联系电话：xxxxx　（必留）
联系邮箱：?????　（必留）
投稿邮箱：icr@cuc.edu.cn

（文章篇幅：8000-10000 字）

延迟命中场景下基于学习方法的缓存性能优化

沈志[1]，江博闻[1]，江波[1*]，林涛[2]
（1. 上海交通大学，上海 200240；2. 中国传媒大学，北京 100024）

摘要：传统的缓存算法大多基于简单的统计信息进行内容替换，在绝大部分场景下都和离线最优算法有着较大的性能差距。当前基于机器学习来设计高效的缓存策略基本都假设请求的大小相等，忽略了真实场景中请求的大小往往不等且大小变化的范围较大。由于传输速度已接近极限，……实验结果表明，ARC-learning+算法在流行度变化较快情况下的时延提升明显优于已有的其他缓存算法。在其他流行度变化情况下也非常接近已有算法的最优时延性能。
关键词：基于学习；非平稳流行度；内容变大小；延迟命中
中图分类号： ?????　**文献标识码：**A

A learning-based method to optimize cache performance with delayed hit

SHEN Zhi[1], JIANG Bowen[1], JIANG Bo[1*], LIN Tao[2]
(1. Shanghai Jiao Tong University, Shanghai 200240, China; 2.Communication University of China, Beijing 100024, China)

Abstract: Most traditional caching algorithms rely on basic statistical data for content replacement, creating a significant performance discrepancy in comparison to offline optimal caching algorithms. However, due to improvements in CPU performance, cache strategies based on machine learning have been developed in recent years to enhance cache performance. …….Experimental results show that the performance of the modified sorting function algorithm is better than other existing caching algorithms when the popularity changes quickly. It is also very close to the optimal delay performance of the existing algorithm under other prevalence changes.
Keywords: learning-based; non-stationary popularity; variable object size; delayed hit

基金项目：国家自然科学基金面上项目（62072302）；地区科学基金项目（62262018）；“媒体融合与传播国家重点实验室（中国传媒大学）”开放课题(SKLMCC2021KF011)
作者简介(*为通讯作者)：沈志(1998-),男,硕士研究生,主要从事缓存技术研究。Email：

shangjiao4313888@sjtu.edu.cn; 江波(1983-), 男, 博士, 副教授, 主要从事系统性能评测与优化研究。Email: bjiang@sjtu.edu.cn

1 引言

随着视频音频等各类应用的发展，…17 万多个服务器上服务数万亿用户请求[1]。

唐慧如等[2]提出了…

CDN 服务器位于用户和广域网(WAN)之间。…这些开销在研究中通常用内容命中率(Object Hit Ratio，OHR)[3-6]，时延(Latency)[7][8]来衡量，….。

2 研究背景介绍

2.1 请求内容变大小

…

3 ARC-learning 算法介绍和改进

3.1 ARC-learning 算法介绍

ARC-learning[9]是基于 LRB(Learning Relaxed Belady)[10]，为适应各种流行度变化设计的算法，该算法有效降低了流行度变化较快场景下，…。ARC-learning 算法框架如图 1 所示。

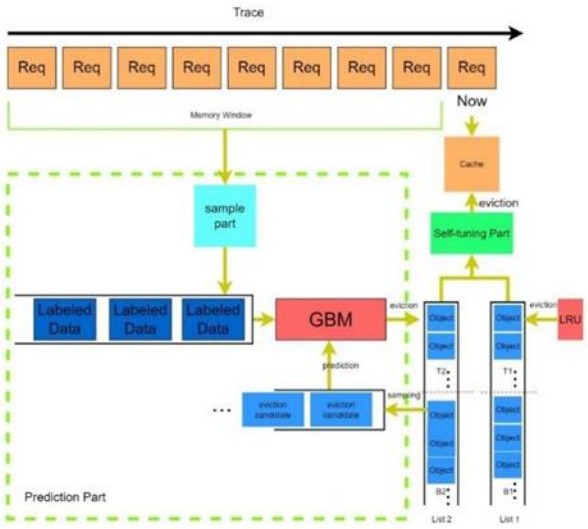

图 1 ARC-learning 缓存替换框架

本节使用了文献[5]在优化时延时所使用的排序函数对 ARC-learning 进行改进。文献[5]对请求未命中时候由完全未命中和延迟命中产生的累计时延……

流行度记为λ_i，内容i在未命中时，从上一级服务器取内容所产生的时延记为L_i，其排序函数定义如公式（1）：

$$\tilde{f}(i) = \frac{Aggregated\ latency}{Mean\ latency\ for\ each\ request} = \frac{D_i}{\bar{D}_i} = \frac{\lambda_i L_i(1+\lambda_i L_i)}{2+\lambda_i L_i} \quad (1)$$

考虑到变大小场景下应当尽可能缓存体积较小的内容以为更多的用户提供服务，上述公

式可进一步优化为:

$$f(i) = \frac{\tilde{f}(i)}{s_i} = \frac{\lambda_i L_i (1+\lambda_i L_i)}{(2+\lambda_i L_i) s_i} \tag{2}$$

排序函数为:　　（注：六个字以内顶格）

$$f(i) = \frac{\lambda_i L_i (1+\lambda_i L_i)}{(2+\lambda_i L_i) s_i} = \frac{L_i (p_i + L_i)}{(2p_i + L_i) s_i p_i} \tag{3}$$

……. （常量用正体，变量用斜）

与….模型进行了对比验证。实验结果如表 1 所示。

表 1 在 NLPCC 数据集上的结果

模型 \ 评价指标	Rouge-1	Rouge-2	Rouge-L	BS
T5 PEGASUS	56.15	39.47	47.87	65.63
SimCLCTS (Ours)	**57.24**	**41.40**	**49.52**	**73.33**

5　结论

……

参考文献（References）:

[1] Song Z，Berger D S，Li K，et al. Learning relaxed belady for content distribution network caching ［C］//17th USENIX Conference on Networked Systems Design and Implementation（NSDI ' 20），2020: 529-544.

[2] 陈嘉猷，鲍怀翘，郑玉玲. 普通话中塞音、塞擦音、嗓音起始时间（VOT）初探［C］//中国声学学会 2002 年全国声学学术会议论文集，2002.

[3] Urdaneta G, Pierre G, Van S M. Wikipedia work-load analysis for decentralized hosting[J]. Computer Networks, 2009, 53(11): 1830-1845.

[4] 唐慧如. 技术资源管理系统资源使用确认及计费设计［J］.现代电视技术，2022（02）:146-148.

[5] Fromkin H L. Affective and Valuational Consequences of Self-Perceived Uniqueness Deprivation［D］. Columbus，OH：Ohio State University,1968.

[6] 曾余洋. 基于深度学习的中文文本情感分析研究［D］. 雅安：四川农业大学，2022.

[7] Snyder C R，Fromkin H L. Uniqueness：The Human Pursuit of Difference［M］.New York：Springer，1980.

[8] 王理嘉，林焘. 语音学教程［M］. 北京：北京大学出版社，1992.

[9] ITU-R. BS.1284-2 ： General methods for the subjective assessment of sound quality [S/OL]. [2019-01-21].https://www.itu.int/rec/R-REC-BS.1284-2-201901-I/en.

[10] 国家广播电影广电总局. GB/T 26252-2010:VHF/UHF 频段地面数字电视广播频率规划准则 [S]. 北京：中国标准出版社，2011.

[11] 中华人民共和国国家卫生健康委员会. 新型冠状病毒肺炎疫情防控疫情通报[EB/OL]. (2022-12-25)[2022-12-28].http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml.

[12] Goodfellow I J，Shlens J，Szegedy C. Explaining and harnessing adversarial examples［DB/OL］. arXiv：1412.6572，2014.

时态要求

英语论文摘要的时态选择.

①如果是描述研究的成果、得出的结论，用一般现在时。 ②如果描述具体的研究过程，则用一般过去时。③如果描述研究结果对未来的影响，用一般将来时。④如果是过去的研究成果，但是对现在得出的结论有影响，用现在完成时。⑤如果是引用已经成为公认的事实，则用一般现在时。