DOI: 10.19364/j.1674-9405.2017.04.007

基于网络爬虫的水利信息检索系统的设计与实现

巫义锐[1,2]，黄多辉[1]，周逸祥[2]

（1. 河海大学计算机与信息学院，江苏　南京　211100；
2. 南京大学计算机软件新技术国家重点实验室，江苏　南京　210093）

摘　要：随着水利信息化进程的快速发展，与水利信息相关的互联网资源不断增多，面对如此巨量和复杂形式的水利信息数据，依靠人工检索、分析的方式已难以满足行业应用的需求。随着大数据信息技术的逐步深入研究，设计与实现可应用于水利信息获取的网络爬虫，成为解决水利信息检索与分析问题的基础。设计应用主题网络爬虫技术的水利信息检索系统，通过水利主题信息爬取，数据格式转化与整理，规范化写入数据库等步骤，实现网络水利数据到格式化数据库数据的自动转化。该系统的实现为多数据源信息的交叉验证与网络应急数据的获取，提供新的思路与可行方案。

关键词：水利信息化；信息检索；网络爬虫；主题爬虫

中图分类号：TP391；TV21　　**文献标识码**：A　　**文章编号**：1674-9405(2017)04-0036-06

0　引言

近年来，随着互联网的快速发展，水利行业与互联网的结合越来越紧密。随着水利信息化与公共信息公开化进程的加快，大量水利信息数据开始在相关网站中，以网页表格形式呈现给公众[1-2]。这些数据往往来源广泛，时效性强。面对这些复杂的网络数据，如何合理地整合与利用，成为水利信息化研究者关注的课题。

传统的水利信息数据收集与检索工作通常依靠人工完成。通过人工采集与整理的水利信息数据往往具有精度高，数据格式规整，可信程度高，但数据量小，来源单一，时效性较差等特点。与此相对应，网络水利信息数据量大，来源广泛，时效性强，但数据格式复杂多变，收集和整理网络水利信息数据需要耗费大量的人力。因此，传统的人工数据采集与整理方法不适用于网络水利信息。在大数据时代，搜索引擎在信息检索方面起着关键性的作用，为人们快速准确地提供所需要的信息。网络爬虫作为搜索引擎的关键组成部分，为信息的准确收集与检索提供了基础[3]。其中，高效率的抓取策略是网络爬虫算法的核心内容，即通过尽可能爬取和用户兴趣相关的网页，提高爬取内容的准确性。1991 年 CHAKABARTI S 等[4]提出了 Focus Project 系统，Focus Project 系统改进了基于关键词表达主题的方式，采用具有相同特性的网页描述关键词。DILIGENT M 等[5]根据网页所占层次的不同，建立一种基于上下文模型的爬虫系统。例如，根网页链接指向第 1 层，以此类推，在每层中依据上下文模型自动生成网页分类器对网页内容进行分类。国内研究网络爬虫的机构也越来越多，例如中国科学院 STIP[6]是一个基于科技文献共享的课题，该系统资源就是采用爬虫实现的。基于网络爬虫技术的蓬勃发展与成功应用案例，提出基于网络爬虫技术的水利信息检索系统，用于大数据网络水利信息的自动采集与整理工作。

基于水利信息检索系统，用户将可以创新性地解决以下问题：

收稿日期：2017-06-13

基金项目：国家自然科学基金面上项目（61370091）；水利部公益性行业科研专项（201501022）；河海大学中央高校基本科研业务费项目（2013/B16020141）；南京大学计算机软件新技术国家重点实验室开放课题项目（KFKT2017B05）

作者简介：巫义锐（1989-），男，四川德阳人，博士，主要从事水利信息化与模式识别研究工作。

1）现阶段所构建的水利信息系统通常具有分布性的特点，这会导致“信息孤岛”现象的存在 [7]。“信息孤岛”指不同水利系统中，对于同一信息，存在不同的数值解释。这种多源数据的不一致性，将极大地降低信息的利用率和效率。针对多源信息不一致的问题，用户能通过水利信息检索系统进行多源数据，特别是不同信息采集渠道数据的收集与整理，进而完成多源数据的交叉验证与整合工作 [8]。

2）针对某些突发的公众性事件，如水体污染等，决策者往往需要在短时间内，得到时效性强的大量相关数据，以利于做出合理的决策。基于水利信息检索系统，决策者能及时通过网络获取水利相关突发事件的多样性信息，进而做出合理判断或决策。

1　总体架构设计

水利信息检索系统总体架构设计是水利信息检索的核心，设计图如图 1 所示。根据网络数据爬取特性与水利表格数据特点，将总体架构设计从左至右分为以下 4 个步骤：

1）水利主题爬虫设计。通过主题爬虫技术判定网页数据与水利主题的相关性，如表格数据与水利相关，则将数据下载至本地，并给予统一编号。

2）数据格式转化。将本地下载的表格转化为通用的数据格式。

3）数据格式整理。根据网络表格的呈现形式，自适应地进行数据的整理与合并，形成规整的数据文档，以利于数据库写入。

4）数据库写入。首先自动建立数据库表，然后为待写入数据添加索引项，最后将规范数据逐条写入本地数据库中，供用户检索使用。

2　方法步骤分析

2.1　水利主题爬虫设计

网络爬虫是一种自动抓取网页并提取网页关键信息的程序，是搜索引擎的主要信息获取渠道。在给定 1 个或多个初始采集点的情况下，网络爬虫从初始网页开始采集，在抓取网页的过程中，不断将新的检测到的网络地址放入待爬行的网络地址队列中，直到满足一定条件（如待爬行队列为空，达到指定爬行数量），停止爬行。在网络爬虫的基础上，主题爬虫按照预定义的爬行主题，应用关键字或主题分析算法，对爬行网页进行内容相关性分析，过滤与主题不相关的网页 [9]。因此，主题爬虫不同于网络爬虫，起始数据采集点必须是预定义的与主题高度相关的页面，主题爬虫仅收集与主题相关的网页。基于主题爬虫的相关特点与水利数据的特性，水利信息检索系统通过以下策略的应用，设计水利主题爬虫，用于网络水利信息的爬取与筛选：

1）主题描述。即如何描述爬取的对象，水利信息检索系统通过字典集合方法定义水利主题。具体而言，收集了 200 个与水利相关的常用业务词语，例如水情、水库、水污染等，并使其构成集合，用于定义水利主题的外延含义。

2）主题爬行策略。主题爬虫需要按照一定的规则抓取网页。具体而言，首先定义起始数据采集点。基于水利信息公开化现状，在现阶段选取了长江航道在线，全国雨水情信息网站，全国重点站点实时雨情、日降雨量和天气网站等 3 类网页作为主题爬虫的起始爬取点。为了适当减少爬取复杂度，采用深度优先的爬取策略，并将深度值限制在合理范围内。此外，由于网页表格中往往含有丰富的数值信息，而这些信息对于水利情况描述具有决定性作用。因此，在爬取过程中仅关注网页中存在的表格信息。

3）主题相关性判断。对于爬取的相关网页，所设计的主题爬虫首先获取页面的文本内容或者页面

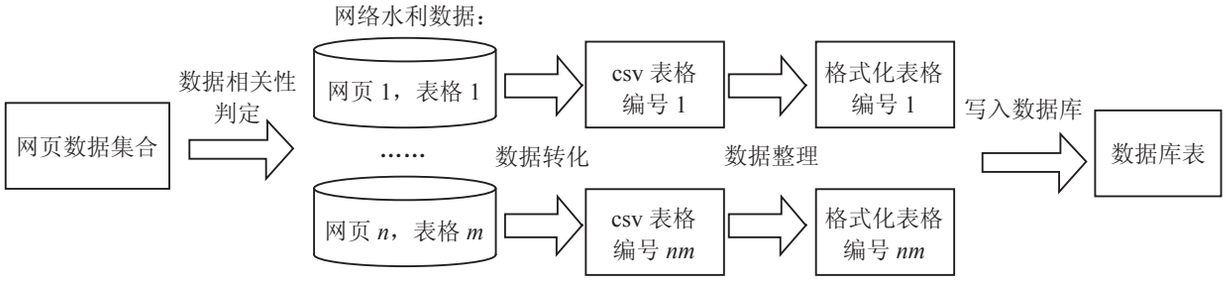

图 1　水利信息检索的总体架构设计图

表格的表头信息；然后，采用关键字匹配方法，与定义的水利主题描述进行对比，以判断所爬取网页是否与水利主题相关。如果数据与水利主题相关，水利信息检索系统将下载相关页面的对应表格到本地服务器，并给予独立编号，编号由网页与网页表格编号 2 部分构成。

2.2 数据格式转化

水利主题爬虫所下载的网页表格数据的格式为 HTML5。HTML5 格式存储的数据成高度离散化状态，即 HTML5 格式对数值与文字描述进行了混合描述。为了后续步骤的顺利进行，水利信息检索系统提出使用数据格式转化，将离散化的数据描述转化为易于提取与处理的通用数据格式，即常用于数据表格描述的 csv 格式。

对 HTML5 数据特征进行观察后，使用基于启发式的算法进行数据格式转化。在对 HTML5 网页所含标签关键字进行处理时，如检测到关键字 table，则将表格数据写入 csv 文档；如检测到关键字 tr，将在 csv 文档中开始写入新的数据行；如检测到关键字 td，将在 csv 文档中写入新的数据列。通过简单的转化规则，水利信息检索系统能够成功地将数据由复杂的 HTML5 格式转化为易于处理的 csv 格式。此外，数据表格的编号将转化为图 1 所示的统一标注，并建立数据内容与编号的对应关系表。在对应关系表中，将具体记录某编号表格的爬取时间、下载表格表头、更新时间等数据来源信息，以供后续的数据库表自动建立与写入步骤使用。

2.3 数据格式整理

由于水利数据的呈现形式没有统一的规范，所获取的标准 csv 文件中存在着大量不规范的数据表现形式。因此，水利信息检索系统提出基于几类常见的表格数据呈现形式，运用自适应检测方法，对数据呈现形式进行整理。

图 2 展示了常用的对水利信息进行描述的表格形式，其中图 a，b，c 中分别展示了完整、缺失表格形式，以及一行多表形式的示例。由图 2 可以发现，数据表格形式复杂多变。为了整理多变的表格数据形式，提出依据以下判断准则进行自适应处理：

1）判断一行是否存在同名项。如存在，则依据某一重复的项，将原数据表分割成多个独立表。该准则主要用于处理图 2 c 所示的一行多表情况。

2）判断某行是否缺失。如存在数据缺失情况，则使用上一行中的同列数据项进行填补。该准则主要用于处理图 2 b 所示的表格情况。

经过自适应处理，能够成功地将非规范的数据整理为规范化数据。规范化数据表中的每一行均为一条完整的水文信息数据。

2.4 数据库建立与数据写入

将整理后的水利数据逐条写入关系数据库将有助于数据的结构化与检索。用户在检索相关数据时，只需设置带关键字的 SQL 查询语句，即可方便快捷地获取相关水利数据信息。为实现水利数据获取与检索过程的自动化，将主要聚焦于建立适应于水利数据写入的数据库。

首先基于对水利数据表格表头信息与首行的分析，建立数据库表。表头信息中一般包含对表格数据信息的概述，例如图 2 a 表格的表头信息为重庆水情信息，将该概述信息作为数据库表名，能够简明扼要地对数据主体进行描述，有利于简化用户后期的检索操作。为此，将表头信息作为数据库表名。

建立数据库表时，还需要定义字段及主键。其中，字段内容来自于对数据表格首行的分析，例如

区(市)县	河流	水文站名	水位/m	流量/($m^3 \cdot s^{-1}$)
江北区	长江	寸滩	161.440	6160.000
北碚区	嘉陵江	北碚(三)	174.520	949.000
武隆县	乌江	武隆	173.000	1670.000
江津区	綦江	五盘	194.150	122.000
石柱县	龙河	石柱	551.900	51.000
黔江区	濯水河	濯河坝	406.540	--
四川省	嘉陵江	武胜(三)	210.580	771.000
四川省	渠江	罗渡溪(二)	204.340	405.000
四川省	涪江	射洪	--	285.000
四川省	涪江	南北堰	102.400	--
潼南县	安居河	泰安	242.620	--
铜梁县	小安溪	虎峰	257.400	0.122

a 重庆江河实时水情表

河名	站名	警戒水位/m	保证水位/m
淮河	息县	41.50	43.00
	王家坝	27.50	29.30
官沙湖	钐岗		
濛进水引河	王家坝闸		
淮河	城西湖闸上		
	润河集(陈)	25.30	27.70
	临淮岗闸上	26.00	28.51
	正 阳 关	24.00	26.50
	鲁台子	23.80	26.10
	峡山口	23.50	25.65
	淮南	22.30	24.65
	蚌埠闸上		23.22

b 安徽实时水情表

地市名	库名	库水位/m	汛限比/m	蓄水量/亿m^3	有效蓄水量/亿m^3	比昨天+增-减/亿m^3	入库流量/(m^3•s^{-1})	出库流量/(m^3•s^{-1})
黄冈	浮桥河	60.87	-4.02	1.67	1.45	-0.01		
	三河口	144.25	-4.75	0.92	0.72	-0.01		
	金沙河	69.57	-2.03	1.20	0.73	0.00		
	明 山	90.62	-2.38	0.95	0.74	-0.01		
	尾斗山	66.63	-2.87	0.63	0.47	0.00		
	白莲河	97.25	-6.75	4.31	2.03	-0.04		
	大 同	115.63	-6.37	1.24	0.89	-0.01		
	花 园	87.87	-4.81	0.48	0.42	-0.00		
	垅 坪	67.68	-5.82	0.58	0.56	0.00		
	天 堂	290.25	-5.75	0.75	0.46	-0.01		
	张家嘴	237.12	-11.88	0.48	0.31	0.00		
	牛车河	73.35	-2.65	0.65	0.35	0.00		
咸宁	南 川	87.56	-14.48	0.23	0.15	-0.00		6
	三湖连江	27.42	-1.08	0.67	0.41	-0.00		

地市名	库名	库水位/m	汛限比/m	蓄水量/亿m^3	有效蓄水量/亿m^3	比昨天+增-减/亿m^3	入库流量/(m^3•s^{-1})	出库流量/(m^3•s^{-1})
襄阳	熊 河	123.49	-1.51	1.14	0.94	-0.00		
	华阳河	134.89	-9.30	0.11	0.09	-0.00		
	云台山	155.24	-7.76	0.59	0.54	-0.00	4.3	9.2
	西排子河	111.16	-0.64	1.12	1.10	-0.02		
	三道河	148.15	-5.85	0.88	0.88	-0.00	5.46	8.45
	石门集	192.91	-2.09	1.04	1.02	0.36	1.79	0.40
	红水河	116.62	-0.38	0.59	0.54	0.00		
	莺河一库	125.84	-6.86	0.33	0.29	0.00		
	孟桥川	134.88	-7.32	0.32	0.29	-0.00		
	寺 坪	312.28	-2.72	2.31	1.29	0.00		
	峡 口	255.06	-9.07	0.86	0.24	0.00	10.6	2.3
	白水峪	193.18	-4.82	0.84	0.37	-0.01	51	127
孝感	郑家河	99.42	-0.88	1.08	1.01	-0.02		
	观音岩	109.20	-0.30	0.68	0.47	-0.00		

c　湖北省大型水库水情表

图 2　常见水利网络信息的表格呈现形式

图 2 a 表格所对应的数据库字段包括区（市）县、河流、水文站名、水位与流量。此外，还将加入一些额外的字段，对数据的爬取属性进行描述，包括爬取与数据更新的时间。其中，数据更新时间来源于网页表头信息分析，部分网页该项可能呈缺失状态，此时将主要依赖于爬取时间对于数据的时效性。

主关键字用于唯一地标识表中的某一条记录。建立适应于数据内容的主键将有助于数据信息去重。在水利信息检索系统中，在数据更新时间项存在时，将主键设置为数据更新时间。假使数据更新时间缺失，则将多数据项作为联合主键。通过如此设计，当爬虫获取重复数据时，数据将因为主键限制而无法写入数据库。数据库中存储的数据将是无重复数据，对于用户的检索使用将提供便利。

在查找到相关水利表格页面时，将数据库建立要素写入数据库脚本，并执行该脚本，以建立数据库表项。在数据库表建立后，将整理好的水利数据内容，逐行写入已建立好的数据库表。水利信息检索系统能够自动将网络上存在的水利数据，通过爬取、数据整理及数据库写入步骤，转化为易于检索的规范化数据。

3　运行实验

将爬取到的内容依据用户的关注程度分为当日与历史信息。其中，当日信息子程序关注于爬取各水利网站更新的每日水利信息，能够保证水利信息检索系统稳定地获取时间连续的网络水利信息。历史信息子程序则提供检索接口用于定向数据查询。对于长江航道在线，全国雨水情信息网站，全国重点站点实时雨情、日降雨量和天气网站，省份雨水情信息公示网站，水利信息检索系统开放时间检索接口；对于欧洲中期天气预报中心，水利信息检索系统开放时间与经纬度检索接口。

基于 C# ASP.NET 平台与 MySQL 数据库系统构建水利信息检索系统。 表 1 是水利信息检索系统在拥有 Corei7 2.2 GHz CPU，6 GB RAM 配置的个人电脑的相关运行时间数据，当日与历史数据下载时间分别代表水利信息检索系统从网络中爬取当日与历史数据的时间。其中历史数据被设定为一日的水利数据，数据整理时间代表水利信息检索系统将爬取到的数据整理为规则化表格数据的时间，数据写入时间代表水利信息检索系统将规则化表格数据写入数据库的时间。在实验中，将网络爬虫的爬取深度设定为 5。值得注意的是，网络爬取数据时间与网络及网页提供服务器的状况高度相关。在某些极端情况下，网页链接可能出现无法访问状况，针对这种情况，设计了重连与失败机制。具体来说，在出现网页链接无法访问的情况时，将在等待一段时间后再次重试访问。假使在一段时间内多次重复访问仍然无法爬取到相关数据，将停止网页链接的访问尝试，并将爬取到的网络链接在系统日志中标为不可访问。此外，省份雨水情信息公示网站所提供的数据为某一省份的平均数据下载、整理和写入时间。

基于表 1 的内容，可分析得到水利信息检索系统中最耗费时间的步骤在于当日与历史数据的下载。因数据整理与数据库写入算法的低复杂度，数据整理与数据库写入所需时间较少。在所爬取的 3 类网站中，发现数据获取总时间长短依次是：全国雨水情信息网站最长，全国重点站点实时雨情、日降

表 1 水利信息检索系统水利数据下载、整理与写入数据库时间 s

起始网站名称	当日数据下载时间	历史数据下载时间	数据整理时间	数据写入时间	数据获取总时间
长江航道在线	1.31	1.27	0.23	0.42	3.23
全国雨水情信息网站	3.82	4.06	0.35	0.53	8.06
全国重点站点实时雨情、日降雨量和天气网站	2.18	3.31	0.28	0.49	6.26
各项任务完成平均时间	2.44	2.88	0.29	0.48	5.85

雨量和天气网站次之，长江航道在线最短。产生这类差异的主要原因是网站结构及网页内容呈现形式的复杂度不同。

4 结语

在综合集成平台支撑下，大数据技术的蓬勃发展导致了水利信息系统的分散特性。如何获取多源信息并进行多源融合及提高传统信息渠道获取的信息时效性成为亟待解决的问题。许多研究者致力于通过数据仓储技术进行多源数据融合与实时更新。通过水利信息检索系统运行的设计与实验证明了基于网络数据进行多源数据融合与实时更新的重要性与可行性。通过网络爬虫与数据整理技术，水利信息检索系统能将水利业务网络信息组件化，最终形成水利业务的描述项。在经过一段时期的积累后，最终可以覆盖整个水利业务应用。通过水利信息检索系统对水利网络信息进行整合，可为用户提供一个可靠、方便、通用的使用环境。通过实际应用，证实了基于网络爬虫的水利信息检索系统可有效解决水利业务应用整合困难的问题，为水利业务应用奠定良好基础。水利信息检索系统的开放性也对水利数据安全性和保密性提出了更高要求，下一步工作将致力于在保证数据安全性与保密性的前提下，解决水利信息化过程中存在的多源数据整合问题。

参考文献：

[1] 莫荣强，艾萍，吴礼福，等. 一种支持大数据的水利数据中心基础框架[J]. 水利信息化，2013 (3): 16-20.

[2] 艾萍，袁定波，边世哲，等. 水利信息化发展状况简要分析方法[J]. 水利信息化，2016 (6): 6-9.

[3] 周德懋，李舟军. 高性能网络爬虫：研究综述[J]. 计算机科学，2009 (8): 26-29，53.

[4] CHAKRABARTI S, DOM B, AGRAWAL R, et al. Using taxonomy, discriminants, and signatures for navigating in text databases [C]// Proceedings of 23rd International Conference on Very Large Data Bases. Burlington : Morgan Kaufmann Publisher, 1997: 446-455.

[5] DILIGENTI M, COETZEE F, LAWRENCE S, et al. Focused crawling using context graphs[C]// Proceedings of 26th International Conference on Very Large Data Bases. Burlington: Morgan Kaufmann Publisher , 2000: 527-534.

[6] 张智雄. Internet 科技信息资源门户网站（STIP）系统的实践研究[D]. 北京：中国科学院文献情报中心，2000.

[7] 严栋飞，陈月妹，张永进，等. 综合集成平台下的多源信息融合及应用整合实例[J]. 水利信息化，2016 (1): 65-68.

[8] 冯钧，佟瑶，陆佳民，等. 跨内外网的数据资源整合与共享关键技术研究[J]. 水利信息化，2016 (5): 1-5，30.

[9] WANG S, ZHOU G M, WANG J. Reviews of relevance algorithm in focused crawler[J]. Computer & Modernization, 2013, 117 (2): 27-30.

Design and implementation of water information retrieval system based on web crawler

WU Yirui [1,2], HUANG Duohui [1], ZHOU Yiyang [2]

(1.College of Computer and Information, Hohai University, Nanjing 211100, China;

2. State Key Labotatory of Computer and Science, Nanjing University, Nanjing 210092, China)

Abstract: With the rapid development of water resource informatization process, the Internet data about water information is growing. Facing complexity and quantity of water information, searching and analyzing with manual work couldn’t satisfy the need of water conservancy industry. Based on the development of big data research, designing and emplying web crawler on water information has been the foundation of solution for water information search and analyzing problem. This paper designs a water information retrieval system based on focused web crawler, which could automatically transform the online water information to formatted database data by online crawling about

water information, data transforming and formatting and properly writing data into database. The proposed system offers a novel and practical solution for cross-validating information from multiple data source and achieving online data for emergency usage.

Key words: water resource informatization; information retrieval; web crawler; focused-crawler

（上接第 35 页）

Research on algorithm for extracting multi-scale drainage network based on TIN

TANG Zhixian

(No.28 Research Institute, China Electronics Technology Group Corporation, Nanjing 210007, China)

Abstract: Multi-scale hydrologic modeling is an important area of hydrological research. Multi-scale river basin is the basis of multi-scale hydrological modeling. While existing extraction algorithm which based on TIN has some shortages such as the definition of river singly, treating only for non-flat areas and un-supporting for multi-scale river basin. This thesis presents a method for extracting multi-scale drainage network based on TIN, defines some features of river basin such as the line, the river valley, the center of gravity - the center of gravity and center of gravity river - the river valley line; determines the flow of the triangular flat areas by using the method of flat back, then determines the river of flat areas further; introduces the concept of watershed area based on TIN to extract multi-scale drainage network. It proposes a binary tree topology with a coding schema to represent the drainage network, providing interfaces for digital hydrology research. The result indicates that the drainage network extracted by this method is basically consistent with the actual river.

Key words: drainage extracting; TIN; multi-scale; gradient

• 简讯 •

水利部组织召开 2017 年国家地下水监测工程项目建设推进视频会

2017 年 8 月 2 日，水利部组织召开 2017 年国家地下水监测工程项目建设推进视频会议，总结水利部国家地下水监测工程项目工作进展和成效，研究分析工程建设中存在的问题，对下一步工作提出明确要求。水利部副部长叶建春出席会议并作重要讲话。水利部水文局介绍了工程项目进展情况，河南、辽宁、湖北、福建、甘肃等省的 5 个单位负责人作了交流发言。

叶建春强调，国家地下水监测工程项目建设是贯彻落实党中央、国务院重大决策部署，保障地下水质量和可持续利用的战略性、基础性工程，具有十分重要的意义。

叶建春指出，在水利部党组的坚强领导下，在各流域机构、各省（区、市）水利、水文部门通力协作、全力推进下，国家地下水监测工程项目建设进展总体顺利。但通过部稽查、审计，项目法人监督检查等，发现还存在一些不容忽视的问题。

叶建春要求，各单位要针对工程项目当前存在的几个方面问题，高度重视，采取以下有力措施，全力推进项目建设实施：1）加快推进工程建设进度。各单位要高度重视，切实加强领导，协调解决好项目建设中出现的问题。要按照目前重新确定的项目建设计划及完成时间要求，组织力量、细化任务、倒排工期，扎实推进项目建设实施。2）确保工程质量。各单位务必高度重视质量问题，切实履行职责，实时跟踪工程进度，检查工程质量。要针对稽查、审计、监督检查中发现的问题，制定整改方案，逐项整改。监理单位要增加人力投入，抓好重点环节和关键工序的监管。3）保证 4 个安全。要深刻认识安全生产形势的严峻性，进一步强化红线意识和忧患意识，强化工程建设、质量、资金管理，确保工程、资金、干部、生产安全。4）加强监督检查。部安监司和各省水利厅切实履行监督检查指导，加强稽查与监管。各级水文部门要加强监督检查，加强施工现场建设监管。5）强化档案管理。要抓好工程档案资料的收集、整理、归档工作，抓紧整改发现的问题。

摘自中国水文信息网