉树的逆向查找过程(式(3)),从节点 $(0, L_1, NaN)$ 出发,快速定位至其所在最大支流的出口河段 $(0, 3, 1)$ (图 5(a)),并将其作为新分区的根节点,赋予新编码值 $(1, 1, 0)$,整个新分区的 RegionIndex 也变为 1(式(4)),然后对该分区重新执行二叉树编码操作(图 5(b)).如果新分区在编码时仍发现存在某节点如 $(1, L_2, NaN)$ 的 BsValue 超值,则要进行支流再分区,得到三级支流,如图 5(c)所示,其 RegionIndex 变为 101(式(4)).对其他各支流执行同样操作,直到全流域所有河段都完成编码为止.需要说明的是,在对各分区进行二叉树编码时,仍然以集水面积大的节点作为兄节点(占据左侧位置),集水面积小的作为妹节点(占据右侧位置)的规则.

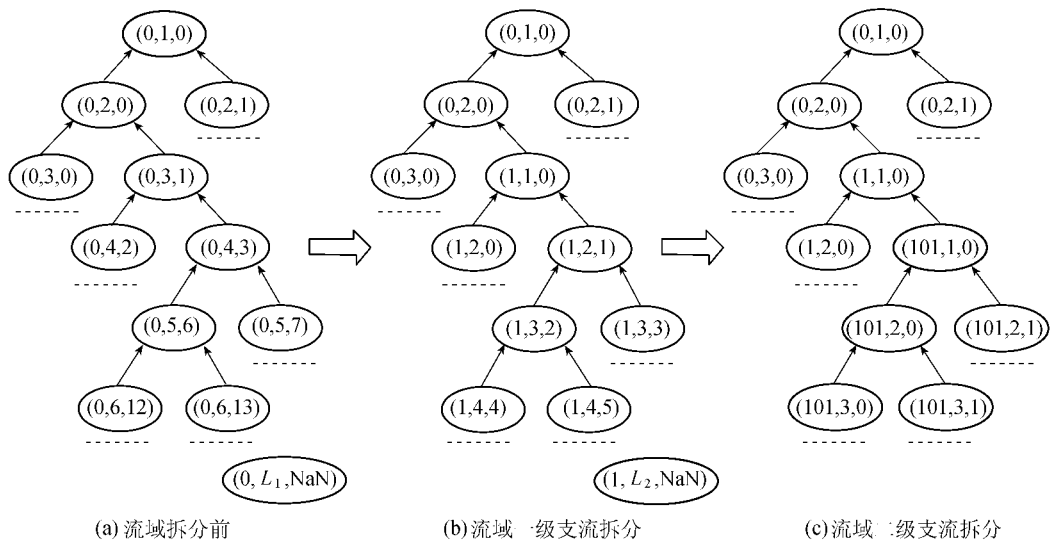


图 5 大尺度流域河网二叉树编码的自动分区

Fig. 5 Automatic partition of binary-tree coding for drainage network of a large-scale basin

水沙过程模拟是针对全流域进行的,不同分区间的耦合关系也一定要得到表达,因为只有这样才

才能实现上下游分区间的信息传递,完成对全流域的计算模拟.图 5(c)中展现的不同分区间的耦合关系如表 2 所示.以第一行记录为例,它表示的意义是:河流级别为 2,RegionIndex 为 1 的支流分区,汇入其下游 RegionIndex 为 0 分区中的 $(0, 2, 0)$ 河段.可以看到,通过表 2 可以表达出不同支流分区间的相互耦合关系,得到全流域一体化的河网编码信息,从而实现全流域从整体到局部再到整体的关系转换.

改进后具有三元编码形式的二叉树编码,可以保证分布式水文模型对大尺度流域能够开展计算模拟,拓展了原有二叉树编码方法的适用维度和应用范围.实际上,三元编码形式对于小尺度流域依然适用.如果流域尺度较小,河网较稀疏,三元编码形式会自动还原为二元形式,以节省内存空间,提高搜索效率.在水文领域,往往需要对具有高精度河网信息的小尺度流域进行精细水文过程模拟,此时虽然流域面积不大,但由于河网精度较高,河段数目很多,支流的纵深也相应很大.该情况实际上等效于大尺度流域情况,只有使用三元编码形式的二叉树编码方法才能完成对全流域的模拟计算.

需要说明的是,改进后的三元编码形式可以彻底描述和刻画地球上任意流域、任意精度的河网信息.随

表 2 流域分区间的耦合关系

Table 2 Coupling relation of subareas

当前子流域级别	当前子流域编码	连接河段 RegionIndex 值	连接河段 BsValue 值	连接河段 BsLength 值
2	1	0	0	2
3	101	1	0	1

着河网密度和精度的加大,支流分区数不断加大,不需再增加第四元编码值.三元编码形式由于增加了编码的复杂性,可能在二叉树遍历过程中增加一些搜索的额外时间(相对于原始二元编码),但是该时间很少,几乎可以忽略不计,这在实际计算中已得到了检验.正如上文所说,在流域河网精度不是很高的情况下,三元编码会自动转化成二元编码形式,以保持原来编码形式的运转高效性.

2 二叉树编码方法在拉萨河流域的应用

拉萨河是雅鲁藏布江流域面积最大的一条支流,发源于念青唐古拉山脉南麓,海拔 5 200 m,流域面积约 3.2 万 km².拉萨河是拉萨市的母亲河与大动脉,对当地的经济、居民生活、生态环境等都有着极其重要的影响.本文使用 TOPAZ 模块对拉萨河流域 85 m 精度的 DEM 数据提取了 59 197 个河段单元(图 6).由于河网精度较高,很多支流的纵深层数已经远远超过 64 层,使用原始二元编码方法已经无法完成对全流域的一体化编码,因此需要采用改进后三元编码形式的二叉树编码方法,以对全流域进行多支流自动分区及相互耦合.本文使用自主开发的大尺度流域河网二叉树编码平台完成了对全流域所有河段的一次性快速编码,全过程仅 1 min.全流域自动划分成 3 级 14 区,为其进行后续的拉萨河流域分布式水沙运动过程的计算模拟奠定了重要基础.

图 6 中显示了对拉萨河流域所有河段进行二叉树编码后的支流分区结果,图上不同的数字即代表不同分区的 RegionIndex,每一个 RegionIndex 都是一个完整的子流域,所有分区联合起来就构成了代表全流域的二叉树河网,通过自动化平台得到的流域不同支流分区间的相互耦合关系如表 3 所示.

可以看到,二叉树编码本身体现了河网的内在结构特征,在该编码方法下形成的局部流域河网与整体流域的关系是与河流本身发育的特点相符的,很好地反映出支流分区与全流域之间局部与整体的关系.河网的分区与各个干支流天然的分界相互统一,同一条河流不会被拆分,断点都在河流交汇处,因此在经过二叉树编码方法标识的河网中,不管是局部还是整体,抑或是从局部到整体的过程,都能体现河流本身从上游到下游,从低级到高级的自然发育特征^[1].

3 结 论

河网二叉树编码方法对于流域水沙过程的分布式模拟极为高效.本文针对 DWM 原有编码机制对大尺度流域处理能力的缺陷,研究了大尺度流域河网二叉树编码的通用方法和有效实现途径,并开发了全自动可视化平台,可以快速实现大尺度流域高精度河网的二叉树编码.该平台应用于中国西藏境内约 3.2 万 km² 的拉萨河流域,实现了 59 197 条河段的一次性编码,并按河网自然拓扑形态和流域天然分界将全流域自动划分为 3 级 14 区,为其进行后续的专业模拟和并行计算奠定了基础.

参考文献:

[1] 王光谦,刘家宏,李铁键.黄河数字流域模型原理[J].应用基础与工程科学学报,2005,13(1):1-8.(WANG Guang-qian,LIU Jia-hong,LI Tie-jian, Digital watershed model of Yellow River[J]. Journal of Basic Science and Engineering,2005,13(1):1-8.(in



图 6 拉萨河流域二叉树编码的支流分区
Fig. 6 Tributary division of Lhasa River basin after bianry-tree coding

表 3 拉萨河流域支流分区间的耦合关系

Table 3 Coupling relation of subareas for Lhasa River basin				
当前子流域 级别	当前子流域 编码	连接河段 RegionIndex 值	连接河段 BsValue 值	连接河段 BsLength 值
2	1	0	1	472
2	2	0	1	434
3	201	2	1	18
2	3	0	1	372
2	4	0	1	294
3	401	4	1	60
3	402	4	1	2
2	5	0	1	185
2	6	0	1	142
2	7	0	1	106
2	8	0	1	54
3	801	8	1	92
3	802	8	1	45

Chinese))

- [2] LIU Jia-hong ,WANG Guang-qian ,LI Tie-jian ,et al. Sediment yield computation of the sandy and gritty area based on the digital watershed mode[J]. Science in China Series E :Technological Sciences 2006 49(6) :752-763.
- [3] 刘家宏 ,王光谦 ,王开. 数字流域研究综述[J]. 水利学报 2006 37(2) :240-246.(LIU Jia-hong ,WANG Guang-qian ,WANG Kai. Review on advancement of study on digital riverbasin in China[J]. Journal of Hydraulic Engineering ,2006 ,37(2) :240-246.(in Chinese))
- [4] WANG Guang-qian ,WU Bao-sheng ,LI Tie-jian. Digital Yellow River mode[J]. Journal of Hydro-environment Research ,2007 ,1(1) :1-11.
- [5] 李铁键 ,王光谦 ,刘家宏. 数字流域模型的河网编码方法[J]. 水科学进展 2006 ,17(5) :658-664.(LI Tie-jian ,WANG Guang-qian ,LIU Jia-hong. Drainage network codification method for digital watershed mode[J]. Advances in Water Science 2006 ,17(5) :658-664.(in Chinese))
- [6] 邓利 ,李铁键 ,刘家宏 ,等. 数字流域河网编码方法应用实例[J]. 泥沙研究 2007 3(3) :1-5.(DENG Li ,LI Tie-jian ,LIU Jia-hong ,et al. Application of the drainage network codification method in digital watershed mode[J]. Journal of Sediment Research 2007 3(3) :1-5.(in Chinese))
- [7] 李铁键 ,刘家宏 和 杨 ,等. 集群计算在数字流域模型中的应用[J]. 水科学进展 2006 ,17(6) :841-845.(LI Tie-jian ,LIU Jia-hong ,HE Yang ,et al. Application of cluster computing in the digital watershed mode[J]. Advances in Water Science 2006 ,17(6) :841-845.(in Chinese))
- [8] SHANG Yi-zi ,WU Bao-sheng ,LI Tie-jian ,et al. Fault-tolerant technique in the cluster computation of the digital watershed mode[J]. Tsinghua Science and Technology 2007 ,12(S1) :162-168.
- [9] 张国义 ,房明惠 ,徐云 ,等. RSI River Tools 系统及其应用介绍[J]. 计算机应用 2002 ,22(8) :38-40.(ZHANG Guo-yi ,FANG Ming-hui ,XU Yun ,et al. A brief view of RSI River tools and its application[J]. Computer Applications ,2002 ,22(8) :38-40. (in Chinese))
- [10] 张超 ,郑钧 ,张尚弘 ,等. ArcGis 9.0 中基于 DEM 的水文信息提取方法[J]. 水利水电技术 2005 36(11) :1-4.(ZHANG Chao ,ZHENG Jun ,ZHANG Shang-hong ,et al. Extraction of hydrological information from digital elevation model with ArcGis 9.0[J]. Water Resources and Hydropower Engineering 2005 36(11) :1-4.(in Chinese))
- [11] JURGEN G ,LAWRENCE W M. TOPAZ overview[R]. Oklahoma :U S Department of Agriculture ,Agriculture Research Service ,1999.
- [12] BRUNO R P. Structures and algorithms[M]. Beijing :Electronic Industry Press 2000 :197-231.

Binary-tree coding for drainage network of large-scale basins

WANG Hao , LI Tie-jian , GAO Jie , FU Xu-dong , WANG Guang-qian

(State Key Laboratory of Hydrosience and Engineering , Tsinghua University , Beijing 100084 , China)

Abstract : The binary-tree coding method for the drainage network based on the binary tree theory in the digital watershed model is able to realize topological operation and direct location for arbitrary river segment. However , the method is probably of coding overflow defect for the drainage network of large-scale basins , greatly reducing its applicability and validity. The general method of the binary-tree coding was studied , and an automatic visualization platform was developed , which could realize the binary-tree coding for arbitrary scale basin. A solid basis was made for the subsequent modeling of water and sediment motion in basins. The present method was used in the Lhasa River basin with an area of 32 000 km² in Tibet , China. It realized the drainage network coding of 59 197 river segments in only one minute and the basin was divided into 3 grades and 14 regions according to the natural topological configuration of the drainage network and the natural dividing line of the basin. The feasibility and effectiveness of the improved method were thus verified by the results.

Key words : large-scale basin ;distributed hydrological model ;drainage network coding ;Lhasa River