

大数据在洪水分析中的应用前景探究

吴美玲, 杨 侃*, 杨 哲

(河海大学水文水资源学院, 江苏 南京 210098)

摘要: 从上个世纪 80 年代开始开展水利信息化至今, 防汛领域已经积累了大量的数据资源并且数据规模同上个世纪相比已经有了新的突破, 水利大数据已经来临。研究阐述了大数据的基本理论包括大数据基本概念、应用阶段、应用层次 3 个方面, 并以秦淮河流域的洪水分析作为大数据技术的试点应用, 介绍了防汛行业领先的大数据应用平台的搭建, 通过平台实现水利大数据的处理分析, 探讨了大数据在洪水分析中的应用, 分析了大数据技术在应用中带来的成果和效益, 最后对大数据在防汛领域的应用加以展望。通过大数据和洪水分析建设关系的研究, 对推进大数据在防汛信息化的建设具有一定的参考价值。

关键词: 大数据; 防汛; 洪水; 信息化

中图分类号: TV122

文献标识码: B

文章编号: 1007-7839 (2018) 06-0013-07

Exploration on the application prospect of large data in flood control

WU Meiling, YANG Kan*, YANG Zhe

(College of Hydrology and Water Resources, Hohai University, Nanjing 210098, Jiangsu)

Abstract: Water conservancy information has been carried out since the beginning of 1980s. A large number of data resources have been accumulated in the field of flood control and the data scale has new breakthrough compared with the last century, and the large data of water conservancy is coming. The basic theory of large data including the basic concept of large data, the application stage and the application level in three aspects were expounded. Taking the flood analysis of Qinhuai River Basin as a pilot application of the large data technology, the construction of the leading large data application platform in the flood control industry was introduced, and the processing and analysis of the large data of water conservancy through the platform was realized. Besides, the application of large data in flood analysis was discussed, and the results and benefits of large data technology in application were analyzed. Finally, the application of large data in flood control field was prospected. Through the study of the relationship between large data and flood analysis, it's of certain reference value for promoting the construction of flood control information with large data.

Key words: large data; flood control; flood; information

0 引言

近几年来, “大数据”这个词在日常生活中出

现频率非常高。“大数据”在 20 世纪 80 年代就由美国著名的未来学家及社会思想家 Alvin Toffler 首次提出, 然而直到互联网、物联网等信息技术的

收稿日期: 2018-01-22

基金项目: 江苏省研究生科研与实践创新计划项目(SJCX17_0124); 中央高校基本科研业务费(2017B750X14)

作者简介: 吴美玲(1994-), 女, 硕士研究生, 研究方向为水资源规划与管理。

通讯作者: 杨侃(1965-), 男, 教授, 博士生导师, 研究方向为水资源规划与管理。

发展和兴起,关于大数据的关注与研究才逐渐受到重视。

目前大数据的案例主要集中在商业、金融、教育、交通、医疗、农业等领域^[1],关于水利大数据的相关研究和应用比较少。为了推动大数据在水利行业的研究和应用,国务院发布了《国务院关于印发促进大数据发展行动纲要的通知》^[2]。基于国家对大数据分析的扩展应用,探索防汛防旱大数据管理新模式,是传统水利管理模式转变的必然趋势,也是当代水利信息化技术发展的必然选择。

1 水利大数据

1.1 水利大数据基本概念

全球的数据规模已经有了新的突破,根据统计分析,全球的数据规模每年以 40% 左右的速度在增长,并预计 2020 年全球的数据将从 2015 年的 8.6 ZB 增加到 40 ZB^[3]。近几年来随着重大项目(国家防汛抗旱指挥系统建设、水资源监控能力建设等)、专项工作(首次全国水利普查、水资源调查评价等)、以及各项日常工作(水利工程、农村水利、水利规划)的积极开展,目前已经产生并积累了大规模的水利数据。据统计,在 2016 年我国水利数据总量已经达到了 300 TB,并以年增量 80 TB 的速度增长,已经呈现出量大、种类多样、处理速度快和价值高的特征,利用常规的 IT 技术和软硬件工具已经无法在可容忍的时间内对其进行感知、管理、处理和服务,需要新的处理模式代替传统的处理模式^[4]。与传统水利数据相比,水利大数据具有以下特点(见表 1)。

1.2 水利大数据应用层次分析

根据对现有相关的水利大数据的研究分析,可将大数据应用分为 3 个层次^[5],分别为直接应用、关联应用以及预测应用。

直接应用即对大量数据进行统计分析,找出问题的原因或数据的规律并加以应用。刘怀利^[6]等提出利用大数据应用技术获得山洪灾害的相关信息并对数据进行统计分析,找出产生问题的主要影响因子,更加精准地预警和预测山洪爆发,为减少山洪灾害提供技术支撑。

关联应用即对两个或是多个有关联的事物之间进行分析,找出它们之间的关系,通过一个事物来预测另外一个事物。刘雄^[7]提出利用 Map Reduce 模型对渍水点历史降雨强度和渍水点的积水深度的关系进行了回归分析进行编程,从而确定了降雨积水关系模型,将降雨预报数据输入到回归模型中,分析出渍水点的积水情况。

预测应用即通过数据来预测未来,并作出决策来调整当前的行为。杨太萌^[8]提出了利用大数据建立水位预测模型,通过模型来预测未来 6 h 的水位,辅助防汛工作人员在汛期来临时根据预测水位能及时制定防洪减灾相应的对策,减少受洪涝灾害对城市造成地损失。LiD 等^[9]研究基于 POT/GP 模型的水流大数据分析方法,并且将这个方法应用对丹江口水库进行洪水预测。欧洲洪水感知系统 EFAS 实现了对 10 d 的极端天气的预测和洪水预警^[10]。

2 防汛与大数据

2.1 南京水利大数据来源及特征

表 1 水利大数据与传统数据的不同点

对比内容	传统数据	大数据
数据量	MB、GB	PB、EB、ZB
数据格式	数据类型单一,以结构化数据为主	结构化、半结构化、非结构化数据
数据来源	来源简单,主要来自业务系统	智能终端、人际互联、传感设备
产生速度	数据产生、变化速度慢	数据产生、变化速度快
处理对象	仅为数据	解决其它领域问题
处理工具	一种或少数几种	结合多种数据处理工具进行处理
存储设备	使用昂贵的数据存储设备价格	使用价格低廉的存储设备

2.1.1 南京水利大数据体系

南京水利大数据来源范围广泛,包括防办、水务局、水文局等管理单位,数据主要包括水务监测数据、水利工程工情数据、空间地理信息数据、气象数据等类别,形成了南京市水利大数据体系(见图1)。南京水利大数据资源横跨多个业务部门,数据系列长,资源丰富,为进行基于大数据技术的洪水分析提供良好数据基础。整合的数据涵盖了防汛相关信息的各个层面以及其它领域的相关数据,构成非常复杂,数据量也异常庞大。



图1 南京防汛大数据体系

2.1.2 水利大数据特点分析

(1) 体量大

南京市水利信息化建设至今,水利数据在监测、采集、传输、存储、管理和分析等方面的基本设备逐渐得到完善,防汛领域积累的数据量越来越大且每年的数据集都在大量增加。这些数据涉及到水文历史数据、防洪工程设施、地理空间数据、防汛物资数据、辅助信息数据、预报调度业务成果数据和社会经济等静态数据,还包括实时工情数据,水旱灾害数据、防汛物资数据、气象数据、旱情数据、水质数据、预警数据等动态数据。

(2) 分散性

水利大数据分散在各个不同的部门并由其管理,如省厅数据中心、下属单位、城区防办、市供水节水管理处、市政府信息中心等多个部门,因此需要水利相关部门积极主动地参与数据共享的工作中,尽量完善水利大数据。

(3) 异构性

水利基础数据的通常的获取途径如下:流域内各个水文监测站点、先进的遥测技术、复杂的物理实验等手段,每个手段获得数据形式各不相同。并且水文数据分别由中央、各流域机构等各个水利部门单独进行管理和存储,且各个部门所采用的数据管理平台以及数据库系统也是截然不同,导致各类水文数据结构存在明显的差异,为

数据整合汇集带来的巨大的工作量。

(4) 周期性

水文现象具有一定的周期性,如受到地球公转的影响,使水文现象具有年周期性;受到地球自转的影响,水文要素具有日周期性;受到月球的影响,使潮水位进行周期性运动(潮汐);受到季节的影响,每年南京市在6月将进入汛期到9月底结束。

(5) 交叉性

水利大数据不仅仅涉及到水资源、水质、防汛抗旱、水行政、水污染等方面的数据,而且还涉及到气象、地理空间、社会经济等方面的数据。

(6) 价值高

水利数据对于水行政部门的规划、建设、管理和社会各界开展与水利相关的活动,以及满足人民群众对水资源信息日益增长的需求具有重要的价值。

2.2 水利大数据应用平台搭建

运用大数据的相关技术,搭建了南京市水利行业领先的大数据应用平台(见图2),开展对分布异构多源的数据进行整合挖掘分析利用,建设高端创新应用,形成具备水利应用系统特色的水务系统应用大数据分析环境。

构建大数据分析平台,并以秦淮河流域为本次大数据分析对象,并基于大数据平台实现秦淮河流域的洪水分析应用,主要实现洪水事件信息多维关联与展示,洪水分析、防洪调度的匹配分析。

2.2.1 大数据分析平台

利用市水务局旧有和新购置的服务器构建分布式大数据应用硬件平台,配置大数据存储、分析、处理等相关平台软件,建立基于HDFS/Hadoop的分布式存储平台,Spark的分布式并行处理环境,构建市水务大数据应用平台。根据南京市防汛防旱指挥系统实际情况并结合大数据支撑服务需要进行二次开发。图3是大数据分析技术架构。

2.2.2 数据流程

水利大数据处理流程如图4所示。

2.2.3 平台功能设计

构建市水利大数据应用平台,并以其为核心,构建大数据等相关的创新应用。从数据存储、数据挖掘分析到应用,整个大数据分析平台基本功能如下:

(1) 海量水利数据存储

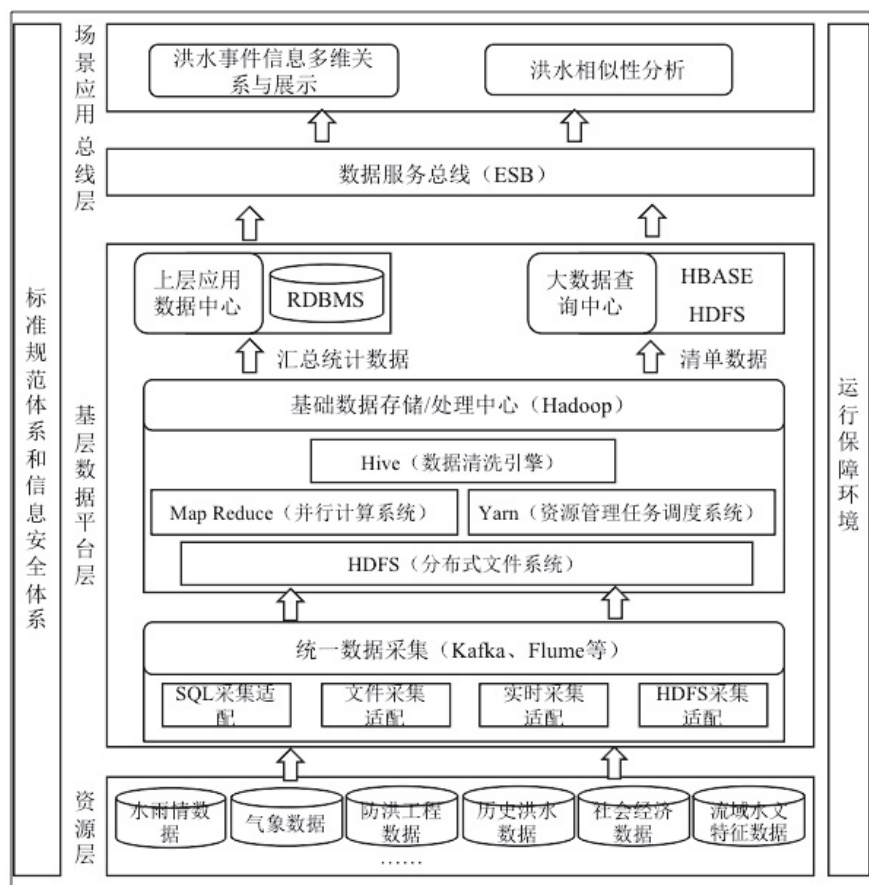


图2 南京水利大数据应用框架图

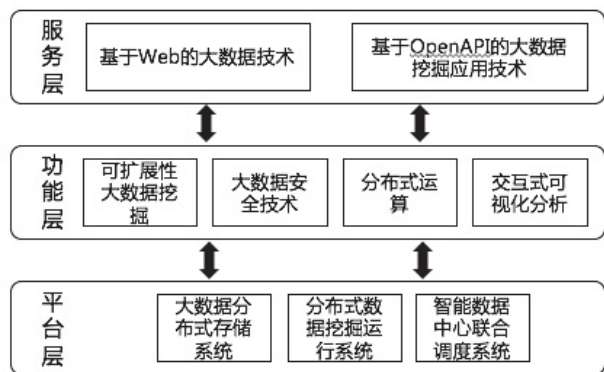


图3 大数据分析技术架构

传统的关系型数据库，适合存储结构化的水利数据，而在进行存储海量的半结构化和非结构化数据过程中暴露出其在性能、扩展性、安全性等方面的不足。Hadoop 提供 HDFS 分布式文件系统和 HBASE 非关系型数据库实现海量异构的防汛数据存储。其中 HDFS 分布式文件系统的冗余备份机制提供了很高的可靠性保证，HBASE 是基于 HDFS 分布式文件系统作为底层数据存储，利用分布式计算的优势，高效地对数据进行随机读

写访问，具有很高的容错性。并且任何数据存储到 HBASE 中的都会带有时间戳，很适合存储传感器采集的相关水雨情信息。

(2) 大数据挖掘

南京水务数据中心为系统提供南京水利大数据存储及数据计算平台，通过平台服务器提供系统调用功能。门户服务中心将整合数据中心存储的数据以及计算资源，并通过 Open API 服务器和 web 应用服务器以 Open API 调用和 web 调用的方式提供大数据存储、管理及挖掘等服务。采用 Spark 组件进行大数据分析，数据挖掘模型以 Spark On Yarn 的方式构建大数据分析引擎。通过调用聚类分析、回归分析、关联规则分析、时间序列分析等算法，进行南京水利大数据分析计算。

2.3 水利大数据应用

以南京市秦淮河流域作为研究区域进行大数据洪水分析应用，提取实时水雨情数据、气象数据、水文历史数据、地理空间数据等海量数据作实时洪水的大数据分析和预测。

基于构建的秦淮河实时洪水业务场景，利用

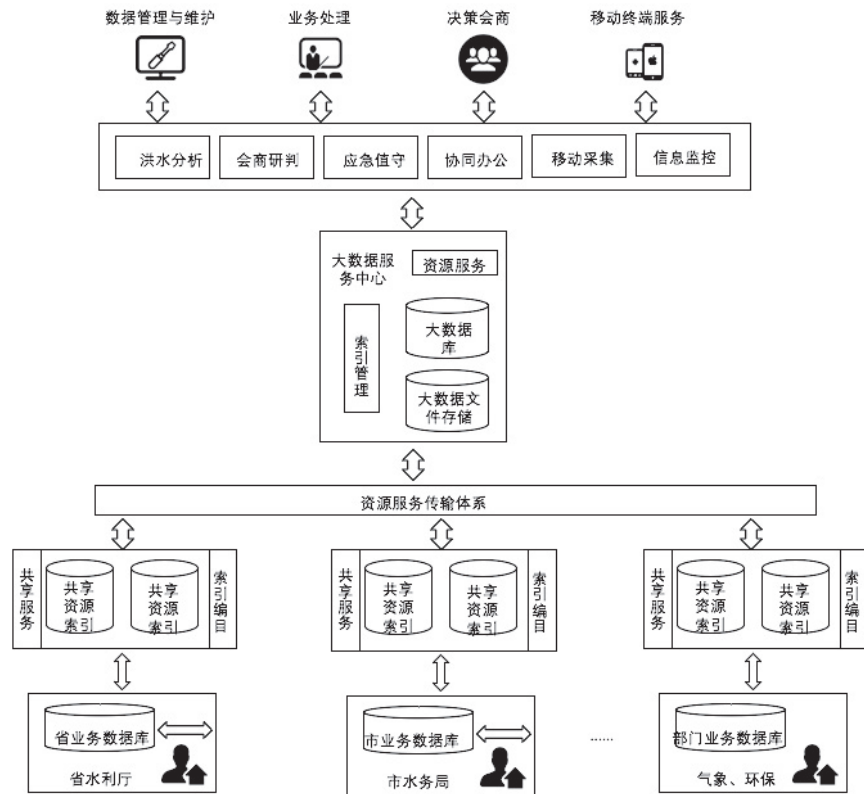


图4 大数据应用平台数据流程

大数据应用平台进行分析:基于分布式服务器集群,通过大数据应用开发平台的数据接入服务实现过程数据采集、传输、存储和加工处理服务。应用大数据挖掘计算服务结合 Spark 分布式计算技术,实现历史数据的相似性分析,通过与历史数据的自动匹配得到秦淮河流域洪水分析成果,为防汛指挥人员能做出科学、合理的防洪调度决策提供信息支持。

2.3.1 洪水事件多维关联

结合大数据分析平台,对当前洪水事件进行信息多维关联分析与展示,从气象、降雨、水情、工情、积水、应急响应、险情等洪水事件的不同维度进行分析与展示,功能结合防汛指挥大屏进行可视化构建,满足实时防汛应用场景需求。图5是基于大数据分析平台对当前洪水事件进行信息多维关联分析与展示的效果图。

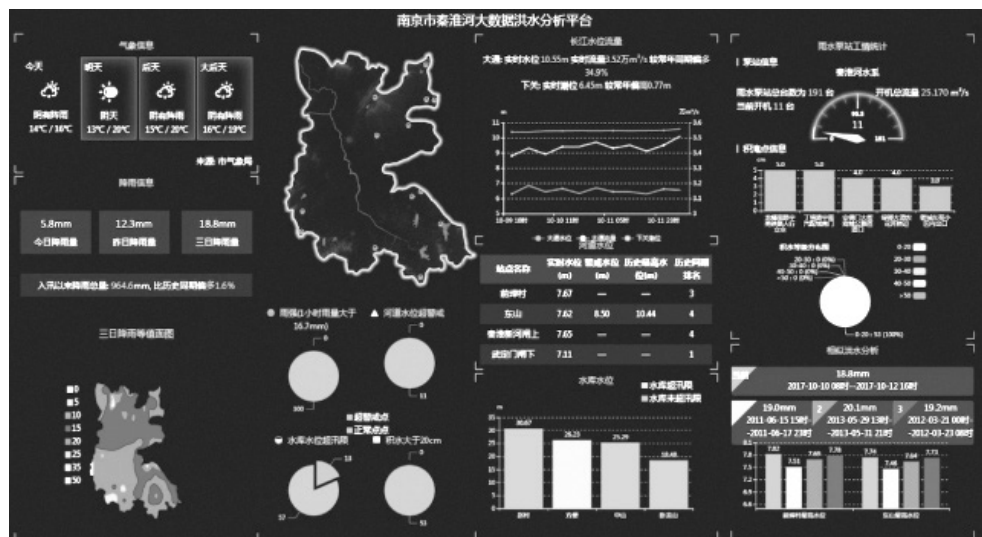


图5 南京市秦淮河大数据洪水分析平台效果图

图5 南京市秦淮河大数据洪水分析平台效果图,系统提供的洪水事件信息的多维关联展示支持防汛指挥人员24小时不间断地对实时汛情信息、实时工情信息、预警信息、险情信息、灾害信息在线监测预警(见图5)。其中,实时汛情信息包括实时雨量、河道水位、流量,水库水位等水雨情信息;实时工情信息主要是泵站的实时运行工况信息;预警信息包括水情监测站点监测预警等;险情信息主要是积水险情信息;灾害信息包括城市受淹情况统计。系统利用图、表的交互方式,为防汛人员提供全面、高效、实时、直观地汛情信息服务,便于其正确分析和判断当前汛情、险情等信息,并制定科学的防汛调度方案。

2.3.2 洪水相似性分析

结合历史洪水信息,利用洪水相似性分析方法,对选定洪水进行历史相似洪水在线分析,根据分析结果进行相似度排序,查看不同相似度情形下历史洪水信息,包括发生时间、发生过程中断面水位对比情况,分析工程历史调度方案采用的控制条件、历史相似洪水相关统计信息等。将实时数据和大数据分析结果利用可视化技术,以图形的形式进行展示。

(1) 洪水相似性分析方法

洪水相似性分析是选取流域主要控制站的水位、流量等为特征指标,采用系统相似性度量中的相关系数法^[11]和非平均距离系数法^[12],建立洪水相似性分析方法,在选定主要控制站后,从数据库中获取这些控制站的历史场次洪水信息,与当前洪水信息进行对比分析,并推荐出相似的历史场次洪水。基于历史洪水资料,利用洪水形成过程中的相似性进行分类以达到更为合理的分类结果。其中,洪水过程相似分析运用相关系数法和加权距离系数法,计算实时洪水与历史洪水的相似系数,进行综合定量分析。

假设两场洪水 X 和 Y 有 n 个相关的特征指标,采用相关系数法可计算 X 和 Y 在同一指标下的相关系数,即:

$$R_{xiyi} = \delta_{xi} \delta_{yi} / \delta_{xi} \delta_{yi}, i=1, 2, \dots, n \quad (1)$$

式中, δ_{xi} 、 δ_{xi} 、 δ_{yi} 分别表示 X 、 Y 系列的均方差。

$$S_{xiyi} = |R_{xiyi}| \quad (2)$$

式中, S_{xiyi} 则是第 i 个特征指标的相似系数。

若有 n 个特征指标则有:

$$S_{xy} = \beta_1 S_{x1y1} + \beta_2 S_{x2y2} + \dots + \beta_n S_{xnyy} \quad (3)$$

式中, β_n 为 X 、 Y 系列在指标 n 下的权重,采用二元比较模糊决策分析法来确定指标权重^[13]。

若计算 m 场洪水之间的相似系数,可将每两场洪水之间的相似系数求和并取其平均值,即:

$$S = \frac{1}{m(m-1)/2} \sum_{m \geq g > h \geq 1} S_{gh}, 0 < S \leq 1 \quad (4)$$

式中, g 、 h 分别为场次洪水的序列号。 m 场暴雨洪水过程之间的相似程度,通过 S 的值来衡量, S 越大 m 场洪水过程的相似程度越高。

加权距离系数法为系统相似性度量的一种方法,根据 X 和 Y 两场洪水间的 n 个特征指标建立一个 n 维欧氏空间,定义 X 和 Y 之间的距离为:

$$d_{xy} = (\beta_1 \sum_{j=1}^k (x_{1j} - y_{1j})^2 + \beta_2 \sum_{j=1}^k (x_{2j} - y_{2j})^2 + \dots + \beta_n \sum_{j=1}^k (x_{nj} - y_{nj})^2)^{1/2} \quad (5)$$

式中, β_i 为第 i 个特征指标的权重,且满足 $\sum_{i=1}^n \beta_i = 1$ 。

则 X 和 Y 的相似系数为:

$$S_{xy} = 1 - d_{xy} \quad (6)$$

结合大数据平台的计算结果,对洪水进行在线分析。

(2) 洪水相似在线分析

结合历史洪水信息,考虑降雨总量、最高水位、干流控制断面(选取前埠村、东山等断面进行分析)水位过程、等主要相似因子,利用洪水相似性分析方法,对选定洪水进行历史相似洪水在线分析,根据分析结果进行相似度排序。可查看不同相似度情形下的历史洪水信息,包括历史洪水发生时间、发生过程中主要控制断面水位、流量等相关洪水统计信息。通过查看相似度较高的历史洪水水位、流量的未来发展趋势,来预测当前洪水未来几分钟、几小时内洪水的可能的发展趋势,为防汛指挥人员做出科学、合理的防洪调度决策提供信息支持。随着时间的推移,流域内各地的实时降雨量、河道的水位、流量等数据得到了更新。系统会利用这些数据信息重新进行洪水相似性分析,获得与当前洪水相似的历史场次洪水。

图6是对2017年8月13日发生的洪水进行洪水相似性分析,以面雨量、重要控制断面的水位流量(前埠村、东山、秦淮新河闸上、武定门闸下和大通水位、大通流量以及下关潮位)以及作为指标。利用上述的公式(1)~(3)分别计算出当前面雨量、控制断面的水位流量与历史洪水资料的相

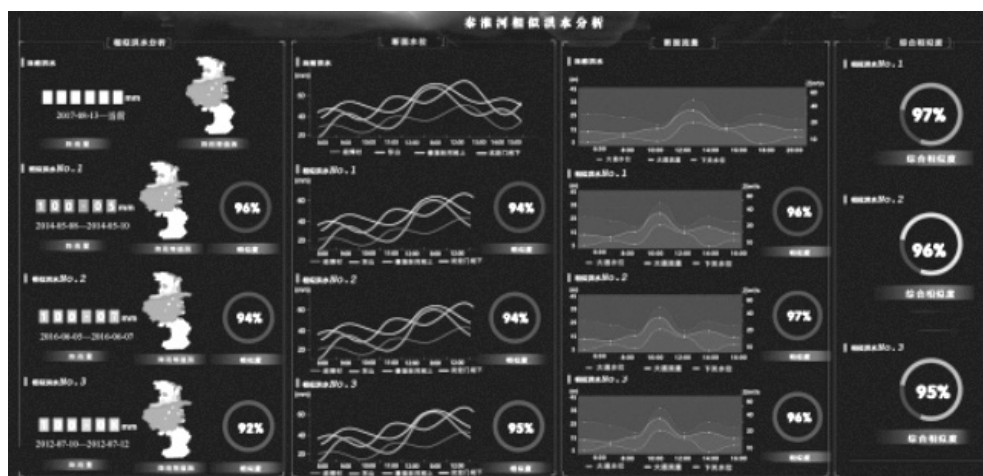


图6 南京市秦淮河2017年8月份的相似洪水分析效果图

似度,并利用公式(5)~(6)计算出相应的洪水综合相似度,系统根据综合相似度大小选出最高的前3场历史洪水。基于大数据平台,获得了与本场次洪水综合相似度分别达到97%、96%以及95%的3场历史洪水。

基于大数据技术匹配出与当前洪水相似度较高的前3场历史洪水,从数据库中获得历史洪水信息,包括洪水发生时间、发生过程中河道断面水位等情况,分析工程历史调度方案采用的控制条件、历史相似洪水相关统计信息等,为防汛指挥人员做出科学、合理的防洪调度决策提供信息支持。

3 成果及效益分析

对南京市水利大数据进行了初步的整合,保证了“一数一源”的管理机制,推进全市水利信息资源的建设。建立的大数据应用平台,对当前洪水事件进行信息多维关联分析与展示,为防汛工作人员提供了全面、高效、实时、直观地汛情信息服务,便于防汛人员正确分析和判断当前汛情、险情等信息,并制定科学的防汛调度方案。

开展秦淮河流域的大数据洪水分析应用,使南京成为区别于其他城市或流域的大数据技术应用用于水文分析的领跑者,是传统方法结合大数据新技术在南京防汛减灾应用的先进尝试,辅助防汛减灾管理工作,提高防汛管理科学水平。利用智能化的大数据技术进行洪水分析,推动了灾后救灾向灾前预防的转变,使防洪工程在汛期能够充分发挥防汛减灾的作用,减少洪涝灾害带来的

影响。

4 结语

大数据技术在防汛领域的应用保证防汛工作的高效率,实现南京市防汛信息地全面汇聚、及时预警、科学决策和高效调度。从数据角度实现了市水利大数据体系的搭建;从框架角度,实现了实现传统防汛管理向现代化管理的跨越式发展;从业务应用角度,基于大数据的洪水分析应用的建设,实现洪水事件信息多维关联分析与展示,探索大数据技术在南京防汛减灾管理领域的应用。大数据技术在防汛领域的成功应用,可以为大数据在“智慧水务”以及“智慧城市”的建设提供大量的基础设施和有价值参考信息。

关于水利大数据的研究的还处在初级阶段,为了深入充分挖掘数据,发现其中的价值,还需要继续进行工作,扩大并完善水利大数据体系;探索当前大数据分析(神经网络、归类分析、聚类等)在水利领域中的应用;探讨深度学习、社会计算、知识学习和大数据可视化等大数据的热点研究方向在水利领域中应用,为防汛减灾活动带来更多的益处。

参考文献:

- [1] Fang Z Y, Fan X W, Chen G. A study on specialist or special disease clinics based on big data. *Frontiers of Medicine*, 2014, 8 (3): 376-381.
- [2] 国务院. 国务院关于印发促进大数据发展行动纲要的

(下转第24页)

- 通知(国发[2015] 50 号)[A]. 北京: 国务院, 2015:15 .
- [3] 曹冲 . 北斗产业引领中国的大数据时代 [J]. 数字通信世界, 2013 (8):8-9 .
- [4] 夏海元 . 面向 Big Date 的数据处理技术概述 [J]. 数字技术与应用, 2012 (3):179-180 .
- [5] 沈可 . 大数据应用的认知和思考 [J]. 科学:上海, 2017, 69 (1):38-40 .
- [6] 刘怀利, 马浩, Liu Huaili, 等 . 安徽省山洪灾害监测预警管理中大数据的研究与应用 [J]. 中国防汛抗旱, 2017 (2):26-29 .
- [7] 刘雄 . 大数据技术在城市洪涝灾害分析预警中的应用研究 [D]. 武汉: 华中科技大学, 2015 .
- [8] 杨太萌 . 基于大数据的城市防汛决策支持系统研究 [D]. 杭州: 浙江大学, 2016 .
- [9] Li D, Guo S, Yin J . Big data analysis based on POT method for design flood prediction[C] . IEEE International Conference on Big Data Analysis, 2016:1-5 .
- [10] 赵琳娜, 刘莹, 党皓飞, 等 . 集合数值预报在洪水预报中的应用进展 [J]. 应用气象学报, 2014 (6):641-653 .
- [11] 欧晓凌, 裘刚, 张毅, 等 . 城市交通流信息相似性分析与研究 [J]. 中南公路工程, 2003, 28 (2): 4-7 .
- [12] 章四兵, 周美立 . 系统相似性度量中的非平权距离系数法 [J]. 合肥工业大学学报(自然科学版), 2004, 27(8): 903-906 .
- [13] 大连理工大学, 辽宁省供水局 . 大伙房水库防洪预报调度方式及其规则设计与应用研究技术报告 [R]. 大连: 大连理工大学, 2007 .