基于 GeoSOT 剖分框架的海量气象海洋数据联合检索技术研究

Research on Joint Retrieval of Large-scale Meteorological and Ocean Data Based on GeoSOT Framework

修义瑞　杨亮　何锡玉　王煦莹

Yirui Xiu　Liang Yang　Xiyu He　Xuying Wang

91001 部队　中国·北京 100166

Troops 91001, Beijing, 100166, China

摘　要：随着水文气象观测手段和预报技术的日益发展，各类水文气象基础保障数据种类日益增多，数据量呈现爆炸式增长态势。论文提出将 GeoSOT 全球剖分组织框架作为气象水文数据组织管理系统的建设基础，从剖分方案、编码方法、组织方式三个方面构建了水文气象数据剖分区位编码模型，从而实现多源数据的联合检索、局部数据的精准汇集。

Abstract: With the increasing development of hydrometeorological observation means and forecasting technology, the types of basic hydrometeorological support data are increasing day by day, and the amount of data shows an explosive growth trend. The paper puts forward that the GeoSOT global subdivision organization framework is used as the foundation for the construction of the meteorological and hydrological data organization and management system, and the hydrological and meteorological data subdivision location coding model is constructed from three aspects of the subdivision scheme, coding method and organization mode, so as to realize the joint retrieval of multi-source data and the accurate collection of local data.

关键词：水文气象；GeoSOT；剖分区位编码；联合检索

Keywords: hydrometeorology; GEOSOT; zonal coding; joint search

DOI: 10.12346/csai.v1i1.6880

1 引言

气象海洋观探测数据作为水文气象预报业务的核心基础，其丰富的数据来源、多样的观测方法、精细的探测手段、海量的数据资源都是水文气象预报水平得以不断提升的有力保障。随着观测手段和预报技术的日益发展，当前水文气象预报基础保障数据种类十分多样化，主要包括常规观探测资料、船舶报、飞机报、卫星遥感资料、雷达资料、浮标资料、海洋站资料、数值预报产品、数值再分析产品等，其数据量呈现出爆炸式的增长态势，已有向 PB 级发展的趋势。水文气象预报和水文气象装备辅助决策系统均是基于对多源气象海洋数据的快速融合应用，具有很强的数据交叉融合特点。目前国际气象海洋中心都已建立了一套健全的数据总库管理机制与运维系统 [1,2]，但针对很多应急任务，需要快速汇集大量资料的需求下，未优化的检索能力已难以很好地满足快速应急保障要求。因此面对庞大的基础气象水文数据，如何实现有效的组织管理、高效的检索和快速的汇集已成为迫在眉睫的需要解决的技术问题。

论文以提高水文气象基础数据共享管理能力为目标，提出以 GeoSOT 全球剖分组织框架作为气象水文数据组织管理系统的建设基础，从剖分方案、编码方法、组织方式三个方面构建水文气象数据剖分区位编码模型，建立基于剖分区位编码的检索汇集规则，实现多源数据的联合检索和局部数据的精准汇集。

2 GeoSOT 全球剖分网格简介

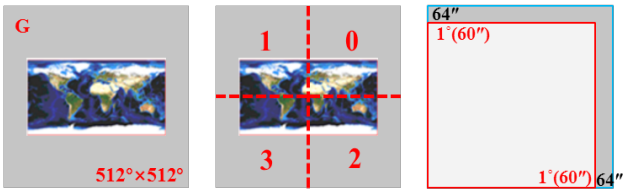
全球剖分组织框架 GeoSOT[3,4] 是由中国一大学的著名教授提出，与中国正在使用的空间信息网络系统有比较强的聚合与关联关系，在继承已有的网络系统前提下，对一些关

【作者简介】修义瑞（1964-），男，满族，中国辽宁鞍山人，本科，高级工程师，从事数值天气预报研究。

键的数据进行多维的立体剖分，面向空间信息提供了统一的区位剖分面片集合框架，从而实现高效的多源数据空间索引与应用。

GeoSOT 网格体系是属于等经纬度的四叉树剖分网格，经纬度数值空间均定义在 [-256°，256°] 上。由于地球经纬度的值域为经度 [-180°，180°]、纬度 [-90°，90°]，故 GeoSOT 网格中部分区域不属于实际的地理空间范围。GeoSOT 剖分 0 级网格定义为以赤道与本初子午线交点为中心点的 512° ×512° 方格，0 级网格编码为 G，含义为全球 Globe，如图 1（a）所示。GeoSOT 剖分 1 级网格在 0 级网格基础上平均分为四份，每个 1 级网格大小为 256° ×256°，网格编码为 Gd，其中 d 为 0、1、2 或 3，如图 1（b）所示。2 级网格在 1 级网格基础上平均分为四份，每个 2 级网格大小 128° ×128°，网格编码为 Gdd，其中 d 为 0、1、2 或 3。

GeoSOT 剖分 9 级网格大小为 1° ×1°，9 级以上网格为 GeoSOT“度”级网格，第 10~15 级网格为“分”级网格、第 16~21 级为“秒”级网格。“分”级面片起始点为 9 级网格的 1° 面片（或 60′ 面片），编号连续，网格的起始数值空间大小由 60′ 外延到 64′，如图 1（c）所示，GeoSOT10 级剖分网格以 64′ ×64′ 大小平均分为四份，每个 10 级网格大小为 32′ ×32′，网格编码为 Gddddddddd-m，其中 d、m 为取值 0、1、2 或 3 的四进制数。10~15 级网格为“分”级网格，剖分大小和编码形式按照上述规则递归。秒级的网格剖分处理方式与分级的相同。最后的 32 级网格大小为 1/2048″ ×1/2048″。



(a) (b) (c)

图 1 GeoSOT 网格剖分示意图

3 气象水文数据剖分区位编码模型设计

3.1 建模思路

在基于原本的 GeoSOT 剖分网格框架基础之下，基于统一的剖分编码实现各类数据、图形和资料与剖分网格的有效关联，并以此编码为索引项，建立剖分索引表或索引文件，通过确定查询范围的剖分层级与网格编码，实现对多源异构数据的高效检索与快速汇集，整体思路如图 2 所示。通过建立统一的检索与表达剖分框架，实现对多源异构数据综合管理与调度，其中涉及的网格剖分编码计算与生成可参见金安提出的剖分编码方法 [6]。

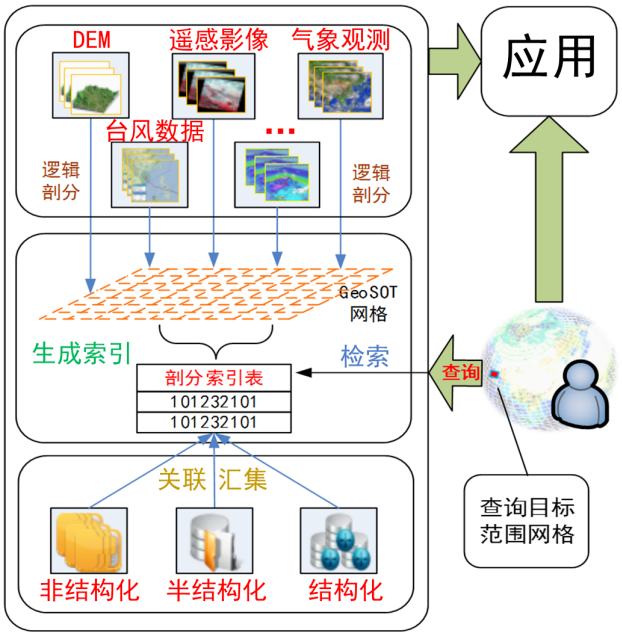


图 2 剖分关联索引示意图

论文选取了两类典型性气象水文数据来建立剖分区位编码模型，一类是单站类型数据（固定陆地站观测报、海洋站观测报、航线天气报告等），另一类是标准格网化类型数据（数值预报产品、再分析产品、卫星资料等）。

3.2 单站类型数据编码模型

3.2.1 剖分方案

单个站体的水文气象数据的空间位置一般以经纬度坐标作为参考基准，根据剖分理论该点必然落入多个具有包含关系的层级面片中，不同的层级代表了不同的尺度，这里规定：针对站点数据合理的剖分面片层级是一个面片只包含一个站点的最大层级，因此选定层级不应大于第 12 层级，否则会出现多点同格的现象。根据选中的水文气象台站的实际分布状况，选出第 14 层（剖分单元为 2′，赤道附近覆盖范围 4 km × 4 km）作为站点数据的剖分层级最为合理。

3.2.2 编码方法

由单站类型数据的编码模型方案可知，此类型数据最大剖分层级为 14 级，最小剖分层级 12 级，且一条数据信息仅需要一个编码索引。如图 3 所示，该类型的区位编码具备必设项与可设项两种属性，由 4 个部分组成，分别是剖分编码、数据类型编码、报文类型编码及扩展属性码。其中，①剖分编码为必设项是区位编码的核心，设计最大长度 30 位，有效使用长度在 24 位至 28 位之间；②数据类型编码为必设项，设计定义为 0000；③报文类型编码为可设项，随着编码应用的细化或条件检索的需要，可参考数据类型编码方法将符合单站类型数据的报文种类进行编码，并记录在此编码段内，理论上现有的此类报文种类尚未超过 30 种，故此编码段预设 7 个二进制位，最大记录种类可达 64（26）个；④扩展属性码属于可设项，且本文的模型中暂未用到，预设

此编码段主要用于满足后续的拓展需求。

3.2.3 组织方式

单站类型数据结构相对简单，一般存储在结构化的关系型数据库中，因此索引编码可通过增加表字段的方式进行存储，同时可利用数据库的索引引擎初步提高检索效率。具体需要增加的字段见表 1。

3.3 标准网格化类型数据编码模型

3.3.1 剖分方案

标准化格网类型的数据主要包括数值预报产品、再分析产品、卫星遥感融合产品等，这类数据都是以等间隔的格网点标记空间位置信息后，再在该点上记录各类属性信息。本节以欧洲中期数值天气预报产品为例进行建模。

同一时次同一要素的全球格点预报资料由 8 个数据点集拼接而成，如 4(a)所示。考虑到实际的应用方式与检索效率，本文将此类格点资料按第 2 级进行剖分，如 4（b）所示，若资料为处理过的合成全球资料则不进行剖分处理，直接定义剖分层级为 0 级，该类数据的剖分以示意性定位为主，具体的区别将在编码方法中定义。

3.3.2 编码方法

标准化格网类型的数据的区位编码包括剖分编码、数据类型编码、报文类型编码、覆盖范围标识、数据特性码和扩展属性码六个部分组成。其中，①剖分编码为不定长编码，记录一个或一组区位编码；②数据类型编码，设计定义为“0001”；③报文类型编码为“000”；④覆盖范围标识为“00”与“01”，“00”表示全球覆盖，“01”表示局部覆盖；⑤数据特性码分为两种情况，若为欧洲格点资料则为“75”与“25”，分别表示 0.75° ×0.75° 与 0.25° ×0.25° 两种分辨率格式；⑥扩展属性码中欧洲格点资料记录文件名信息。

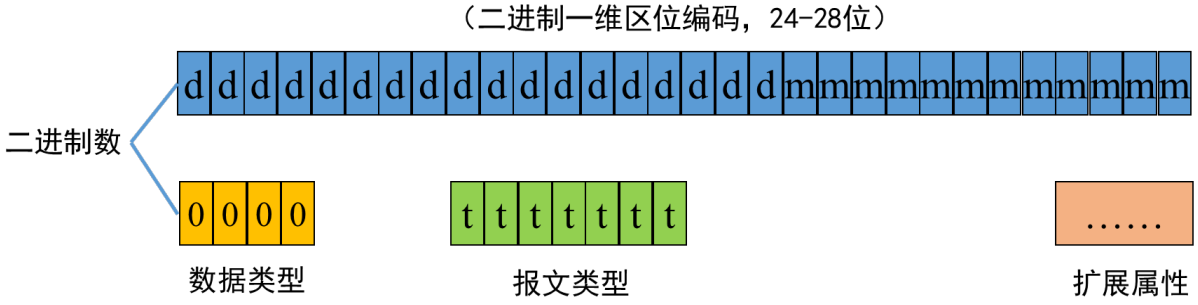
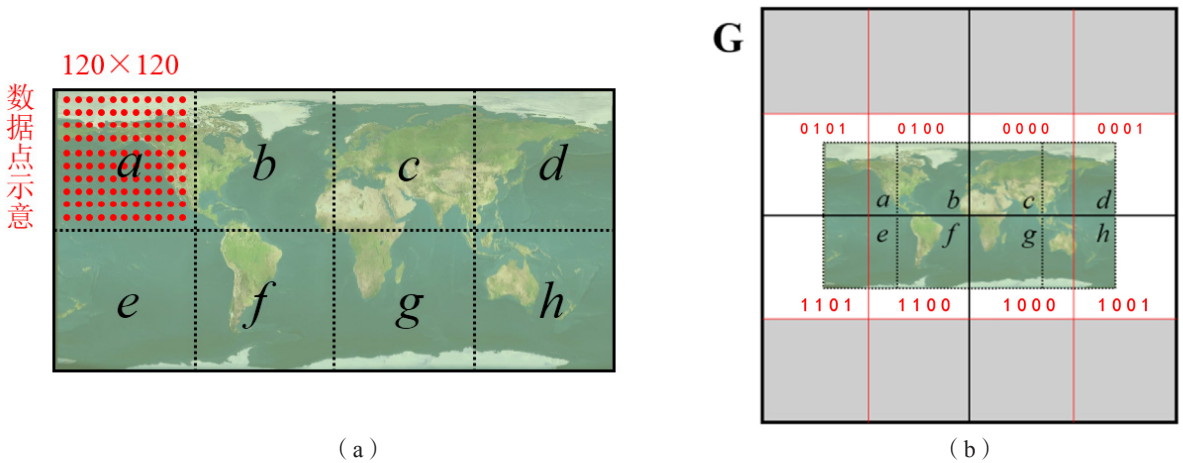


图 3 单站类型数据编码方法示意图

表 1 单站类型数据表字段信息

序号	字段名称	数据类型	字段长度	是否可空	字段描述	内容描述
1	code_geo	varbinary	30	否	区位编码	
2	type_data	binary	4	否	数据类型	0000
3	type_m	binary	7	是	报文类型	预留字段
4	others			是	扩展属性	



（a）　　（b）

图 4 格点资料示意及剖分效果图

3.3.3 组织方式

数值预报产品有存储一般分为数据库存储和文件存储两类。针对存储在数据库中的数值预报格点资料（以欧洲格点报为例），通过对数据表添加字段的方式，实现剖分区位编码信息的嵌入，表 2 中设计了数据库表中需要增加的字段信息。

针对原始的欧洲中心数值预报产品文件，建立一个专属的区位编码文件，文件前名称为产品文件的全名，文件后缀为“*.geo”。“*.geo”文件以符合标准的“xml”语言组织要素信息。

3.4 试验验证

为了验证剖分区位编码模型的可行性与高效性，论文实验以本地主频为 2.4 GHz 的双核 CPU、4 GB 内存的 PC 机为载体，使用全国固定陆地站观测资料 50 GB 和数值预报数据 360 GB 进行编码生成速度量算，冗余增量统计以及汇集效率测试。

编码生成速度及冗余增量测试结果如表 3 所示。由测试结果可看出，编码生成方案对单点经纬度的结构化数据效率最高，而对面状数据需要有较多的判断与剖分层级变化，相对平均耗时要高一些；数据冗余方面，由于采用二进制数进行编码存储，故极大地节省了存储空间，尤其是格点数据本身数据量较大，导致数据增加几乎可忽略不计，从本结果不难发现，利用 GeoSOT 进行编码化预处理的时间和空间代价都是极小的。

针对某一指定 2 × 2 km 区域范围进行剖分编码检索汇集与经纬度汇集对比测试，分别比较检索汇集耗时与汇集有效数据量，得出结果如表 5 所示。从测试结果不难发现通过剖分编码进行检索的速度明显远超过经纬度检索方法，分析巨大差异产生的原因，以单点数据为例，经纬度检索需要进行两次经度大小判断与两次维度大小判断，而剖分编码不同层级间空间位置的包含关系也体现在编码位的包含关系上，如编码为 G00101101 的面片必包含于编码为 G001011 的面片中，故而在进行编码包含关系判断时，仅需要一次判断即可。且剖分编码间的运算是基于二进制数展开的，在计算机层面减少了转换环节，尽可能地对运算效率实现了优化。通过大量的测试，其结果可以看出有效数据条目也比经纬度检索出的要多出一些，需要进一步完善（表 4）。

表 2 数值预报产品表字段信息

序号	字段名称	数据类型	字段长度	是否可空	字段描述	内容描述
1	code_geo	binary	4	是	区位编码	NULL 表示剖分层级为 0 级，即全球覆盖
2	type_data	binary	4	否	数据类型	0001
3	type_m	binary	3	否	报文类型	000
4	area	bit	1	否	覆盖范围	0：全球 1：局部
5	character	enum		否	数据特性	25、75
6	others	varchar	25	是	扩展属性	文件名

表 3 编码生成速度统计

序号	名称	数据大小（GB）	数据量（万条）	总耗（s）	平均耗时（ms/ 条）	冗余（MB）	增量比（‰）
1	气象单站数据	50	100	160	0.16	87	1.7
2	标准网格数据	360	45	113	0.25	33	0.09

表 4 汇集对比测试结果

序号	名称	检索汇集总耗时（min）		汇集有效数据量（条）	
		经纬度检索	剖分检索	经纬度检索	剖分检索
1	气象单站数据	25	3	7159	11159
2	标准网格数据	67	2	63290	63290

4 总结与展望

基于 GeoSOT 全球剖分组织框架，建立了单站类型数据编码模型和标准网格化类型数据编码模型，实现对海量水文气象数据的联合检索和局部数据的精准汇集。实验验证表明：基于剖分网格的多源数据快速汇集方法以具有明显位置包含关系的面片编码作为联合索引项，极大地提高了检索效率，同时通过剖分编码的生成，将确定空间范围的大量耗时工作在非紧迫时间或数据生产时提前完成，这种预处理与检索分步的方式成功地实现了应急状态下的高效检索，通过剖分编码进行检索的速度远超过传统经纬度检索方法。

GeoSOT 全球剖分组织框架具有全球覆盖、无缝无叠、面片逐级嵌套、剖分尺度灵活可变、剖分编码生成快捷、区位信息转换简单和包含关系明显等诸多特点与优势。本文建立的基于 GeoSOT 构建的各类数据类型编码模型和软件模块均以扩展插件的形式接入数据共享管理平台，在后台实现数据的编码生成，索引信息记录和联合检索汇集等功能，可成为现有数据管理平台的纯绿色插件。因此，该技术可以成为解决海量气象水文数据联合检索效率难以提高的一个有效手段，可成为水文气象与其他行业领域实现联合检索、交叠展示和综合应用的有力支撑平台，具有较高的实际应用价值。

参考文献

[1] 程承旗,付晨.地球空间参考网格及应用前景[J].地理信息世界,2014,21(3):1-8.

[2] 程承旗,任伏虎,濮国梁,等.空间信息剖分组织导论[M].北京:科学出版社,2012.

[3] 杨宇博,程承旗,郝继刚,等.基于全球剖分框架的多源空间信息区位关联与综合表达方法[J].计算机科学,2013,40(5):8-10.

[4] 金安,程承旗.基于全球剖分网格的空间数据编码方法[J].测绘科学技术学报,2013,30(3):284-287.