【本文献信息】包力伟．基于大数据的互联网视听节目监管系统建设[J]. 广播与电视技术，2020,Vol.47(9).

基于大数据的互联网视听节目监管系统建设

包力伟

（江苏省广播电视监测台，江苏 210001）

【摘　要】本文梳理了江苏省广播电视监测台建设的互联网节目监管系统的整体架构。系统基于大数据，运用了智能识别调度技术，结合了大数据采集、大数据搜索、图像识别处理、自有作业调度等关键及创新技术，实现了数据采集、传统视听监测、敏感舆情监测、综合业务管理等诸多功能，有效维护了省属地内清朗的网络空间。

【关键词】互联网，大数据，视听节目，监测监管

【中图分类号】TN915.4　　【文献标识码】B　　【DOI 编码】10.16171/j.cnki.rtbe.20200009025

Construction of Internet Audiovisual Program Monitoring and Supervision System Based on Big Data Technology

Bao Liwei

(Jiangsu Broadcasting and Television Monitoring Station, Jiangsu 210001, China)

Abstract　This paper thoroughly analyzes overall architecture of Internet Audiovisual Program Monitoring and Supervision System, which is constructed by Jiangsu Broadcasting and Television Monitoring Station. The system is based on big data technology, uses intelligent recognition and scheduling technology, integrates key and innovative technologies such as big data acquisition, big data search, image recognition processing, and self-owned job scheduling. Many Functions such as data acquisition, traditional audiovisual monitoring, sensitive consensus monitoring, and integrated business management are realized in this system, which effectively maintains a clear network space in the province.

Keywords　Internet, Big data, Audiovisual program, Monitoring and supervision

0 引言

当前，互联网用户规模和技术都已经进入快速发展的时期。互联网发展重心从“广泛”转向“深入”，网络应用对大众生活的改变经历了从点到面的过程。互联网对用户在生活中全方位渗透的程度进一步增加,从而推动了用户生活的不断“网络化”。随着互联网的快速发展，网络视听节目传播愈发多样化，互联网视听节目的服务活动越来越活跃。

1 项目背景

根据广电总局对于互联网等新兴媒体的监测监管要求，江苏省广播电视监测台改进监管方式，突破台内多年前的旧有互联网监管系统功能上的局限，重新开发建设一套针对省内互联网站点播出视音频节目做到全面发现、监测和管理的互联网视听节目监管系统。系统要实现以下六个目标：一是实现对互联网视听节目实时准确搜索、分析、判别、排重和研判，实现对包括视听节目在内的各类海量节目采集和数据快速分析处理；二是能够准确掌握管辖范围内视听网站各类节目传播的数量、动态、范围、影响等情况，实现实时动态监测；三是通过疑似违规分析、人工审核与取证的日常处理流程，实现违规视听节目的下载取证和自动分类、专项任务的紧急处理；四是实现热点视听节目传播和分析；五是实现日常互联网视听及舆情监控情况的报表统计；六是实现系统的综合管理，使之融入我台全媒体一体化监管维护体系中。

2 系统综述

互联网视听节目监管系统遵循总局第 56 号令《互联网视听节目服务管理规定》要求，依托江苏省属地 ICP 域名网站数据和分布式采集爬虫、深度学习的图像分析、关键词分析、大数据存储分析技术对省属地的视听节目相关的舆情信息进行监

测监管。系统采用模块化结构设计，具有良好的可扩展性，技术成熟，稳定可靠，实用高效，操作便捷。通过 24 小时不间断扫描指定区域，及时发现各视听节目网站、视听节目和传播源，从而做到对互联网视听节目传播主体、传播内容和传播活动的全面监管。首先，系统对省内的持证网站进行重点监测，及时了解持证网站数据更新情况，根据关键词分析技术发现重点人或重点节目是否在该持证网站上传播。其次，系统对省内非持证网站进行监测，通过分布式爬虫、深度学习的图像分析、关键词分析技术自动发现传播视听节目或色情违规内容的网站，完成属地网站上视听节目相关舆情信息排查，发现违规信息，净化属地网络环境。

3 系统整体架构

3.1 系统逻辑结构

系统逻辑结构包括硬件云平台、统一数据中心、互联网视音频监测系统几大部分，以及配套的安全管控功能、用户管理功能、数据共享服务、消息和日志服务等模块。系统逻辑结构如图 1 所示。其中，统一数据中心是其中的关键，包括四大模块：统一采集引擎、统一计算和存储引擎、数据加工和检索引擎以及数据分析管理引擎。统一采集引擎由数据采集系统和对外接口构成，对指定类型目标网站进行自动采集，包括互联网文字、音视频；数据采集系统主要由任务调度中心、代理 IP 绑定机制、要素采集配置和管理后台组成。统一计算和存储引擎主要包括分布式文件系统、NoSQL 数据库、关系数据库和离线计算、实时计算等多种分布式计算引擎等。数据加工和检索引擎由数据加工和检索系统及对外接口构成，可以对采集的数据提供 ETL 服务和检索服务，并结合数据分析管理引擎完成数据加工和数据分析功能，实现包括实体抽取（节目名、地名、机构名等）、自动去重、自动摘要、信息分类、内容元信息抽取管理等功能。数据分析管理引擎通过完成自然语言处理、图像识别和机器学习算法，实现包括话题聚类、热点 / 热词计算、正负面情感分析、传播溯源分析、各种专题分析的功能。

3.2 系统硬件架构

系统构建在我台自主搭建的一套在线虚拟化云平台上，满足基于互联网业务监测监管系统的上线要求，为全台提供基础计算资源和存储的支撑。通过对系统建设的总体目标分析，系统硬件组成包括虚拟化后的 WEB 服务器、采集服务器集群、接口服务器、音视频分析分布式集群、海量数据存储服务器集群、中文语言智能分析处理服务器，以及实体数据库服务器、交换机、防火墙、IPS 等网络设备。系统虚拟化后的硬件架构如图 2 所示。采集服务器通过分类，包含了重点网站搜索、播客节目搜索、全媒体视听节目搜索等业务。WEB 等各类服务器虚机通过 Vmware 管理平台实现热备份、热迁移；数据库服务器通过热备软件实现热备份。

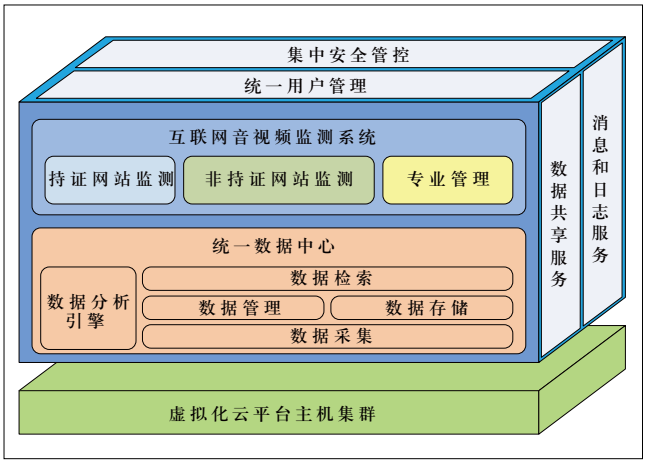

图1　互联网视听节目监管系统逻辑结构图

3.3 系统业务流程

系统通过人机结合方式实现对我省互联网网站的日常监测。首先是对省内所有持证网站进行重点监测，将持证网下发到持证网站采集爬虫，通过定制的采集模板采集持证网站上传播的节目信息。其次是依托我省属地 ICP 域名库，通过核心词匹配技术自动发现属地范围内疑似传播视听节目的网站，再将疑似视听网站下发到非持证网站采集爬虫，抓取非持证网站的节目信息。第三是通过图像和文本分析技术自动发现疑似违规的节目，通过人工方式查看节目是否违规，支持人工录屏方式进行违规节目取证，并支持生成电子文案后发送给上级部门进行查处。系统业务流程如图 3 所示。

4 系统功能

系统基于 B/S 结构，使用浏览器进行登录。系统包括数据采集部分、传统视听监测部分、敏感舆情监测部分、综合业务管理部分等。

4.1 数据采集

系统数据采集主要采用多台搜索机并行抓取的形式，对监管范围的所有音视频网站进行全网搜索，并对重点网站的视音频进行增量抓取，同时抓取相关的文本信息，对爬虫搜集到的视听节目进行下载取证。系统采集数据范围覆盖互联网全媒体包括新闻、论坛、博客、微博、电子报刊等媒体类型的数据收集。系统实现数据采集主要分为以下四个模块。

4.1.1 采集控制与调度

采用集群服务器来完成对网络的分布式搜索，建立统一任务调度平台，将大量复杂搜索和下载任务分派到不同的节点，

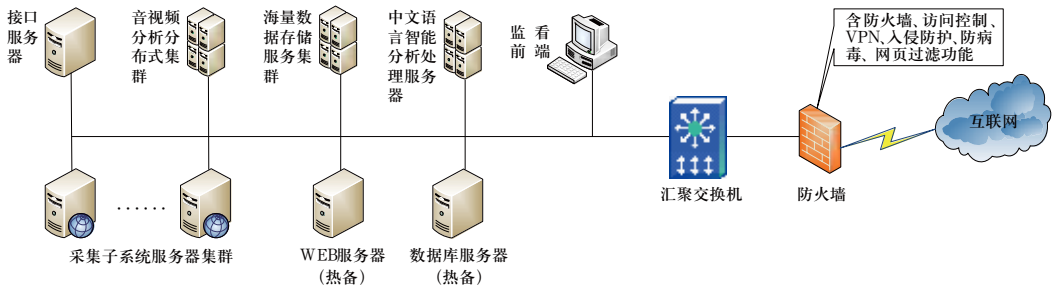

图2　互联网视听节目监管系统硬件（虚机）架构图

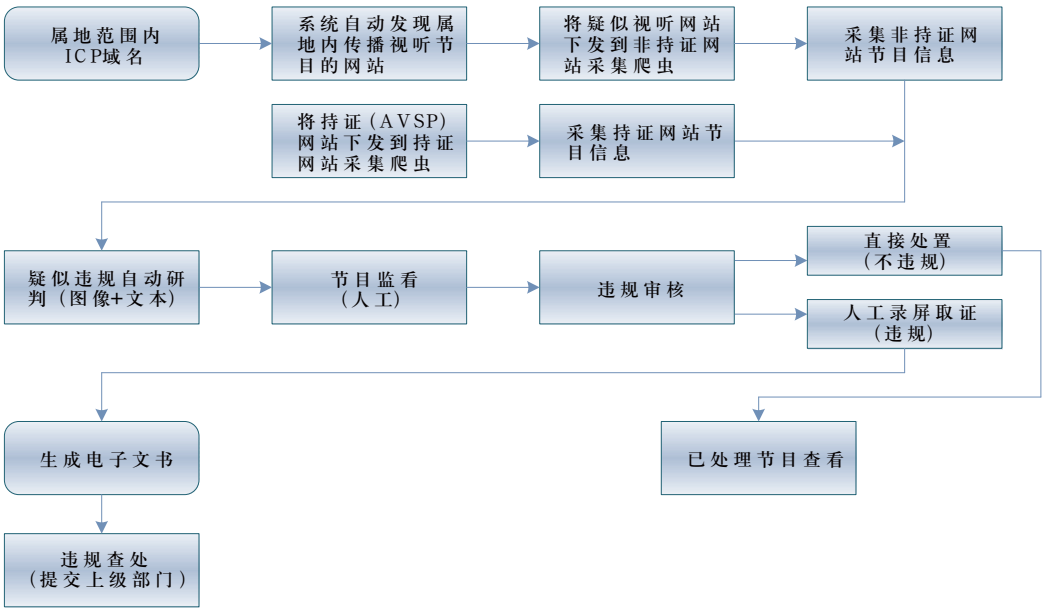

图3　互联网视听节目监管系统业务流程图

实现负载平衡，使资源利用率达到最大。

4.1.2 网页快照采集

系统搜索服务器采集相关网页的文本信息，作为快照保存，并与相应的节目进行关联。

4.1.3 代理 IP 池策略

由于系统监测的目标网站可能具备自我保护机制：当某一个 IP 地址频繁访问该网站的时候，目标网站就会对用户访问做一些访问频次限制，超过设定即拉黑。针对这种情况，采集通过调用云中心的代理 IP 池，用不同的 IP 地址以轮询的方式访问该网站，确保了数据抓取成功，也保证了监管目标无遗漏。

4.1.4 采集管理系统

网站的采集监测模块监测各节点目前的调度模式、总调度次数、发现资源数量、采集资源数量、所耗时间及任务状态等各种关注的数据。

4.2 传统视听监测

监管系统发现属地网站后，通过种子部署，实现广度采集，从而发现节目。发现的节目通过和常规任务中设置的关键字进行匹配，将该节目打上“疑似违规”的标志，然后通过人工审核、取证（截图或录屏等）进一步置为违规节目。与此同时，系统将该节目所在的网站自动置为违规网站。最后，系统以网站、节目的角度生成各种报表。系统监管页面如图 4 所示。

4.2.1 网站监管

网站监管功能主要对各类网站进行详细的监控，实现“发现传播源头、追踪传播内容、监测传播主体”的职能。网站监管主要包括网站发现、网站监控、网站管理、网站查处、效果验证、已处理网站和统计报表等模块。

4.2.2 节目监看

违规节目一经发现，节目监看模块随即进行监测、检查、取证、研判。流程上具体包括节目查询、节目分析、节目取证、

节目查处、效果验证、节目下载、查处简报、统计报表等工序。

4.2.3 节目检索

节目检索主要包括文本检索、音视频检索和节目内容摘要等模块。文本检索可对本地采集结果和搜索云数据资源进行即时查询，可基于标题或正文的关键词进行搜索，实现基于重点网站的深度采集和基于全网的广度采集。音视频检索将系统中的音视频与上传的样本进行比较，并将匹配到的节目加以统计。节目内容摘要主要是将系统中采集到的所有视频节目信息以时间降序排列展示，可以快速浏览采集到视频节目的标题、摘要、违规状态、审核状态、类型、网站、发布时间等基本信息，可以对节目进行违规 / 非违规审核、视频下载、贴文收藏、加入报表、正负面内容设置等操作。

4.2.4 专业管理

专业管理通过对系统网站采集的节目进行常规任务、专项任务、敏感节目、违规节目的关键字设置，可对匹配的疑似违规节目进行搜索、收藏、自定义简报导出等操作。

4.3 敏感舆情监测

敏感舆情监测部分主要是对舆情信息进行监测，展示出设定关注的敏感信息，对违规网站和违规节目列表详情集中进行监管。敏感舆情监测部分展现出采集于各主流搜索引擎的热点事件列表、热点舆情信息和报告文件。敏感舆情监测的内容，涵盖了关注的敏感信息、违规网站、有害节目、热点视频、热搜词、热点事件等方面。

4.4 综合业务管理

综合业务管理部分主要包括综合信息、统计分析、业务管理、系统管理和统计报表五部分。

4.4.1 综合信息

综合信息是对网站和节目信息进行综合管理、对黑名单网站和黑名单节目进行集中监管。系统对所有已添加关注的网站和本地爬虫采集到节目信息进行列表管理，对系统中列入黑名单的网站和节目以列表的形式进行展示，同时对系统中的节目下载情况进行管理。

图4　互联网视听节目监管系统页面

4.4.2 统计分析

统计分析是对系统网站、节目数据进行总体统计、日周月统计、工作报表以及对系统各种用户的登录次数和工作量的统计。

4.4.3 业务管理

业务管理主要包含管理系统的通知通告信息，流程上包括对系统所发布的所有通知通告信息的集中展示、上报 / 下发通告信息等操作。

4.4.4 系统管理

系统管理主要是对系统后台运行项目进行管理，主要执行对系统重要用户权限设置、单位角色管理、系统操作日志展示、数据库备份等操作。系统管理包括常用的按用户、角色级别设置权限功能，以及对所有不同权限的用户进行统一紧急通告的功能。系统提供日常监测数据和信息，可方便地进行数据查看和相应操作。监管人员也可在高级配置选项中对系统各项参数、本地热点、爬虫配置、同步关键字等进行配置。

4.4.5 统计报表

统计报表主要从网站数据、违规网站、查处网站和节目数据四个方面进行详细的数据统计，具备按需导出各类报表功能。

5 系统关键技术

系统关键技术来源于大数据的应用，体现在对大规模数据信息的采集和搜索上，包括对多源异构数据大规模采集的实现，以及海量数据搜索引擎的开发应用。

5.1 多源异构数据的大规模采集

系统的采集使用内嵌自动抽取算法和代理 IP 服务的海量数据采集及存储技术，根据不同数据源的数据特征，组合多种数据采集策略实现大规模爬取多数据源的信息。数据采集框架如图 5 所示。采集模块支持配置化写入多种存储介质，如 kafka、redis、mysql、hbase 等；支持自定义数据流动和分发规则；支持配置化抽取网页结构、关注的核心区域和字段；同时保留基于接口编程开发的能力，可处理复杂的页面。

5.2 海量数据搜索引擎

海量数据搜索引擎采用开源的分布式搜索引擎 ElasticSearch，并针对检索数据的需求对 ElasticSearch 的使用配置进行优化，包括对索引 schema 的设计、索引存储的策略的优化、重建索引的性能优化等，在保留搜索结果高亮功能的前提下，降低了索引大小，缩短了索引的重建时间，提高了索引线上查询的响应速度。

6 系统特色及技术创新

系统的技术特色及技术创新体现在以下两方面：其一来自

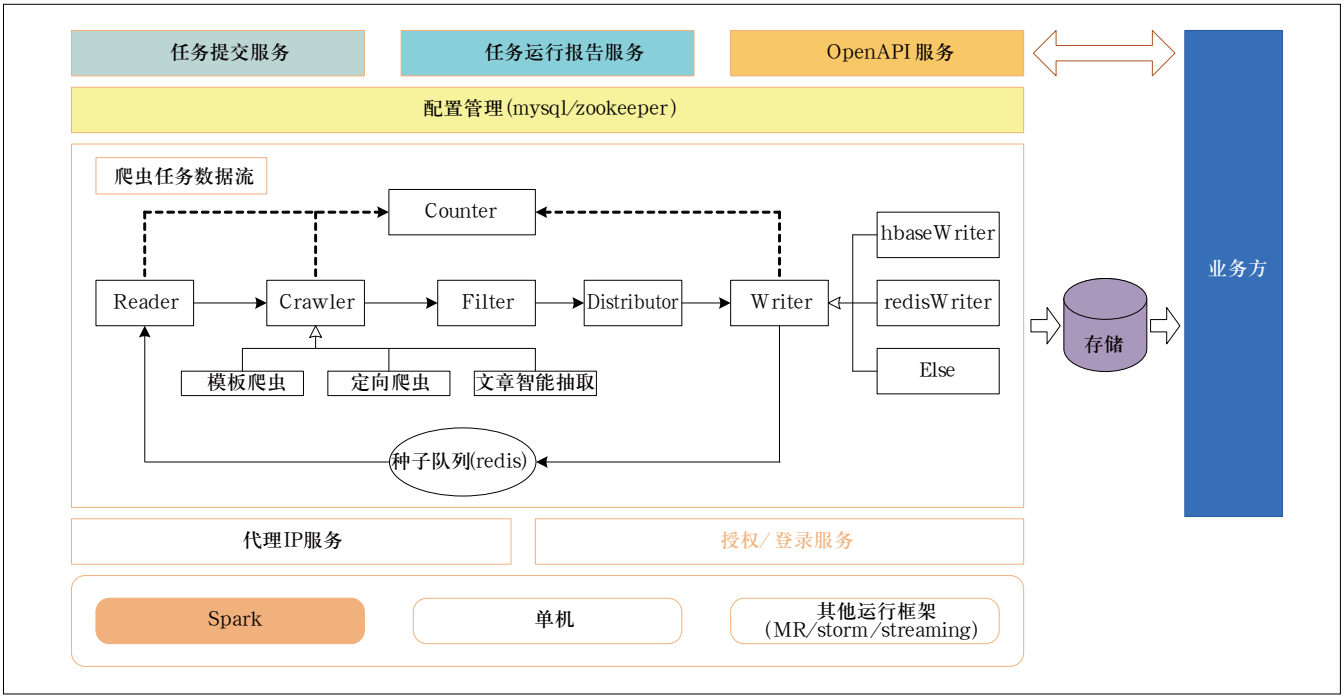

图5　互联网视听节目监管系统数据采集框架

系统前台，为基于深度学习的图像识别技术的运用；其二来自系统后台，为自有作业调度技术的使用。

6.1 采用基于深度学习的图像识别技术

由于传统的图像识别算法依赖于准确标注的训练集，因此仅能处理有限的图像类别。然而在互联网上，往往是提供了大量的标注和半标注图片。如何利用这些半标注信息通过增量方式自动学习新的类别已成为当前需要解决的一个难题。系统采用前沿的基于深度学习的图像识别技术，运用大规模自动基元学习方法，将每个图片看成由一系列可视单词组成，每个可视单词也被视作是图片的构成基元。由于本系统基于云计算平台，可以实现从超大规模的图像中自动学习、识别基元，完成在海量图像数据中准确、快速的信息识别工作。

6.2 采用自有作业调度技术

随着监管业务需求量的激增，原先的监管系统的多任务处理功能面对数据处理任务量的几何级增长已经显得力不从心。在现阶段，完整的海量数据处理任务需要多个具有严格的前后依赖关系的子任务共同协作完成，而多个子任务之间也存在着并行协作要求。针对当前需求的特点，系统采用自有作业调度技术。首先，自有作业调度实现了各类任务之间串行、并行以及串、并行模式的自由组合；在多种类型的任务同时调度时，实现针对每一种任务分别设置优先级，并以此为依据的统一调度。其次，在日趋复杂的业务硬性要求不断升级下，自有作业调度做到调度平台与业务代码完全分离，支持多种计算框架任务的运行，如 Spark 基于内存的分布式计算框架以及 Storm 实时流处理，满足现阶段海量数据处理业务的需求。第三，自有作业调度保证了任务调度的稳定运行。在执行大批量、高并发任务时，保证服务不跌宕更新；在添加新业务流程时，协调多设备间的负载，确保集群的负载均衡。

7 系统社会效益

我台互联网视听节目监管系统建成后，较之于原先的系统，无论在设计思想、大数据处理、文本语义分析、深度学习的图像分析处理、语音转文本等技术上，还是在对互联网网站视听节目的搜寻、检索、内容分析、预警、发布等效果上，都实现了非常大的跨越。新建的系统将多种技术及应用进行整合，并经过持续的优化改进，实现对于全省 26 家持证网站的遍历时间缩短到每天一轮，实现对于全省 40 万家左右活跃的有江苏 ICP 备案但无 AVSP 网站的遍历时间由原来的 10 天一轮缩短到 5 天一轮，从而大大增强了我省对于互联网视听节目的监测监管力度，提高了监管效率。系统建成投入使用至今，已封堵关停 100 多家无证网站，对互联网视听节目传播秩序的监管起到了明显的成效，净化了我省网络空间，为上网的用户提供了良好的视听环境，具有重大的社会效益。

8 结束语

习近平总书记指出，网络空间天朗气清、生态良好，符合人民利益。网络空间乌烟瘴气、生态恶化，不符合人民利益。江

【本文献信息】楼昶，李仲祥，高峰．广电省级 5G 平台建设发展的分析和思考 [J]. 广播与电视技术，2020,Vol.47(9).

广电省级 5G 平台建设发展的分析和思考

楼昶，李仲祥，高峰

（宁波华数广电网络有限公司，浙江 315000）

【摘 要】本文以 5G 高速发展为背景，从广电实际出发，通过分析广电发展 5G 的优劣势，阐述广电 5G 发展的目标和路径。同时，以广电省级 5G 平台建设为切入点，进行技术要点和成本收益分析。

【关键词】5G，B 端，网络切片，700M

【中图分类号】TN949.197 【文献标识码】A 【DOI 编码】10.16171/j.cnki.rtbe.20200009026

Analysis and Thinking on Construction and Development of Provincial-level 5G Platform of Radio and Television

Lou Chang, Li Zhong Xiang, Gao Feng

(NingBo WaSu Radio and Television Network Co.,Ltd. ZheJiang 315000, China)

Abstract Based on the background of 5G high-speed development, and starting from radio and television reality, this paper analyzes advantages and disadvantages of 5G development in radio and television, and expounds its goal and path At the same time, taking provincial-level 5G platform construction of radio and television as starting point, the paper analyzes technical points and cost-benefit.

Keywords 5G, B end, Network slicing, 700M

1 广电5G发展现状分析

2019 年，工信部正式为广电发放 5G 运营牌照，这就意味着，广电作为第四大移动通信运营商，正式进入 5G 移动通讯领域。但是通过广电自身分析，在 5G 发展的道路上，可谓任重而道远，优势和劣势并存。

广电 5G 发展优势主要包括 700M+4.9G 融合组网，覆盖半径更远；零起步建网，建网成本相对较低。广电 5G 发展劣势也很明显，包括无法实现 700M 的多天线输入输出（MIMO）；OTN 核心网互联互通能力不足；PTN/SPN 接入网基础薄弱；5G 建设和运营经验不足等。

2 广电5G发展目标及路径

2.1 5G 发展目标

广电没有 4G 移动通讯的基础，属于零起步建网，建议以

苏省广播电视监测台建设的互联网视听节目监测系统，基于大数据，采用先进的大数据分析、智能图像识别处理和任务调度技术，大大缩短了寻找省属地网站上暴恐、色情、反动等违规音视频的时间，降低了互联网违规视听信息对社会的负面影响，为营造我省清朗的网络空间，维护公众利益和国家利益，促进互联网视听节目产业健康有序的发展提供了有力保障。RTBE

参考文献：

[1] 谢伟．视听节目内容识别技术在三网融合中的应用 [J]. 广播与电视技术．2011，38，(2)：94—97.

[2] 李晓东，王仝杰．互联网视听节目监管的关键技术及系统简介 [J]. 广播与电视技术．2012，(5)：125—127.

[3] 谢燕燕．互联网视听节目监管系统关键技术研究及方案设计 [J]. 广播与电视技术．2015，42，(5)：120—124.

作者简介：

包力伟，男，1983 年生，硕士，工程师，江苏省广播电视监测台，主要从事各种项目建设、系统运行维护等技术工作。