

# 面向水文数据整合的一体化水文索引模型研究

张 文<sup>1</sup>, 邱文东<sup>1</sup>, 孟诣卓<sup>1</sup>, 陈雅莉<sup>2</sup>

(1. 武汉大学遥感信息工程学院, 湖北 武汉 430079; 2. 长江水利委员会水文局, 湖北 武汉 430010)

**摘 要:**随着水文行业信息化程度的不断提高,对数据的需求逐渐从单一、具体化向多源、综合化方向发展。当前,在水文行业内,对数据按照大类分而治之的管理模式已渐与用户需求脱节,给数据的综合利用带来困难。针对该问题,梳理了现有数据结构与管理模式,提出一种水文数据整合与索引模型(Hydrological Data Integration and Index Model, HDIIM),利用该模型在多个水文数据库与水文应用之间搭建水文数据整合层,建立多库统一的一体化水文索引,实现对多库信息的整合与快速查询。该方法有效屏蔽了多个水文数据库之间的异构性,支持跨库的数据检索与应用,从而提高水文数据的获取与使用效率。实验证明, HDIIM 能够有效提高水文数据的检索效率,特别是针对批量检索操作, HDIIM 的表现较现有方法更加高效和稳定。

**关键词:**水文数据; 信息整合; 水文索引; 树形索引

**中图分类号:** P337-3      **文献标识码:** A      **文章编号:** 1000-0852(2018)05-0006-05

## 1 引言

信息化的发展推动了我国水文行业的巨大变革,水文数据作为重要的基础国情信息,对其存储管理与应用模式提出了新的要求<sup>[1]</sup>。各水文数据库表结构与标识标准使得大部分流域机构基本实现了水文数据的现代化管理,进入新技术应用与试点研发阶段<sup>[2]</sup>。

如今水文数据产品类型日趋丰富,如何利用好已有数据、提高水文数据的存取与共享效率已成为国内外学者与业内人员的主要研究课题。美国<sup>[3]</sup>、欧盟<sup>[4]</sup>、澳大利亚<sup>[5]</sup>等国家与地区均已建成比较成熟的水文数据管理与服务系统,用以满足日常工作与科学研究需求<sup>[6]</sup>。相对而言,我国水文数据管理方面研究尚未完全成熟。其中,史铮铮等<sup>[7]</sup>通过对基础水文数据库的解释和转化,提高了基础水文数据读取的人机交互与分析效率;陆宇庆等<sup>[8]</sup>研发了基于 OLAP 模型的水文数据多维分析系统,提高了大体量水文数据的查询速度与分析能力;曾楷等<sup>[9]</sup>结合 workflow 技术提出了一种水文数据索引模型,实现了水文数据自动化服务系统;王露等<sup>[10]</sup>基于云 GIS 提出了一种水文信息资源共享机制。上述成

果为提高我国水文数据共享与使用效率提供了理论与实践基础,在一定程度上实现了水文数据的有效管理。

本文针对我国水文数据种类繁多、体量巨大的管理现状,提出水文数据整合与索引模型(Hydrological Data Integration and Index Model, HDIIM),以立体的树形结构对水文数据进行整合并建立索引,达到提高水文数据的检索效率与检索易用性的目的。实验证明, HDIIM 方法与传统检索方式相比具有更高的检索效率。应用 HDIIM 模型建立的长江水文数据资源整合与索引系统,能够提供高效的水文数据检索服务,满足水文数据的现代化应用需求。

## 2 水文数据整合模式研究

当前,我国根据水文数据的不同特征,对其采取分库治理的管理方式,并制定了相应库表结构与标识标准<sup>[11-13]</sup>进行维护。此外,由于历史沿革问题,例如个别测站的停测、迁移等,部分数据产生断档与混乱现象。这不符合数据一致性规范,也给水文数据的综合管理与应用带来困难。我国水文数据时间跨度长,地理范围广,涉及的管理机构众多<sup>[14]</sup>,对数据进行批量修

收稿日期:2017-04-19

基金项目:国家重点研发计划“国家水资源立体监测体系与遥感技术应用”(2017YFC0405806)

作者简介:张文(1980-),女,湖北十堰人,博士,主要研究方向为遥感与GIS的水利应用。E-mail:wen\_zhang@whu.edu.cn

改、合并甚至重新测量需要投入大量人力物力,不符合科学发展观理念。因此,如何通过少量数据梳理、建立合理的索引来实现数据的综合管理与应用是不可或缺的工作。水文观测数据大多由测站观测得到,测站是数据获取的来源,其观测位置还赋予了水文信息重要的地理空间属性。测站直接与所观测的水文要素关联,是业务数据库建立的基础。也为水文数据跨库整合提供了可行的思路,即以测站为中心建立各类不同水文观测信息的信息关联模型。

通过对不同水文数据的表结构标准研究发现,以测站为中心,关键信息可分为三类:测站属性、测站编码与观测要素。其中,测站属性是业务操作中所产生的实用属性,例如实际承接的观测任务、被划分到的具体项目等内容;测站编码表征测站的观测类型;观测要素是水文信息的主体,包含于各业务数据库中。观测要素与测站密切相关,是以测站为中心进行水文数据整合工作的重要前提。图1为水文数据整合模式示意图,可以看到,整合与索引层作为中间层分别与水文应用、数据库进行信息传递,改变了水文应用与数据库之间直接交互的传统数据应用模式。此外,整合与索引层还应具有高度可扩展性,能够满足多种水文数据的整合需求。

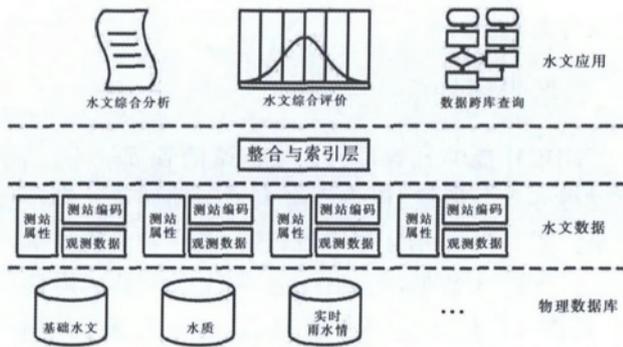


图1 水文数据整合模式示意图  
Fig.1 The model of the hydrological data integration

### 3 树形水文数据索引模型

本文针对水文数据的不同形式与性质,采用三种树状子模型进行管理,分别为动态属性模型(Dynamic Attribute Model, DAM)、测站索引模型 (Station Index Model, SIM) 与观测要素索引模型 (Observation Index Model, OIM)。水文数据与树状模型的对应关系,以及树状模型间的关联如图2所示:

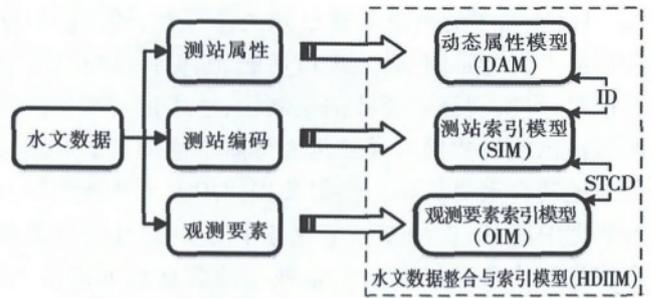


图2 模型示意图  
Fig.2 Diagram of HDIIM

图中 ID 表示 HDIIM 中测站所代表的唯一标识符,STCD 为各原始数据库中的测站编码。

#### 3.1 动态属性模型

测站属性是水文数据检索的重要入口。测站属性分为不变属性与可变属性两类。不变属性是一个测站固有的信息,已添加到业务数据库中,例如管理机构往往不会变更。可变属性则不具有普遍性与确定性,产生于用户定制或项目划分等过程,是根据实际工作需要产生的实用信息。这类信息具有动态性、无序性、不确定性等特点。此外,多数属性还存在子属性。例如,九江站包含于自动报讯站项目,属于水文与水质常规断面测站。此外,九江站还可能包含于其他特定的研究、观测项目。这些可变属性难以通过关系表有效组织,但是对于数据检索有着重要意义。

这里提出动态属性模型(DAM)实现测站属性的动态管理。DAM从结构上为树形集合模型,其节点均为测站集合,代表一种测站属性。子节点属性为父节点属性的从属属性。例如,水文站、水位站和雨量站均从属于水文站网,因此在 DAM 中水文、水位、雨量站将会成为水文站网的子节点。DAM 的中间节点的集合元素由其子属性的性质决定,如果其子属性为相容性质,则取子节点集合交集,否则取交;末端节点直接为测站集合,代表具有该属性的所有测站。DAM 的 UML 类图如图3所示。

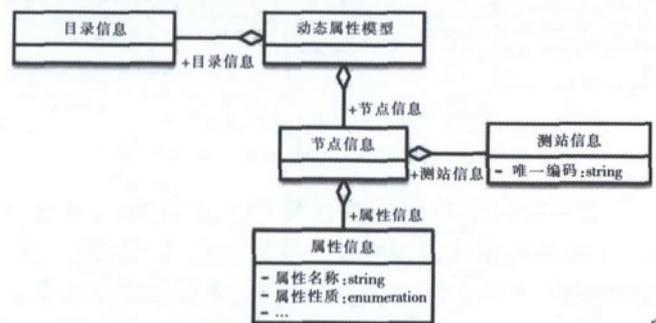


图3 属性管理树 UML 类图  
Fig.3 UML class diagram of the DAM

DAM 除了能有效管理已知测站属性,还支持根据用户需求自行定制属性。通过增删、编辑目录信息与节点信息,可实现测站属性的定制化、动态化。立体的树形结构也使得数据的可视化变得简单可行,树形结构很好地迎合人类认知,使得数据使用人员从抽象的二维数据中脱身,更加专注于业务本身。DAM 提供的集合查询能力能够满足用户多样化的数据查询需求,极大地提高数据检索效率。

### 3.2 测站索引模型

我国水文数据的分库管理,使得同一测站在不同业务数据库中以不同的测站编码(STCD)维护。水文观测过程中的测站停测、迁移等情况,造成数据的断档与混乱,在数据库中产生一站多码、异站同码等问题。此外,个别测站在不同业务数据库中还存在重名、更名、错名等情况。这些问题造成水文数据检索正确率及效率低,亟需水文工作者解决。为此,本文提出测站索引模型(SIM),用以解决上述问题。

SIM 以测站为单位,为实际测站生成三层树形结构,以梳理测站编码。该树形结构的顶层节点为测站的唯一标识 ID;顶层节点的子节点作为树形结构的第二层,用以存储该测站位于各业务库中当前使用的数据标识 STCD;如果在同一业务库中包含有该测站的其他站码,例如历史站码、废弃站码、比测站码等,会在该二层节点下生成第三层子节点。图 4 为 SIM 的 UML 类图。

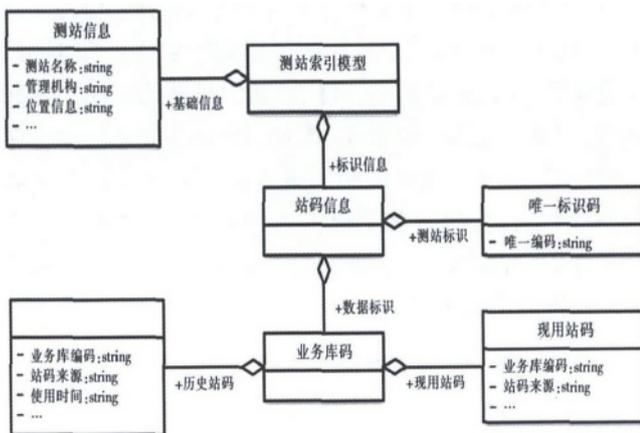


图 4 测站索引树模型 UML 类图  
Fig.4 UML class diagram of the SIM

SIM 的优点在于不需要修改原始数据结构与内容,不需要重新编辑观测要素数据,可直接实现历史数据的整理,并支持现势数据的整合。同时以测站为中心的索引结构加强了不同数据库数据间的联系,便于数据在水文项目上的综合运用。除此之外,该索引从逻辑

上优化了原始数据的数据结构,将平面的关系型数据转变为立体的树形结构,这使得数据的灵活性大大提升,能够满足水文业务数据检索快速、多样化的需求。

### 3.3 观测要素索引模型

观测要素索引模型(OIM)用于索引库内原始水文业务数据,统计各业务库内的观测数据,为它们生成索引信息,包括摘要信息、峰值、覆盖时间等内容,方便用户查询。OIM 以 STCD 为中心,其 UML 类图如图 5 所示。

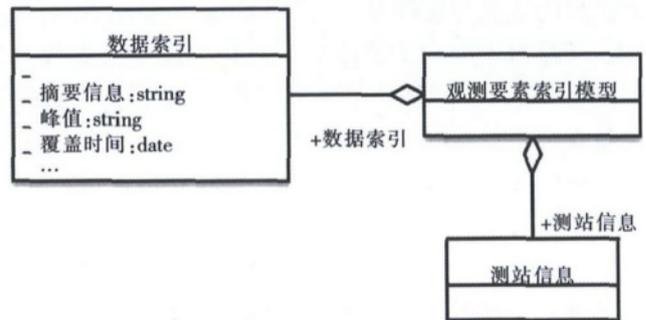


图 5 观测要素索引 UML 类图  
Fig.5 UML class diagram of the OIM

由于观测要素数据是持续更新的,其索引也需要相应更新。这就要求 OIM 保持独立,不会因为其更新而影响到其他子模型。OIM 不改变原始数据结构,仅通过 STCD 与测站索引树关联,保证了各模型相对独立。该结构既满足了用户的易用性需求,又满足了数据的有效组织。

## 4 树形索引抽取方案

HDIIM 能够有效提升水文测站的管理效率,其生成过程亦十分便捷,仅需少量人工干预即可达到预想效果。图 6 为 HDIIM 生成流程图,其完整流程如下:

首先接入数据库,除了注册数据库基本连接信息,还需要构建索引的映射关系。对于数据库结构符合我国现行水文数据库结构规范的直接入库,则自动地建立索引映射关系,开始索引抽取工作。否则,需要数据库管理人员通过用户接口交互地完成索引映射关系构建,这部分工作保证了水文数据整合与索引的可扩展性,不仅能够满足现有水文业务数据库的整合需求,也充分考虑了未来水文业务库的扩展需要。

数据库接入工作完成后,扫描该数据库的测站一览表,检查其测站编码信息。对于信息正确完整的测站,则自动抽取其编码信息,生成/更新对应 SIM;对存在多码、错码、错名等问题的测站,首先按照预定规则对其进行修复、整理,而后工作人员对该操作进行人

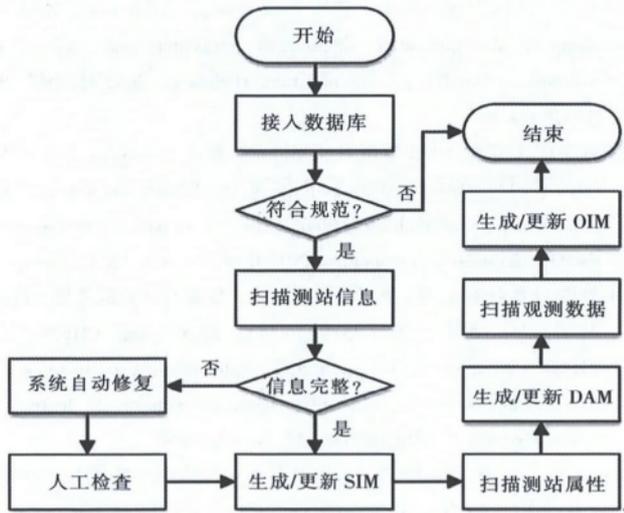


图6 HDIIM 生成流程图  
Fig.6 The workflow of HDIIM generation

工审核。由于测站信息规律已知,系统自检能够自行修复大部分数据问题,减少人工工作量。该步骤不修改底层水文数据,不会影响数据部门日常维护工作。

测站编码抽取工作完成后,分别扫描测站属性信息与观测数据,完成 DAM、OIM 的生成工作。此外,为了应对测站数据的动态性特点,工作人员可以根据自身需求对其进行自定义,随时更新或添加 DAM。

### 5 实验分析与应用系统实现

#### 5.1 跨库水文数据检索效率测试

在效率测试中,设置两个对照组,分别利用 HDIIM 检索与数据库直接检索方式对库中由长江水利委员会水文局直接管辖的全部 495 个测站进行检索测试。实验均在 Windows Server 2008 系统环境中运行。为了测试两种方式对于不同复杂程度目标的检索效率,这里设置两组实验进行验证。其中实验一分别检索 495 个测站的所有业务库站码,用于模拟用户查询测站编码信息,复杂度较低;实验二分别检索 495 个测站的所有在测项目,用于模拟用户查询观测要素信息,复杂度较高。两实验均执行 495 次检索,每次检索均记录其耗时,通过对检索用时的对比验证 HDIIM 索引检索的性能。表 1 所示为部分实验结果,为便于展示说明,我们以 50 次检索操作为间隔列出两种检索方式所用的时间。

从两组实验结果均可看出, HDIIM 的耗时远小于数据库检索方式。随着检索测站数的增加,数据库检索方式的耗时增加更明显。本文实验主要针对两种检索方法的机器效率进行测试。须知,数据检索的实际用时

表1 水文数据检索实验用时数据  
Table1 The time consuming of hydrological data searching

检索次数	实验一		实验二	
	HDIIM 组/s	数据库组/s	HDIIM 组/s	数据库组/s
50	0.2174	1.2440	3.3006	110.6888
100	0.4269	2.2600	6.0795	189.0990
150	0.6450	3.4450	9.1056	318.1379
200	0.8622	4.7540	12.1150	425.3477
250	1.0813	6.0180	15.0421	603.3525
300	1.2846	7.1240	18.8697	776.9532
350	1.4920	8.3680	22.1703	887.6420
400	1.7015	9.3839	24.9492	966.0522
450	1.9200	10.5690	27.9753	1 095.0911
495	2.1147	11.7300	30.7101	1 194.3337
平均用时	0.0043	0.0237	0.0620	2.4128

主要由人工用时与机器用时两部分组成。随着用户需求的变化,工作人员需大量修改检索代码,传统的数据库检索需要工作人员从底层编写 SQL 语句对数据进行检索,需要花费更长的人工用时,使人工检索用时增加。

#### 5.2 水文数据整合与索引系统实现与应用

本文以 HDIIM 为基础,遵循上文所述索引生成流程的基本原理,响应长江水利委员会水文数据高效检索的实际需求,设计并构建了水文数据索引系统。该系统采用 B/S 模式,提供 HDIIM 索引的生成、管理与检索等功能,并支持对生成索引有效性的人机交互检查与修正。系统将 HDIIM 的三个子模型可视化到浏览器页面上,其主要界面如图 7 所示。



图7 长江水文数据资源整合与索引系统主页面  
Fig.7 The main interface of the system

目前,该系统在长江水利委员会水文局已正式上线,在功能、性能以及安全性等方面均令人满意,达到预期效果。此外,基于该平台所建立的 HDIIM 已被长江水利委员会其他检索系统使用,并收到较好评价。

### 6 结论

本文针对水文数据综合利用效率低、检索困难等问题,阐述水文数据的分库管理现状及应对方案,提

出了整合与索引模型(HDIIM)。该模型将测站设为中心,利用三个子模型对各业务库的水文数据进行整合,以立体的树形结构帮助工作人员更快更好地检索水文数据。通过检索效率实验可知,利用 HDIIM 进行水文数据检索的效率比传统方法高出 90% 以上。应用 HDIIM 建立的水文数据整合与索引系统现已上线,并达到令人满意的效果。

目前,HDIIM 仅对测站空间信息进行重组和存储,还未涉及到与其他工具相结合应用。将水文数据索引技术与 GIS 技术相结合,为用户提供更加便捷快速的数据检索服务将会是水文数据管理领域未来的研究方向。

参考文献:

- [1] 曹国荣. 建立我国水文资料共享制度的分析与思考 [J]. 水文, 2008,28(5):78-80+93. (CAO Guorong. How to establish system of hydrologic data sharing in China [J]. Journal of China Hydrology, 2008, 28(5):78-80+93. (in Chinese))
- [2] 史芳斌, 陈雅莉, 陈春华. 关于水文数据库建设问题的探讨 [J]. 人民长江, 2009,40(24):70-72. (SHI Fangbin, CHEN Yali, CHEN Chunhua. Discussion on establishment of hydrological database[J]. Yangtze River, 2009,40(24):70-72. (in Chinese))
- [3] USGS Water Data for the Nation [EB/OL]. <https://waterdata.usgs.gov/nwis/>.2016-12-12.
- [4] WISE. The Water Information System for Europe [EB/OL]. <http://water.europa.eu/>. 2016-11-21.
- [5] Government of Western Australia. The Department of Water [EB/OL]. <http://www.water.wa.gov.au/>. 2016-11-15.
- [6] 张鹏程, 蒋蓉, 张文, 等. 国外水文数据库建设现状研究 [J]. 水文, 2010,30(6):57-60. (ZHANG Pengcheng, JIANG Rong, ZHANG Wen, et al. A survey of hydrology database construction of foreign countries[J]. Journal of China Hydrology, 2010,30(6):57-60. (in Chinese))
- [7] 史铮铮,陈雅莉,张文,等. 面向水文数据的自动化信息整合与分析[J]. 水文, 2015,35(6):42-49. (SHI Zhengzheng, CHEN Yali, ZHANG Wen, et al. Auto mation information integration and analysis of hydrological data[J]. Journal of China Hydrology, 2015,35(6):42-49. (in Chinese))
- [8] 陆宇庆,万定生. 水文数据多维分析系统设计与实现[J]. 信息技术, 2016, (2):47-50+60. (LU Yuqing, WAN Dingsheng. Design and implementation of multidimensional analysis system for hydrological data[J]. Information Technology, 2016,(2):47-50+60. (in Chinese))
- [9] 曾楷,陈雅莉,张文,等. 基于工作流的水文数据自动化服务机制研究与实现 [J]. 水文, 2015,35(5):46-53. (ZENG Kai, CHEN Yali, ZHANG Wen, et al. Research and implementation of hydrological data automation service mechanism based on workflow[J]. Journal of China Hydrology, 2015,35(5):46-53. (in Chinese))
- [10] 王露,王志坚,高祥涛,等. 水利地理信息云平台的探讨[J]. 水利信息化,2015,10(5):6-10.(WANG Lu, WANG Zhijian, GAO Xiangtao, et al. Research of water conservation geographic information service cloud platform[J]. Water Resources Information, 2015,10(5): 6-10. (in Chinese))
- [11] SL323-2011, 实时雨水情数据库表结构与标识符 [S]. (SL323-2011, Structure and Identifier for Real-time Hydrological Information Database[S]. (in Chinese))
- [12] SL325-2014,水质数据库表结构与标识符规定[S]. (SL325-2014, Regulations for Table Structure and Symbol of Water Quality Database[S]. (in Chinese))
- [13] SL324-2005, 基础水文数据库表结构及标识符标准 [S]. (SL324-2005, Standard for Structure and Identifier in Fundamental Hydrological Database[S]. (in Chinese))
- [14] 章树安,吴礼福,林伟. 我国水文资料整编和数据库技术发展综述[J]. 水文, 2006,26(3):48-52. (ZHANG Shu'an, WU Lifu, LIN Wei. Development of the technology of hydrologic data processing and database in China [J]. Journal of China Hydrology, 2006,26(3): 48-52. (in Chinese))

## Research on Hydrological Data Integration and Index Model

ZHANG Wen<sup>1</sup>, QIU Wendong<sup>1</sup>, MENG Yizhuo<sup>1</sup>, CHEN Yali<sup>2</sup>

(1. School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China;

2. Bureau of Hydrology, Changjiang Water Resource Commission, Wuhan 430010, China)

**Abstract:** With the development of information service in hydrological industry, the related applications are calling for more comprehensive and multisource data. The existing data management methods that organize the data separately by its category fail to meet the demand of the users and hinder the integrated application of the data. To address the problem, this paper investigated the state-of-the-art data management methods and proposed a Hydrological Data Integration and Index Model (HDIIM). The proposed model established an integral hydrological data index between the hydrological database and applications by building the data integration layer, which has achieved efficient retrieval and integration. By shielding the heterogeneity of multiple databases, HDIIM can effectively support inter-database applications and data acquisition. The experiments demonstrate that the proposed model significantly promotes the retrieval efficiency, especially for batch operations.

**Key words:** hydrological data; information integration; hydrological data index; tree index