Hadoop大数据处理用户行为记录的创新应用

杨秉学　张至柔　刘俊龙　吴　娟
（华北电力大学控制与计算机工程学院，北京　102206）

摘　要：随着计算机和网络应用的广泛深入，网络教学成为教育领域的重要组成部分。当前，网络教学视频存在教师与学生之间交互性、实时性、反馈客观性等方面的缺陷。基于此，本文利用Hadoop大数据实现对学生观看视频全过程的行为监控、记录和反馈，客观地分析教学视频的重难点、学生的掌握情况等信息，辅助教学双方。

关键词：大数据；网络视频教学；时间轴；用户行为监测与反馈
中图分类号：TP311.13　　**文献标识码**：A　　**文章编号**：1003-5168(2020)08-0037-03

Research on Innovative Application of Hadoop Big Data Processing User Behavior Record

YANG Bingxue　ZhANG Zhirou　LIU Junlong　WU Juan
(School of Control and Computer Engineering, North China Electric Power University, Beijing 102206)

Abstract: With the extensive application of computer and network, network teaching has become an important part of education. At present, online teaching videos have defects in interaction, real-time, and feedback objectivity between teachers and students.Based on this, this paper used Hadoop big data to implement behavior monitoring, recording and feedback on the entire process of watching videos by students, objectively analyzed the important and difficult points of the teaching video, the students´ mastery, and other information to assist both sides in teaching.

Keywords: big data; online video teaching; timeline; user behavior monitoring and feedback

随着信息技术的发展，互联网越来越贴近人们的生活，人们的衣食住行、娱乐与学习等方面都有互联网的身影[1]。依据国务院印发的《“十三五”国家信息化规划》，我国提出了大数据战略的重大决策，开启了信息化发展的新征程。教育部在2018年4月发布的《教育信息化行动计划2.0》也明确提出了促进信息技术在教育领域的广泛应用，推动教育的改革和发展，培养适应信息社会要求的创新人才以及促进教育现代化的目标。

1　研究意义

视频学习方式本身具有一定的局限性，传统网课需要全部播放完才能获得用户的反馈意见[2]，很多学生在通过视频自学的过程中并不了解课程的难点和重点，导致虽然看视频学了很长时间，效果却并不理想。同时，进行视频教学的教师也无法了解学生掌握的情况。因此，有必要处理用户在视频学习过程中产生的大量行为数据，从中获取用户观看的教学视频的难点、重点，反馈给视频提供方和教师，促进他们调整教学内容和方式，从而提高网络教学质量。

每个用户在观看教学视频时都可能根据自己的需要和已掌握的相关知识重点看自己需要的部分，跳过不需要的部分，即对视频进度条进行向前、向后拖动或倍速播放，这就形成了观看视频时的用户行为，产生大数据分析的“滤镜效应”，即定位观众的热点，通过后台的服务器自主进行计算，获得用户观看视频的行为数据[3]。由于观看视频的用户数量巨大，这种行为数据的量也极大，因此数据处理对计算、存储的要求很高。Hadoop是对大量数据进行分布式处理的软件架构，包含了当前主流的大数据处理技术，适合作为对用户行为数据进行计算、存储、管理的平台，因此笔者在Hadoop平台上研发了教学视频

收稿日期：2020-02-03
作者简介：杨秉学(1999—)，男，本科在读，研究方向：计算机科学与技术。

的用户行为处理系统。该系统可将前台(视频播放器)提交的用户观看视频的行为数据通过计算转换为每个视频中每秒视频片段的播放次数统计,并存储于Hadoop文件系统HDFS中,作为该视频的播放情况记录。在前台需要时,这些数据将以曲线形式展示到播放界面上,作为新用户或视频提供方的参考,客观展示教学视频中的重点和难点。

2 系统架构和算法设计

2.1 系统介绍

系统的Hadoop节点部署为1个master节点、6个slave节点,其中slave2、slave3、slave5、slave6均为DataNode节点,形成分布式存储数据的HDFS文件系统。前端播放器提交的JSON格式的用户行为数据由系统发送至HDFS中,之后运用Java语言编写Job函数与MapReduce函数,分布式处理用户的倍速播放的片段起始、结束位置和播放速度,前进、回退播放的片段起始、结束位置行为痕迹数据,获得每个视频以秒为单位的播放次数结果,将计算结果形成文件并进行压缩,存储到HDFS中,成为某视频播放情况记录文件,并在前端提出请求时反馈给前端。

2.2 算法设计

程序总体流程设计如图1所示。

2.2.1 MapReduce数据处理算法。在Hadoop平台对大量用户行为数据进行分析与处理,设计MapReduce算法将不同用户观看不同视频的用户行为数据装入多个Mapper里,由Mapper将每一条用户行为数据转化为该用户观

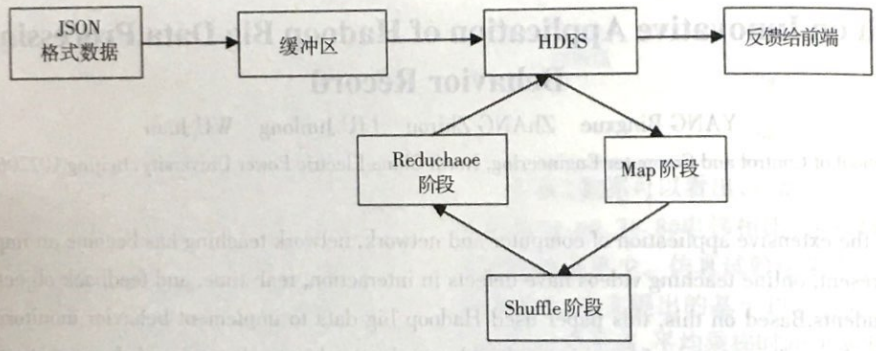

图1 程序总体流程设计

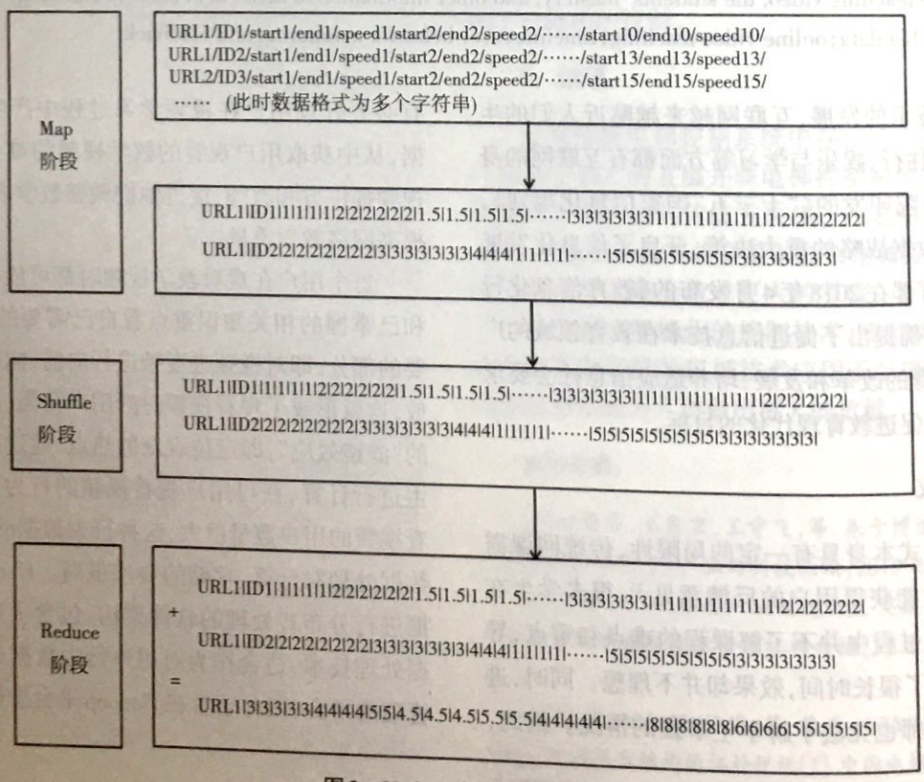

图2 Hadoop程序算法设计

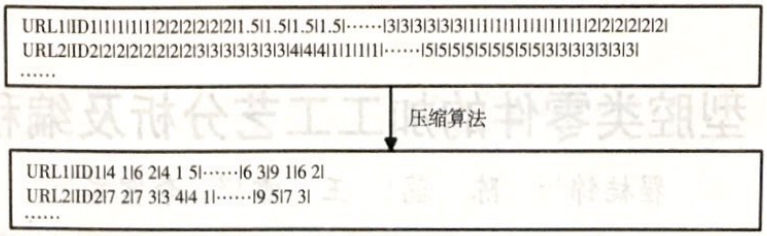

图3 压缩算法设计

看某个视频中以秒为单位的视频片段次数，形成一条记录，这些记录再根据视频的URL分配给多个Reducer，由这些Reducer将这些记录合并计算，得到各视频以秒为单位的总体播放情况，并用这些数据与HDFS中存储的该视频原总体播放情况数据累加，更新该视频总体播放情况数据。具体算法设计如下。

2.2.1.1 Map阶段。各个Mapper将前端提交的每一条用户行为数据，按照每个播放片段的起始时间、结束时间及播放倍速，对整个视频以秒为单位形成的数组进行加权计算。例如，一条用户行为数据为该用户从视频的第1秒到第30秒以2倍速进行播放，则该视频数组第1秒，第2秒，直至第30秒的数据都加0.5。最终得到多个观看次数数组，数组的key值为视频URL，value值为某个用户在观看该视频的过程中根据观看倍速对每秒视频进行加权的数值。

2.2.1.2 Shuffle阶段。它是MapReduce算法的关键环节，Mapper的计算结果进行“洗牌”，将key值相同的数据分到一类，并交给同一个Reducer处理。

2.2.1.3 Reduce阶段。将key值相同的信息中的value值累加在一起，更新每个视频所有用户观看行为的总记录。系统的程序算法设计如图2所示。

2.2.2 处理结果压缩算法。由于前述MapReduce算法计算出的视频总体播放行为数据量比较大，在输出到HDFS文件系统存储时，I/O开销比较大，各主机节点之间交互频繁。为了提高网络利用率和处理速度，人们可以将计算结果进行压缩后再输出到HDFS中，这样既节约了存储空间，又节约了网络带宽。

为了保证传输过程是无损传输，后续数据处理是正确的，发送方压缩后得出校验码，接收方接收数据后进行校验，使得数据传输准确无误。

本文采用文件字符里面的重复字，用“数字（即重复次数）+字符”代替原来重复字符的方式进行压缩。压缩前后数据结构如图3所示。由于视频文件每秒片段被播放的次数相同的概率很大，这样压缩出来的文件很小，压缩率很高，可以大大减少读写HDFS的开销。

校验采用java.util.zip.CheckedInputStream里面的getChecksum()方法进行校验。

3 结论

用Hadoop大数据将单个用户的视频观看行为痕迹进行计算、分析、存储，转化为各教学视频的总体播放情况数据，这种方法以数字化的手段直观地记录用户的视频观看信息，从大数据的角度监测、存储和分析用户观看行为数据，使得教学信息的反馈方式更加实时、客观、可靠，对提高网络教学效果具有重要意义，会成为促进网络教学水平提高的有力工具。未来，人们会进一步丰富获取用户行为数据的内容和方式，采用更丰富、灵活的方式反馈教学情况，帮助师生在网络上高效获取知识。大数据分析技术为快餐式观看视频提供技术支持，虽然大数据的作用很大，但是它仅仅是一种手段，不能完全替代认真观看的地位[4]。

参考文献：

[1]詹昕蕊，张至柔，胡柳静，等.基于时间轴的用户播放行为监测播放器研究[J].科学与信息化，2019(19)：123-124.

[2]张蓝姗.网络视频观看模式的创新与影响：以“绿镜”智能观看模式为例[J].当代传播，2017(4)：105-106.

[3]徐方.大数据时代下的影视业革新[J].西部广播电视，2014(9)：8.

[4]刘融.基于大数据的影视剧创新[J].中国新通讯，2015(1)：32-33.

科学开拓视野 技术引领未来

主管：河南省科学技术厅
主办：河南省科学技术信息研究院
出版：河南《创新科技》杂志社

河南科技

Henan Science and Technology

创新驱动

总第 706 期

08

03月中

2020

基于Winkler基体的滚柱丝杠降维接触模型研究

◎机械加工技术在汽车发动机曲轴制造中的应用

◎深埋大直径污水管带水迁改施工技术

◎基于交通安全的市政道路绿化设计

◎新一代天气雷达电磁波辐射测试研究

ISSN 1003-5168

9 771003 516201 08

中国核心期刊（遴选）数据库源期刊　中国期刊全文数据库源期刊

万方数据知识服务平台全文收录　中文科技期刊数据库源期刊

2020年第08期
总第706期

主管单位
河南省科学技术厅
主办单位
河南省科学技术信息研究院
出版单位
河南《创新科技》杂志社

国际标准刊号　ISSN 1003-5168
国内统一刊号　CN 41-1081/T
河南省自然科学一级期刊

CONTENTS 目　次

资　讯 CONSULTATIONS >>>

技　术 TECHNOLOGY >>>