ISSN 1000-1239
CODEN JYYFEY

# Journal of Computer Research and Development

# 计算机研究与发展

第53卷 第4期 2016年4月
Vol.53 No.4 Apr.2016

主办 中国科学院计算技术研究所 中国计算机学会 出版 科学出版社

中国计算机学会会刊

# 计 算 机 研 究 与 发 展

(Jisuanji Yanjiu Yu Fazhan)

第 53 卷 第 4 期 2016 年 4 月


## 目 次




责任编辑:沈宝丽

JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT
Vol. 53 No. 4 April 2016

## CONTENTS

# 《计算机研究与发展》编委会




计算机研究与发展
Jisuanji Yanjiu yu Fazhan
(月刊，1958年创刊)
第53卷　第4期　2016年4月

主　　管　中国科学院
主　　办　中国科学院计算技术研究所
　　中国计算机学会
编　　辑　《计算机研究与发展》编辑部
　　中国科学院计算技术研究所
　　地址：北京中关村科学院南路6号
　　邮政编码：100190
　　电话：+86 (10) 62620696(兼传真)
　　+86 (10) 62600350
　　E-mail:crad@ict.ac.cn
　　http://crad.ict.ac.cn
主　　编　徐志伟
出　　版　科学出版社
　　地址：北京东黄城根北街16号
　　邮政编码：100717
印刷装订　北京佳艺恒彩印刷有限公司
国内总发行　北京报刊发行局
订 购 处　全国各邮电局
国外总发行　中国国际图书贸易总公司
　　北京399信箱
　　邮政编码：100044

**Journal of Computer Research and Development**
(Monthly, Started in 1958)
Vol.53 No.4 Apr.2016

**Supervised by**　Chinese Academy of Sciences
**Sponsored by**　Institute of Computing Technology, Chinese Academy of Sciences
　　China Computer Federation
**Edited by**　Editorial Office of *Journal of Computer Research and Development*
　　Institute of Computing Technology, Chinese Academy of Sciences
　　Add: 6 Kexueyuan South Road, Zhongguancun, Beijing 100190, China
　　Tel: +86(10) 62620696 (also Fax)
　　+86(10) 62600350
　　E-mail: crad@ict.ac.cn
　　http://crad.ict.ac.cn
**Editor-in-Chief**　Xu Zhiwei
**Published by**　Science Press
　　Add: 16 Donghuangchenggen North Street, Beijing 100717, China
**Printed by**　Beijing Jiayi Hengcai Printing Co.Ltd
**Distributed by**　Beijing Bureau for Distribution of Newspapers and Journals
**Domestic**　All Local Post Offices in China
**Foreign**　China International Book Trading Corporation
　　P.O.Box 399, Beijing 100044, China

ISSN 1000-1239

CN 11-1777 / TP　国内邮发代号：2-654　国外发行代号：M603　定价：68.00元

计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20151143
53(4):742-751,2016

# 基于流行度预测的互联网+电视节目缓存调度算法

朱琛刚　程　光　胡一非　王玉祥
(东南大学计算机科学与工程学院　南京　211189)
(教育部计算机网络和信息集成重点实验室(东南大学)　南京　211189)
(gcheng@njnet.edu.cn)

# A Caching Strategy for Internet Plus TV Based on Popularity Prediction

Zhu Chengang, Cheng Guang, Hu Yifei, and Wang Yuxiang
(*School of Computer Science and Engineering, Southeast University, Nanjing* 211189)
(*Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing* 211189)

**Abstract**　Internet plus TV tends to excessively consume storage space to achieve higher cache hit ratio. A novel cache schedule algorithm called PPRA(popularity prediction replication algorithm) is proposed in this paper based on programs popularity forecast. Firstly, according to statistical analysis from actual measurement, we apply random forests (RF) algorithm to construct a forecasting model of programs popularity. Subsequently, we use the principal component analysis (PCA) to overcome dimensionality curse and accelerate the forecasting process. Finally, we validate PPRA with authentic behavior data of a certain cable operator's 1.3 million users in a period of 120 days. Our experimental results show that PPRA only consumes 30% storage space to achieve a fixed cache hit ratio compared with LRU and LFU algorithms, therefore the cost of Internet plus TV platform is saved.

**Key words**　Internet plus TV; popularity prediction; random forests (RF); caching strategy; dimensionality curse

**摘　要**　针对互联网+电视平台为提高热点节目命中率而过渡消耗存储空间的问题,提出一种基于流行度预测的节目缓存调度算法PPRA(popularity prediction replication algorithm).首先,在对实际测量数据进行统计与分析的基础上,使用随机森林(random forests, RF)算法构建节目流行度预测模型.同时,针对所选特征存在的"维数灾难"问题,利用主成分分析法(principal component analysis, PCA)实施特征降维处理,以实现视频流行度预测值的快速计算.然后基于节目流行度预测数据调度缓存中的节目.最后以某广电运营商130万用户120 d的收视数据为例,对PPRA算法进行实验.实验结果表明,在保证一定缓存命中率前提下,与LRU,LFU算法相比,PPRA算法仅需30%的存储空间,可有效降低互联网+电视平台的建设成本.

**关键词**　互联网+电视;流行度预测;随机森林;缓存策略;维数灾难

**中图法分类号**　TP393

**收稿日期**:2015-12-21;**修回日期**:2016-02-16

**基金项目**:国家"八六三"高技术研究发展计划基金项目(2015AA015603);江苏省未来网络创新研究院未来网络前瞻性研究项目(BY2013095-5-03);江苏省"六大人才高峰"高层次人才项目(2011-DZ024)
This work was supported by the National High Technology Research and Development Program of China (863 Program) (2015AA015603), the Prospective Research Program on Future Networks of Jiangsu (BY2013095-5-03), and the Six Industries Talent Peaks Plan of Jiangsu (2011-DZ024).

**通信作者**:程光(gcheng@njnet.edu.cn)

为克服传统广播电视要求用户在固定时间及地点收看特定电视台播出的特定节目,无法满足当今用户空间移动化、时间碎片化和内容个性化需求的缺点,互联网+电视网络通过大数据分析技术,从海量用户行为数据中分析当前用户的收视兴趣,有的放矢地向用户提供符合用户兴趣的节目内容,为用户提供个性化的电视节目,给用户带来更加人性化的体验.为此,互联网+电视网络系统通过将热点节目缓存在边缘服务器上,有效降低网络访问流量和播放时延.但为保证一定的节目缓存命中率,现有节目缓存调度算法需要消耗大量存储空间,增加了缓存建设成本,阻碍了互联网+电视的进一步推广.

针对此问题,本文通过对实际互联网+电视平台数据分析,发现节目流行度与用户点播量、播出时间、节目内容、制作方等因素密切相关,在定性分析这些关系的基础上,提出一种基于流行度预测的缓存调度算法 PPRA(popularity prediction replication algorithm).该算法首先使用随机森林(random forests, RF)算法建立了节目流行度的预测模型,采用主成分分析法(principal component analysis, PCA)对模型输入实施特征降维处理,实现流行度预测值的快速计算;然后通过比较各节目的预测流行度决定调入缓存的节目内容;最后,以某广电运营商的 130 万云媒体电视用户行为数据为例,对所提出的算法进行测试.实验结果表明,在保证一定缓存命中率的前提下,与近期最少使用(least recently used, LRU)[1]算法、最近最不常用置换算法(least frequently used, LFU)[2]算法相比,PPRA 算法仅需 30%的存储空间,可有效降低互联网+电视平台建设成本.

## 1 相关工作

### 1.1 流行度预测

节目流行度是节目收视量的表现,收视量越大流行度越高.节目流行度作为反映节目热度的重要指标,不仅能够帮助运营商在购买节目版权时做出决策,指导广告商合理分配广告投入,而且能够作为边缘内容分发网络(CDN)缓存调度的重要依据.目前针对在线视频、图像、音乐、微博、话题的流行度分析和预测是当前研究的热点和难点.近期的研究工作大多通过在具体情境下引入更多的分析特征,以提高预测的准确率和覆盖率.

Wang 等人[3]通过分析用户在腾讯微博中转发优酷视频链接的行为,采用神经网络算法构建视频流行度模型,在不依赖历史收视数据的情况下取得了良好的预测精度.Vallet 等人[4]综合社交网络和视频发布平台的数据,提出基于传染病传播的流行度预测模型,实现了对 YouTube 上点播量瞬间爆发视频的预测.孔庆超等人[5]研究了影响网络讨论帖流行度的动态因素,并提出一种基于动态演化的讨论帖流行度预测模型.Ding 等人[6]通过分析社交网络中流行图片和非流行图片的特征,提出一种用于预测图片流行度的模型.Figueiredo[7]研究了 YouTube 视频流行趋势,给出了 UGC 视频特征和节目流行度的关系.Pinto 等人[8]利用节目上线初期的点播数据,提出了一种适用于 YouTube 平台的内容流行度预测算法,有效地降低了预测偏差.Ahmed 等人[9]使用内容自身流行度的比例变化与用户关注其他内容流行度的比例变化之间的相似程度,得出内容在不同时间段聚类组的转换关系.Sanner 等人[10-11]利用 Twitter 数据对 YouTube 平台内容流行度进行分析,采用转移学习算法有效预测出视频流行度发生突变的情况.Chang 等人[12]采用自动回归模型捕获用户行为变化的特征,实现了对网络连续剧内容的质量分析和流行度预测.McParlane 等人[13]为解决预测内容缺乏标签、评论等交互数据导致的冷启动问题,使用图片、用户上下文信息以及视觉感受预测 Flickr 图片的流行度,取得了良好的效果.

### 1.2 缓存调度

CDN 网络中如何高效地管理边缘服务器的缓存空间是当前国内外研究的热点问题.缓存策略主要由缓存替换算法和缓存决策算法 2 部分构成.缓存替换算法负责在有限的缓存空间下寻找出最合适的内容数据进行缓存;缓存决策算法为需要缓存的数据寻找合适的存储节点.因此,如何在有限的缓存空间下设计高效的缓存策略,从而提高缓存命中率、降低网络带宽消耗,是缓存管理的关键.

Abrahamsson 等人[14]通过分析一个在线时移系统的用户点播记录,发现 20%的节目产生了 80%的点播请求.如果能够合理地缓存热点节目,可以很好地降低系统的带宽消耗.Balachandran 等人[15]分析了 3000 万的视频会话数据,发现用户在使用 VOD 服务时存在同步收视的特征,采用混合 P2P-CDN 的方式能有效降低 CDN 的建设成本.Zhou 等人[16]

研究了 VOD 系统中不同类型节目流行度的变化规律,并提出一种 FIFO 和 LFU 混合缓存策略,取得了较好的效果. Gouta 等人[17]提出了一种 ICN 网络中动态自适应流行度变化的流缓存策略. Abrahamsson 等人[18]研究了影响缓存效果的诸多因素,包括节目流行度、节目大小、缓存策略等. Nencioni 等人[19]通过在机顶盒硬盘上提前录制用户可能收视节目,有效降低高峰时段的网络流量. Agrawal 等人[20]通过缓存不定长的影片片段取得了相比 LRU 算法更好的命中率. Akhtar 等人[21]提出了一种用于缓存在线影片的分层过滤算法,该算法复杂性类似于 LRU,但在命中率、替换率和缓存吞吐量上均有明显提升. Park 等人[22]提出了一种自适应流行度分布的二元缓存算法,在缓存较小时取得了良好的效果. 王永功等人[23]提出了一种基于预过滤的改进算法,通过对原始内容的整形使得内容更容易被后续缓存命中.

以上有关缓存策略的研究中,各边缘节点在进行决策时均未综合考虑内容的流行度和缓存容量等因素. 一些算法片面追求缓存命中率,没有考虑缓存空间和内容总量的关系,造成整个系统建设与运营成本高昂,性价比较低.

## 2 互联网+电视节目的流行度特征

目前流行度的研究对象主要集中于视频、图像、微博、分享链接等. 流行度的定义与具体的应用相关. 本文研究对象是互联网+电视平台中的节目,因此将节目在某段时间窗口内的点播量以及节目点播量在总片库中的排名数据作为节目流行度的度量.

### 2.1 数据描述

本文数据采集是某广电运营商互联网+电视平台 2015-01-01—2015-04-30 共计 120 d 的用户时移回看记录和电子节目指南数据. 本文主要关注基于 IP 技术的互联网+电视,采集数据不包括传统直播业务的收视记录. 通过对服务器 RTSP 日志的清洗,用户回看记录、电子节目指南都作为一条记录存储在 ORACLE 数据库中. 表 1 总结了数据集的具体情况. 本文共收集了 130.9 万个用户 2.01 亿次回看记录,节目数量总计 42.3 万个. 每条回看记录包含机顶盒序列号、回看时间、节目名称、频道名称;每条电子节目指南包括频道名称、节目名称、节目类型、节目时长、播出日期、拍摄日期、导演、演员信息.

**Table 1 Internet Plus TV Data Set in Figures**

**表 1 互联网+电视收视数据总体情况**

| Item | Requests | Programs | Clients |
|---|---|---|---|
| Daily Max | 2 300 747 | 20 447 | 407 133 |
| Daily Min | 1 066 706 | 17 552 | 225 290 |
| Daily Median | 1 678 506 | 19 098 | 318 052 |
| Total | 201 420 717 | 423 254 | 1 309 381 |

通过对 120 d 的点播行为数据(如图 1 所示)进行分析,发现在线用户数(clients)和节目数(programs)基本稳定在相对固定的范围,而收视次数(requests)随时间呈现周期性震荡. 点播次数在每周六通常会出现一个高峰. 数据采集周期内,平均每天点播次数达 167.5 万次.

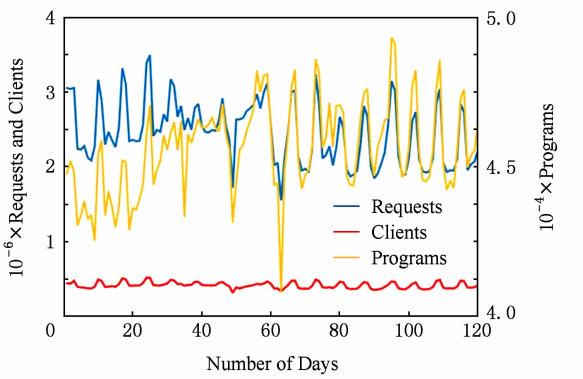



Fig. 1 Number of requests, active clients, and distinct programs requested over 120 days.

图 1 点播量/用户/节目在采集周期内的总体情况

### 2.2 流行度与时间关系

在观测周期内,点播量以周为单位存在明显的周期变化. 周末的点播量明显高于工作日的点播量. 图 2 展示了一周之内每小时的点播量数据. 通过图 2

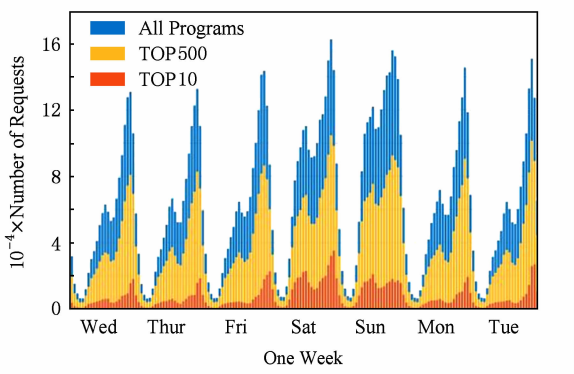



Fig. 2 Number of requests of top10, top500 and all programs for an hour in a week.

图 2 一周内每小时 top10,top50 和所有节目的点播量

可以发现,一周内的所有节目点播次数呈现出明显的周期性,其中周六和周日在一周中的点播量最大,而在每天晚间 20:00—21:00 是收视高峰期;此外在每天中午的 12:00—13:00 会出现一个小高峰,而在下午会有所回落,直到傍晚 16:00 之后,点播量开始逐渐上升.总体看来,Top10 和 Top500 的节目也呈现出同样的特征.

### 2.3 流行度与内容关系

观测周期内,平均每天有 4.5 万个节目被点播.排名最高的节目累计点播量接近 20 万次,大部分节目的累计点播量不足 1000 次.图 3 是收视次数与节目排名的累积分布,可以看出排名前 10%的节目占据了 75%的点播量,排名前 20%的节目占据了 90%的点播量,呈现出明显的长尾效应.

本文将节目分为电影、电视剧、新闻、少儿和综艺.由于收视人群的不同以及各期节目之间的关联性,造成不同类型节目的流行度呈现不同的变化趋势.

图 4(a)为一部热播电影上线以来的点播量以及每天的点播排名变化情况.可以发现,电影节目在上线初期点播量很大且排名很高,随上线时间变长,其点播量和排名均会同步下降;在周末,点播量会出现一个明显的回升.

图 4(b)为综艺节目点播量以及点播排名变化情况.可以看出节目在上线初期的点播量较高;而后出现下降的情况;与其他类型节目不同,真人秀节目会周期性出现一些点播的高点,而这一周期正好与真人秀节目的播放周期一致.

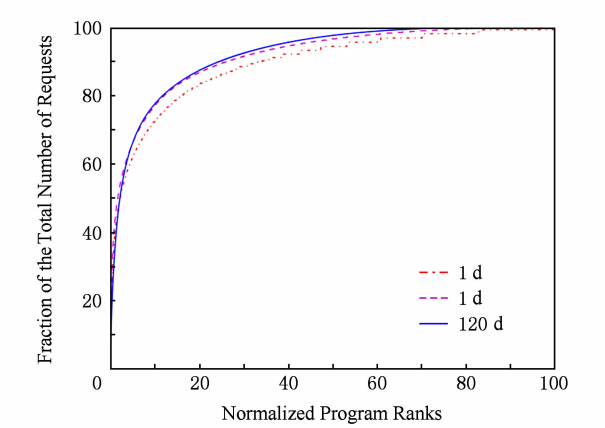



Fig. 3 Cumulative distribution of requests to programs.

图 3 收视次数随节目排名的累积分布

图 4(c)为单条新闻发布 48 h 内的点播量以及点播排名变化情况.可以发现,新闻刚发布,即带来较高的点播量,其排名也较高,但点播量和排名在 48 h 内都会快速下降.

图 4(d)为儿童动漫节目点播量以及点播排名变化情况.可以发现,动漫节目在上线初期的点播量特别大,然后会下降;与其他类型节目不同的是,动漫节目在上线一段时间后仍会保持比较稳定的点播

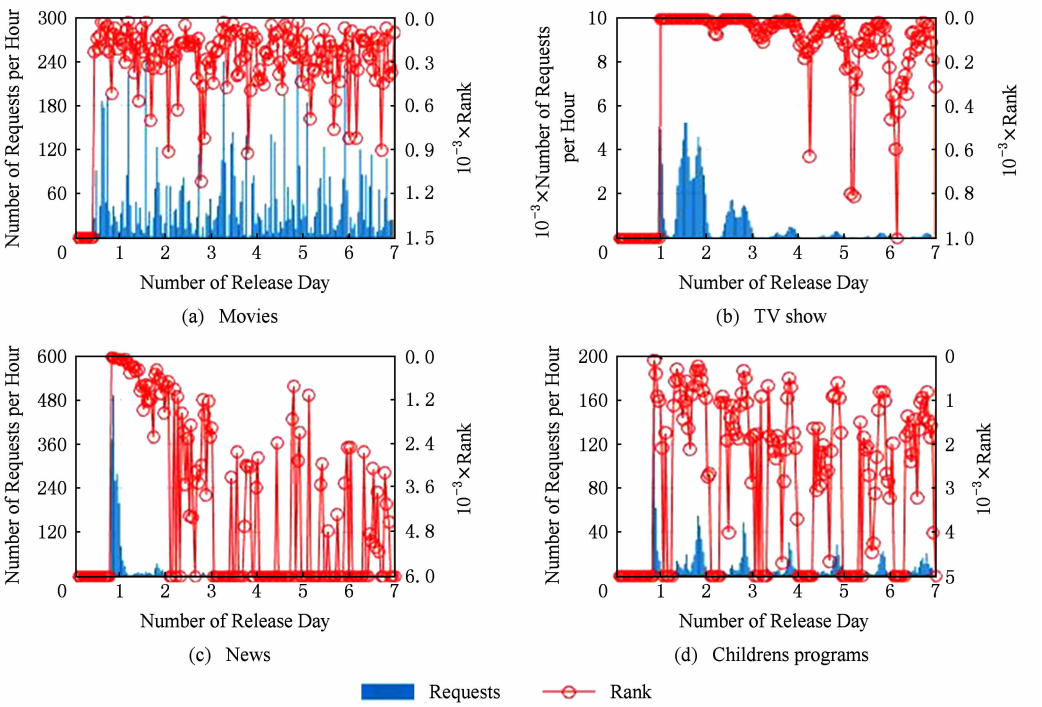



Fig. 4 Requests and rank for movies, TV show, news and children's programs for seven days.

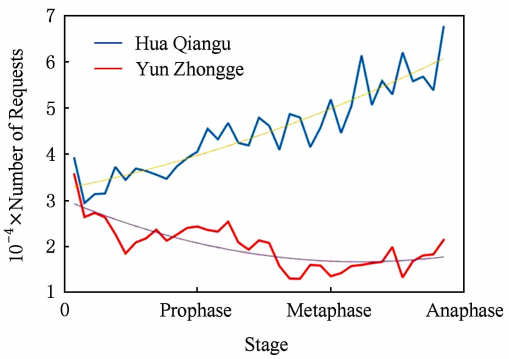
图 4 电影、综艺、新闻和卡通 7 日点播量和收视排名变化

量和点播排名;此外,动漫节目在周六和周日的点播量会明显高于一般工作日.

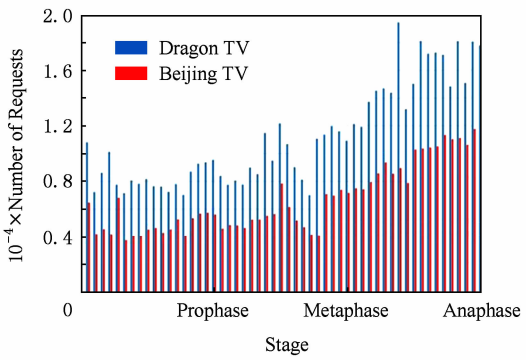
### 2.4 流行度与节目制作方关系

同样内容和类型的节目,由于导演、演员、制片公司、宣发频道和首播时间的不同,点播量和排名变化也呈现出不同的趋势.

节目是否符合当前观众的兴趣口味,节目本身的制作水平是关键的影响因素,与导演风格、演员的当红程度、制作公司能力息息相关.在观测的 120 d 内,本文整理了 445 位导演和 2008 位演员的节目数据,发现 16%的导演制作的节目产生了 80%的点播量,15%的演员参演的节目产生了 80%的点播量.图 5(a)为在相同宣发频道、相同内容主题的前提下,不同制作方拍摄节目的流行度随时间的变化.可以看出,不同演员和导演制作的节目会产生截然不同的流行度.




(a) Influnce of different films productions




(b) Influnce of different distribution channels

Fig. 5 Influence of film production and distribution corporation on programs popularity.

图 5 不同制作方和宣发频道对节目流行度的影响

宣发频道的实力和定位也影响着节目的点播量和排名变化.宣发频道是否有足够实力在节目播出前在观众群中为节目预热、宣发频道本身带来的收视惯性,以及该宣发频道的定位是否与节目风格符合都直接影响节目上线后的流行度.图 5(b)为相同节目在不同宣发频道上的流行度变化,可以看出,宣发频道对节目流行度有直接的影响.

## 3 PPRA 算法

从第 2 节分析可看出,可通过选取多个不同特征来描述一个节目,如上线日期、播出时间、节目类型等,而节目流行度与所选取的节目特征之间有很强的关联性,因此本文选用一系列有监督学习的方法来对节目的历史数据进行分析和学习,从而提取影响节目流行度的关键特征,并构建节目流行度随时间变化的模型,根据某个节目的特征和该节目上线以来的历史数据预测该节目未来 7 d 的流行度,为节目的缓存调度提供合理的理论依据.

### 3.1 数据预处理及输入特征选取

在本文第 2 节中,已经描述了若干对节目流行度有显著影响的特征,但是由于这些特征并不能够直接作为机器学习算法模型的输入,所以需要对其进行预处理.

1) 上线时间.图 4 为电影、综艺、新闻和卡通 4 类电视节目 7 d 的点播量和收视变化,可以看出,随着节目上线时间的增长,节目的点播量和收视排名均会出现下降.因此,新上线节目往往会比老节目更加受欢迎,本文用“距离首次上线的天数”这一指标衡量上线时间带来的影响,如式(1)所示:

$$\Delta T = T - T_0, \tag{1}$$

其中,$T$ 为观测日期,$T_0$ 为节目首次上线时间.

2) 频道、节目类型.图 5 为不同制作方和宣发频道对节目流行度的影响,可以看出宣发频道和节目类型对节目的流行度有直接的影响.同时,此类特征属于类型特征,每一个特征都有 $m$ 个不同的取值,但同时只有一个是有效的.本文将其转换成采用 One-Of-$K$ 编码模式的特征,即如果一个特征有 $m$ 个不同的取值,则用一组长度为 $m$ 且只有一位为 1 的二进制数来表示.

3) 节目标签、导演、演员.本文的观测数据集中,包含 445 位导演和 2008 位演员,通过分析发现 80%的点播量由 16%的导演制作的节目产生,15%的演员参演的节目产生了 80%的点播量.此类特征和节目类型有相似之处,都会有 $m$ 个不同的取值,但不同的是有效取值可以同时存在多个.对其采用类似的处理方法,转换为长度为 $m$ 的二进制位后,将每一个有效的位用 1 来表示,其余则为 0.需要指出的是,导演和演员的数量众多,如果不加处理直接

使用会导致样本属性维度过大、训练结果过拟合.因此在训练前,筛选贡献80%流行度的导演和演员,将导演和演员的数量分别从445位和2 008位减少到223位和1 202位.

4) 播放时间.从图2可以看出,不同时间段的点播量存在较大差异.本文将一天按24 h划分为24个时间段,根据节目播放的时间,为其标记1—24当中对应的值,然后再根据类型特征的预处理方法进行处理.

此外,节目时长以及是否为周末/节假日对影片的点播率也有很大的影响,将这些因素全部作为特征加入到机器学习的模型中,将有效提升模型预测结果的准确性.

对本文设计的模型而言,一条样本$\{I_v^T, O_v^T\}$由输入特征向量($I_v^T$)和未来7 d该节目的流行度($O_v^T$)组成.$\boldsymbol{O}_v^T$是一个7维向量,每一维代表该节目在接下来某一天的流行度.

### 3.2 流行度预测模型的训练

Biau[24]对RF算法的适用场景和计算模型进行了分析和改进,改进后的RF可以用来做分类、聚类、回归和生存分析.本文选用RF模型的原因有3点:

1) 经过预处理后的特征维度会达到2 000的数量级,即使经过一系列的降维处理,仍会有100.而RF算法得益于算法中双重抽样的环节,在面对高维特征时仍有较好的性能和较高的准确性.

2) RF算法在学习完成后可以给出特征的重要性,这对于分析特征对点播率的影响有很大的作用.

3) 本文的目标是根据节目的特征预测出未来7 d的流行度,属于回归问题,RF算法相比于传统线性回归算法更为简洁高效.

使用随机森林算法进行流行度预测的过程如下:

设$\boldsymbol{X}$为$R\times M$的矩阵,$X_{ij}$表示第$i$个样本的第$j$个特征;$\boldsymbol{Y}$为$R\times 1$的向量,$Y_i$表示第$i$个样本的输出值.

Step1. 对数据集进行行、列双重采样,并对采样后的数据建立决策树.行采样采用有放回的采样方式,也就是在采样的集合中有可能有重复的样本.设输入样本有$N$个,每次采样的个数为$n(n<N)$个.这样使得每一棵树的输入样本都不是全部样本,在训练过程中相对不容易出现过拟合的情况.对于列采样,从$M$个特征中随机选取$m(m=\sqrt{M})$个,由于每棵树的特征并不是全部特征,所以使算法对数据缺失有很好的适应性.采样的次数即随机森林中决策树的棵数$T$视具体情况而定.

Step2. 采样完成后对数据使用完全分裂的方式建立决策树.决策树的训练$DecisionTreeTrain(\boldsymbol{X},\boldsymbol{Y})$过程如下:

1) 如果$\boldsymbol{X}$中的所有样本值都相同或$R<2$,则跳转Step2,否则跳转Step3.

2) 产生一个叶节点,该节点的值为$\boldsymbol{X}$中任意样本的值.

3) 搜索分裂变量和分裂点.将空间划分为2个节点$R_1, R_2$.假设分裂变量和分裂点分别为$j$和$s$,定义一对半平面如式(2)所示.

$$R_1(j,s)=\{X_{ij}\mid X_{ij}\leqslant s\}\ \text{且}\ R_2(j,s)=\{X_{ij}\mid X_{ij}\geqslant s\}. \tag{2}$$

4) 搜索分裂变量$j$和分裂点$s$的目标函数为

$$\min_{j,s}\left[\min_{c_1}\sum_{X_{ij}\in R_1(j,s)}(Y_i-c_1)^2+\min_{c_2}\sum_{X_{ij}\in R_2(j,s)}(Y_i-c_2)^2\right], \tag{3}$$

其中:

$$c_1=ave(Y_i\mid X_{ij}\in R_1(j,s))\ \text{且}\ c_2=ave(Y_i\mid X_{ij}\in R_2(j,s)). \tag{4}$$

5) 对分裂出的2个节点$R_1$和$R_2$再次调用算法

$$DecisionTreeTrain(R_1,Y_1),\quad DecisionTreeTrain(R_2,Y_2).$$

Step3. 决策树训练完成后,输入需要预测的数据,每一棵树会给出一个预测值$p_i$,计算最终结果

$$p=\frac{p_1+p_2+\cdots+p_T}{T}. \tag{5}$$

### 3.3 模型准确性评估

用来训练的数据集是2015-03-04—2015-04-15之间某广电运营商互联网+电视平台共103 721条各类节目的节目特征,130万用户在节目上线7 d的点播量及节目的点播排名.

输入特征向量$\boldsymbol{I}_v^T$由包括节目时长、播放时间、演职员信息在内的1 872个特征组成.在这些特征当中,虽然导演、演员特征都已经经过优化,规模得到了缩减,但特征维度仍然过大.为提升算法运行效率,本文借用Candès等人[25]提出的PCA算法在最大保留特征规律的前提下对输入特征向量进行降维优化.

在模型训练的过程中,原有的103 721条样本经过清洗,去掉异常值和空值后剩余91 636条,从中随机打乱抽取15%的样本作为模型训练的交叉验证集,防止在样本空间中出现过拟合的现象,影响预测的准确率.在多次尝试和调整之后,发现将RF算法中每次抽样的树个数设定为100棵,通过PCA将特征降维至50维时,模型可以取得最佳的输出结果.

通过总样本抽取的15%的交叉验证集(共13 746条)来验证上述模型的准确性. 采用 $R^2$ 值和均方误差(mean square error, MSE)来衡量预测结果的准确性. 将本文的RF模型与线性回归模型(linear regression, LR)以及梯度提升决策树(gradient boost decision tree, GBDT)模型进行对比.

图6显示了RF与上述2个模型比较的结果,无论是 $R^2$ 或是MSE,本文提出的RF模型在结果上都要优于GBDT和LR两个模型.

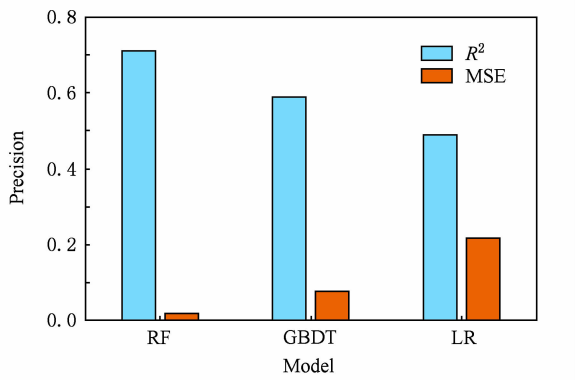



Fig. 6 Comparison of precision between RF, LR and GBDT.

图6 RF,LR与GBDT模型准确度比较

RF模型可以计算出特征对预测结果的影响力. 在模型学习之后,每个特征都会被赋予一个影响因子. 从图7可以看出,上线时长的影响因子最大,节目越新,流行度越高;导演和演员对节目的流行度也存在较大影响;其余特征按影响力大小依次为节目类型、节目时长以及节目内容.

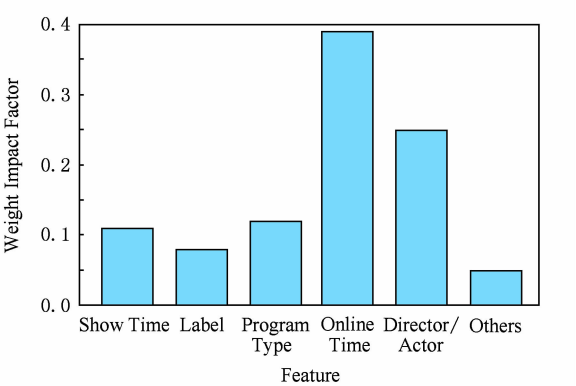



Fig. 7 Weight of features in RF model.

图7 RF模型中各影响因子权重

### 3.4 基于流行度的缓存策略

在互联网+电视平台中,用户对某个节目的首次收视请求将被提交给中心服务器处理,随后由缓存替换算法决定该节目是否调度到边缘服务器缓存中,以降低后续收视对网络带宽的消耗,同时也可以缩短服务的请求响应时间. 相对互联网+电视整个片库的容量,边缘服务器的缓存容量十分有限. 因此,系统需要从片库中选择最合适的节目存储到缓存中,同时需要将用户不感兴趣的节目从缓存中删除以释放缓存空间.

掌握节目流行度的变化,可以有效地提升缓存命中率,对提升服务质量有重要作用. 本节利用节目流行度预测模型的输出结果,作为节目缓存调度的依据,设计了一种缓存调度算法.

好的缓存策略应在有限的空间里尽可能多地缓存最近用户可能收看的节目. 常见的算法包括LRU算法和LFU算法. LFU算法记录了最近一个时间窗口内某个节目被播放的次数,时间窗口的长度作为一个可调参数. 每次当新节目上线后,CDN服务器会将上一个时间窗口内访问最少的节目丢弃. 如果节目流行度是静态,且时间窗口选择的足够大,LFU算法在理论上是能够选择出最流行的影片进行缓存的. 与LFU算法类似,LRU算法保存了最近用户收看的节目. LRU算法记录了缓存中每个节目最近一次的访问时间,并用最新播放的非缓存节目代替长时间无人访问的节目.

互联网+电视平台的节目流行度是高度动态变化的,这点在新闻类节目流行度的变化上尤其明显. LRU算法和LFU算法无法预测新上线的节目流行度,也无法适应节目流行度的剧烈变化. 根据之前的观测数据,部分节目的流行度在上线后迅速爬升,无法被LFU算法捕捉到. 因此,互联网+电视的缓存策略一方面要能侦测到流行度相对稳定的老节目,另一方面也要在历史收视数据很少的情况下捕捉到用户感兴趣的新节目. 为实现上述目标,采用了节目流行度预测数据,作为缓存调度的依据.

**算法1.** 基于流行度的缓存替换算法.

输入:置换窗口大小 $W$、缓存节目集 $CT$、节目集 $S$、置换节目集 $RT$;

输出:置出节目集 $OT$、置入节目集 $IT$.

初始化:根据流行度将 $CT$ 中的节目进行升序排序.

① for $p$ in $sorted(CT)$
② if ((size of $RT$)$<W$)
③ $RT=RT+p$;
④ else:
⑤ break;
⑥ end if

⑦ end for
⑧ for $t$ in $(S-CT)$
⑨ for $q$ in $RT$
⑩ if $(P_t>P_q)$
⑪ $OT=OT+q$;
⑫ $IT=IT+t$;
⑬ else
⑭ continue;
⑮ end if
⑯ end for
⑰ end for
⑱ return $OT$, $IT$.

基于流行度预测的缓存算法 1 所示,该算法设置大小为 $W$ 的缓存窗口,缓存窗口内的节目是缓存中流行度排名靠后的 $W$ 个节目.在缓存置换时,仅将主存内的节目与缓存窗口内的节目进行流行度的比较,将比置换窗口中节目流行度高的节目调入缓存,并将相应的低流行度节目调出缓存,这样可以有效地减少比较次数,提高缓存替换算法的响应速率.

## 4 性能评估

采用某广电运营商互联网+电视平台的用户行为记录作为性能评估的原始数据.以天为单位计算节目收视次数和调度缓存内容的时间间隔,比较 LRU,LFU,PPRA 三种缓存调度策略命中率随缓存数目的变化.其中,命中率是指节目请求的命中率,命中率的计算周期为 7 d,实验结果如图 8 所示.从图 8 可以看出,缓存算法的命中率与缓存大小之间正相关,而且,当缓存较小时缓存的增大对于命中率的提高幅度较大,但随着命中率的提高,缓存大小对于命中率的提升效果减弱.从 2.1 节的数据集介绍中可以得出,每天的平均电视节目请求数为 19 098,为了比较不同缓存策略的差异,取日均节目请求数的 5% (955 个)作为观测点.从图 8 可以看出,当缓存总请求量 5%的电视节目时,LRU 算法的缓存命中率为 57%,LFU 算法的缓存命中率为 59%,PPRA 的缓存命中率为 79%,而且,在缓存节目大小相同时,PPRA 算法的缓存命中率明显优于 LRU 算法和 LFU 算法的缓存命中率;同时,为了达到一定的缓存命中率,PPRA 策略所需的缓存空间要远小于 LFU 算法和 LRU 算法,例如要达到 80%的缓存命中率,LRU 和 LFU 需要缓存 2 987 个节目,而 PPRA 算法仅需缓存 1021 个,从而可以有效地节省缓存空间、降低系统建设成本.

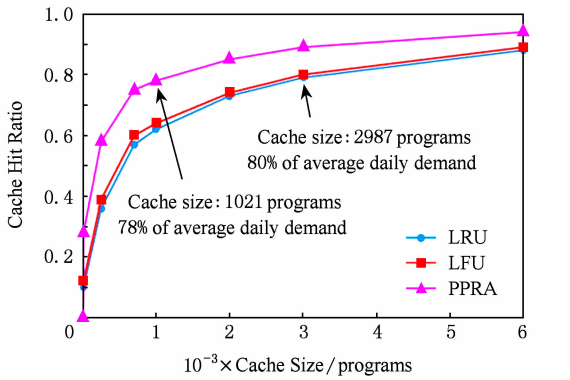



Fig. 8 Cache hit ratio vs cache size for three algorithms.
图 8 不同算法缓存命中率和缓存容量的对比

同时,选取日均节目的 5%作为缓存的容量,使用 120 d 的用户收视数据检验 PPRA 缓存算法缓存命中率,并将 PPRA 算法的缓存命中率与 LRU 算法和 LFU 算法的缓存命中率进行比较.

基于 17 周的观测数据,我们对 PPRA 算法、LRU 算法和 LFU 算法的缓存命中率进行了对比,发现 3 种算法的缓存命中率均随时间发生一定的波动.图 9 展现了 17 周中 3 种算法缓存命中率最高一周的波动情况.在整个 17 周中,PPRA 算法最大缓存命中率是随时间发生周期变化的,且大部分时间高于 LRU 算法和 LFU 算法,在系统负载最大时达到最高值,在黄金时间(每天 20:00—21:00)命中率超过 80%;而 LRU 算法和 LFU 算法的缓存命中率稳定在 50%~60%.实验结果表明,PPRA 算法能够有效地提高最大缓存命中率.

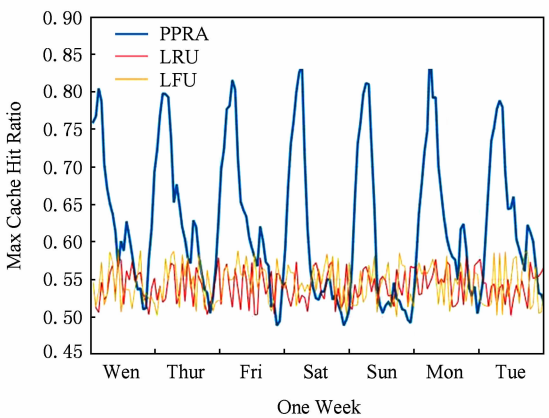



Fig. 9 Comparison of max cache hit ratio.
图 9 最大缓存命中率的比较

图 10 展现了 17 周中 3 种算法缓存命中率取得中间值时一周的波动情况.在整个 17 周中,PPRA 算法缓存命中率是随时间发生周期变化的,且大部分时间高于 LRU 算法和 LFU 算法,在系统负载最

大时达到最高值,在黄金时间(每天 20:00—21:00)命中率超过 70%;而 LRU 算法和 LFU 算法的缓存命中率稳定在 45%～55%. 实验结果表明,PPRA 算法相比 LRU 算法和 LFU 算法,对缓存命中率中值有很好的提升.

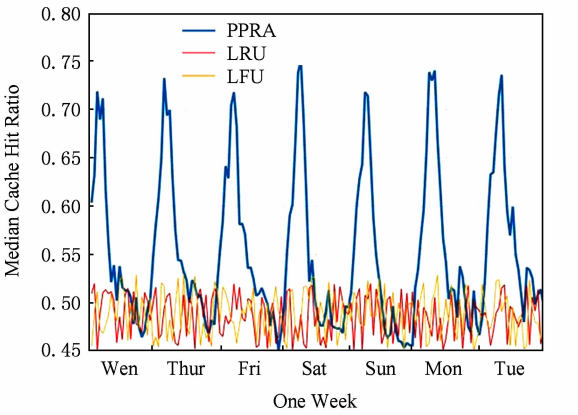



Fig. 10 Comparison of median cache hit ratio.

图 10 缓存命中率中值的比较

图 11 展现了 17 周中 3 种算法缓存命中率最低一周的波动情况. 在整个 17 周中,PPRA 算法最大缓存命中率是随时间发生周期变化的,在系统负载最大时达到最高值,在黄金时间(每天 20:00—21:00)命中率超过 80%,远高于同期的 LRU 算法和 LFU 算法最小缓存命中率;而 LRU 算法和 LFU 算法的缓存命中率稳定在 40%～50%. 实验结果表明,PPRA 算法的最大缓存命中率优于 LRU 算法和 LFU 算法.

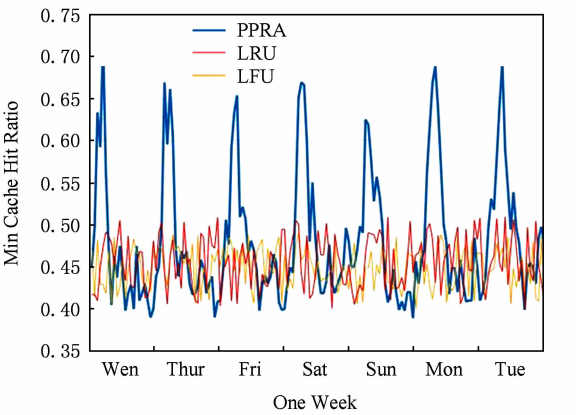



Fig. 11 Comparison of min cache hit ratio.

图 11 最小缓存命中率的比较

综上所述,可以得出以下结论:

1) PPRA 算法采用日均节目的 5%作为缓存空间,可以达到平均 70%的缓存命中率.

2) PPRA 算法消耗的缓存空间明显低于相比传统算法(LRU,LFU),达到 80%缓存命中率时只消耗相当于传统算法 30%的空间.

3) 在互联网+电视系统负载最高时,PPRA 算法的命中率达到最高,较好地契合了互联网+电视节目流行度随时间的变化规律.

## 5 结 论

为了解决互联网+电视平台以提高热点节目命中率而过渡消耗存储空间的问题,本文首先从互联网+电视平台真实数据中揭示了与节目流行度关系紧密的若干因素,定性分析了相应的关系,并根据这些关系选取节目上线时间、类型、标签、播放时间作为特征,基于 RF 算法和 PCA 算法,构建了节目流行度预测模型,并提出一种基于节目流行度的缓存调度算法——PPRA. 该算法在保证缓存命中率的同时能有效节省存储空间. 实验结果表明,在相同节目缓存命中率下,PPRA 算法的空间开销最多可缩减至 LRU 算法的 30%,降低了互联网+电视平台的建设成本.

作为后续研究工作:拟分析用户其他行为(如使用社交网络、移动终端)对节目流行度的影响,进一步挖掘用户收视兴趣,精细化该预测模型,提高预测准确度,并部署在真实的互联网+电视平台上进行验证和应用.

## 参 考 文 献

[1] Fricker C, Robert P, Roberts J. A versatile and accurate approximation for LRU cache performance [C] //Proc of the 24th Int Teletraffic Congress. Krakow, Poland: International Teletraffic Congress, 2012: 1-8

[2] Hendrantoro G, Affandi A. Early result from adaptive combination of LRU, LFU and FIFO to improve cache server performance in telecommunication network [C] //Proc of Int Seminar on Intelligent Technology and Its Applications (ISITIA). Piscataway, NJ: IEEE, 2015: 429-432

[3] Wang Z, Sun L, Wu C, et al. Guiding Internet-scale video service deployment using microblog-based prediction [C] // Proc of IEEE INFOCOM 2012. Piscataway, NJ: IEEE, 2012: 2901-2905

[4] Vallet D, Berkovsky S, Ardon S, et al. Characterizing and predicting viral-and-popular video content [C] //Proc of the 24th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2015: 1591-1600

[5] Kong Qingchao, Mao Wenji. Predicting popularity of forum threads based on dynamic evolution [J]. Journal of Software, 2014, 25(12): 2767-2776 (in Chinese)
(孔庆超, 毛文吉. 基于动态演化的讨论帖流行度预测. 软件学报, 2014, 25(12): 2767-2776)

[6] Ding W, Shang Y, Guo L, et al. Video popularity prediction by sentiment propagation via implicit network [C] //Proc of the 24th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2015: 1621-1630

[7] Figueiredo F. On the prediction of popularity of trends and hits for user generated videos [C] //Proc of the 6th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2013: 741-746

[8] Pinto H, Almeida J M, Gonçalves M A. Using early view patterns to predict the popularity of YouTube videos [C] //Proc of the 6th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2013: 365-374

[9] Ahmed M, Spagna S, Huici F, et al. A peek into the future: Predicting the evolution of popularity in user generated content [C] //Proc of the 6th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2013: 607-616

[10] Roy S D, Mei T, Zeng W, et al. Towards cross-domain learning for social video popularity prediction [J]. IEEE Trans on Multimedia, 2013, 15(6): 1255-1267

[11] Yu H, Xie L, Sanner S. Twitter-driven YouTube views: Beyond individual influencers [C] //Proc of the ACM Int Conf on Multimedia. New York: ACM, 2014: 869-872

[12] Chang B, Zhu H, Ge Y, et al. Predicting the popularity of online serials with autoregressive models [C] //Proc of the 23rd ACM Int Conf on Information and Knowledge Management. New York: ACM, 2014: 1339-1348

[13] McParlane P J, Moshfeghi Y, Jose J M. Nobody comes here anymore, it's too crowded; predicting image popularity on Flickr [C] //Proc of Int Conf on Multimedia Retrieval. New York: ACM, 2014: 385-397

[14] Abrahamsson H, Nordmark M. Program popularity and viewer behaviour in a large TV-on-demand system [C] //Proc of the 2012 ACM Conf on Internet Measurement. New York: ACM, 2012: 199-210

[15] Balachandran A, Sekar V, Akella A, et al. Analyzing the potential benefits of CDN augmentation strategies for Internet video workloads [C] //Proc of the 2013 ACM Conf on Internet Measurement. New York: ACM, 2013: 43-56

[16] Zhou Y, Chen L, Yang C, et al. Video popularity dynamics and its implication for replication [J]. IEEE Trans on Multimedia, 2015, 17(8): 1273-1285

[17] Gouta A, Hong D K, Kermarrec A M, et al. HTTP adaptive streaming in mobile networks: Characteristics and caching opportunities [C] //Proc of the 21st IEEE Int Symp on Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS). Piscataway, NJ: IEEE, 2013: 90-100

[18] Abrahamsson H, Bjorkman M. Caching for IPTV distribution with time-shift [C] //Proc of 2013 Int Conf on Computing, Networking and Communications (ICNC). Piscataway, NJ: IEEE, 2013: 916-921

[19] Nencioni G, Sastry N, Chandaria J, et al. Understanding and decreasing the network footprint of catch-up TV [C] //Proc of the 22nd Int Conf on World Wide Web. New York: ACM, 2013: 965-976

[20] Agrawal K M, Venkatesh T, Medhi D. A dynamic popularity-based partial caching scheme for video on demand service in IPTV networks [C] //Proc of the 6th Int Conf on Communication Systems and Networks (COMSNETS). Piscataway, NJ: IEEE, 2014: 1-8

[21] Akhtar S, Beck A, Rimac I. HiFi: A hierarchical filtering algorithm for caching of online video [C] //Proc of the 23rd Annual ACM Conf on Multimedia. New York: ACM, 2015: 421-430

[22] Park J G, Choi H, Lee B C. Content caching with bi-level control for efficient IPTV content steaming service [C] //Proc of 2014 Int Conf on Information Networking (ICOIN). Piscataway, NJ: IEEE, 2014: 439-443

[23] Wang Yonggong, Li Zhenyu, Wu Qinghua, et al. Performance analysis and optimization for in-network caching replacement in information centric networking [J]. Journal of Computer Research and Development, 2015, 52(9): 2046-2055 (in Chinese)
(王永功,李振宇,武庆华,等. 信息中心网络内缓存替换算法性能分析与优化[J]. 计算机研究与发展, 2015, 52(9): 2046-2055)

[24] Biau G. Analysis of a random forests model [J]. Journal of Machine Learning Research, 2012, 13(1): 1063-1095

[25] Candès E J, Li X, Ma Y, et al. Robust principal component analysis? [J]. Journal of the ACM, 2011, 58(3): 1-37

**Zhu Chengang**, born in 1982. PhD candidate. Member of China Computer Federation. His research interests include networking measurement and behavior analysis, future networks.

**Cheng Guang**, born in 1973. PhD, professor, PhD supervisor. Senior member of China Computer Federation. His main research interests include networking measurement and behavior analysis, future networks.

**Hu Yifei**, born in 1990. Master candidate. His main research interests include networking measurement and behavior analysis, future networks.

**Wang Yuxiang**, born in 1990. Master candidate. His main research interests include networking measurement and behavior analysis, future networks.