文章编号：1003-0077(2018)10-0087-11

基于网络小说热度预测的 CDN 内容分发策略研究

赵礼强，姜 崇，靖 可

（沈阳航空航天大学 经济与管理学院，辽宁 沈阳 110000）

摘 要：内容分发网络(CDN)以推(Push)和拉(Pull)作为两种主要的内容分发策略。拉是服务器对用户请求给予回应的方式，在面对高并发请求时，以推方式预分发到服务器上的内容将有效辅助改善服务器过载的情况，并实现热度内容的主动推送。预分发内容的热度高低直接决定了内容的边缘命中率和用户的启动延迟，有效减少预分发内容替换的频率和对源服务器访问时的网络堵塞，因此在网络小说服务器中依据网络小说热度对内容分发策略的部署就显得更为重要。目前针对网络小说服务器的内容分发策略的研究较少，缺乏对网络小说热度科学有效的评价标准。以服务器管理人员的主观经验判断和低命中的预分发内容的不断替换来实现内容推送的策略，不仅主观性强，同时内容的不断替换更极大地增加了服务器负担。针对这一问题，该文通过定义网络小说热度概念，对在起点中文网爬取到的数据进行预处理，数据拟合显示数据符合幂律分布规律，并建立热度等级评价标准，分别采用贝叶斯网络、随机森林算法与 Logistic 回归建立预测模型，对网络小说热度预测进行对比研究。结果显示，随机森林算法的预测正确率达到 97.097%，均方误差为 0.112 8，分类预测效果更优，且误差率更低。因此选用随机森林算法，依据网络小说热度评价标准，能够有效解决网络小说 CDN 系统内容部署不准确而导致低命中内容的不断替换和用户访问延迟的问题，为内容分发策略提供有效指导，从而提高内容命中率，提升网络小说 CDN 系统运作效率。

关键词：网络小说热度；幂律分布；内容分发策略；随机森林算法；预测

中图分类号：TP391　　**文献标识码**：A

Content Delivery Strategy Based on Popularity Prediction of the Network Novels

ZHAO Liqiang, JIANG Chong, JING Ke

(College of Economics and Management, Shenyang Aerospace University, Shenyang, Liaoning 110000, China)

Abstract: At present, there are few researches on the content distribution strategy for the internet novel servers, and lack of scientific and effective evaluation criteria for the popularity of the network novel. This paper proposes to measure the popularity of network novels on the novels retrieved from Qidian (www.qidian.com). Bayesian network, random forest algorithm and Logistic regression are applied to establish the prediction model, and the random forest out-performs with 97.097% accuracy(subject to 0.112 8 MSE). With this method, the problem of inaccurate deployment of low-hit novels in CDN system and user access delay can be alleviated, so as to provide effective guidance for content distribution strategy and improve content Hit rate.

Keywords: network novels heat; power law distribution; content distribution strategy; random forest algorithm; forecast

0 引言

随着网络信息的海量爆发，受限于互联网服务器本身网络带宽处理能力，面对海量信息传输与分享，需要多次网络转发，导致传输延时高且不稳定，降低响应速度。内容分发网络(content delivery network，CDN)就是为了有效解决此类问题，在现

收稿日期：2017-11-21　**定稿日期**：2018-03-17

基金项目：国家自然科学基金(71301108)；教育部人文社科规划基金(17YJA630139)；辽宁省教育厅重点项目(L201706)

有的互联网基础上通过放置节点服务器于网络各处,从而构成一层智能虚拟网络架构。CDN系统通过分布式缓存/复制、负载均衡、流量控制及客户端重定向等技术[1],当用户对业务内容发起请求时,将请求重新导向距离用户最近的服务节点上,更快、更精准地触发信息和触达每一位用户,为用户带来更优越的使用体验。CDN在保障信息连续性的前提下,尽可能减少资源的转发、传输、链路抖动等操作,有效解决网络传输拥堵和用户访问延迟的问题,在流媒体与动态内容传输方面得到了广泛应用。

当对CDN进行内容副本部署时,过多的副本部署会增大空间占有率,降低空间有效利用率,过少的部署则会降低服务质量。只有部署流行度更高的内容副本时,才能提高系统效率[2]。内容副本部署策略不佳而进行的后期服务器调整会增加I/O负担,耗费周期长,缺乏经济性[3]。因此合理优化内容副本部署策略是实现CDN优质服务的前提[4]。从用户角度分析内容接收者的特征规律是一种优化内容部署策略的研究思路。例如,借助信任度将信任机制引入对用户的内容推送模型中,实现推策略[5]。以往的研究多是从用户兴趣的角度出发,将用户兴趣与内容内外部流行度结合,对用户群体有针对性地进行推送服务[6]。相关研究采用聚类的方式挖掘用户之间的关系,向用户推送相似内容[7],或通过挖掘节点随机运动中隐藏的用户社交特征和兴趣特征,结合信息需求量,实现最大效用的内容推送[8]。这种基于用户兴趣来挖掘用户关系和相似性从而实现内容推送的策略在微博的内容部署[9]和新闻内容部署[10]中都得到了应用。

在流媒体服务器的内容分发策略部署中,针对视频持续时间长、文件大等特点,芮兰兰等人[11]结合内容流行度和节点中心度的缓存策略解决了缓存冗余的问题,合理分配资源,提高了整体效用。但研究的重点集中在缓存的技术层面,对内容流行度的判别及与排名匹配标准的研究不足。熊庆昌等人通过研究用户访问规律,根据内容内外部流行度的分布情况而提出影片生存期的缓存技术,对内容分发策略的部署具有很好的参考价值[12];在杨传栋等人的研究中同样指出,由于流媒体内部流行度差距巨大,而提出采用不同的分段方法对流行度不同的内容进行部分推送策略[13]。综上可以发现,虽然对用户兴趣研究的角度不同,但最终落脚点依旧是内容的流行度,说明内容的流行度才是内容分发部署策略的关键。一方面,推技术虽然更适合内容请求集中的多媒体热度内容,但由于缺乏对内容的预测机制,当用户请求没有被预分发内容命中,请求远端源服务器时产生的网络堵塞现象势必会对用户体验造成负面影响[14]。另一方面,虽然流媒体服务器的研究对网络小说服务器的内容分发有启示作用,但针对流媒体持续时间长、文件大等特点提出的缓存策略并不适用于网络小说。因此,从网络小说热度作为切入点研究内容分发策略就显得更加适用和重要。目前通过结合微博[15]与网络搜索[16]对电影票房和电视剧点播量进行预测,并挖掘票房和点播量影响因素的研究较多,但针对网络小说热度的研究以定性研究为主,缺乏科学的热度评价标准,更缺乏针对网络小说服务器内容分发策略而对网络小说热度进行预测的研究。

本文通过定义网络小说热度概念,建立网络小说热度评价标准,采用分类算法对网络小说热度进行预测,旨在为高热度网络小说副本以合理优化的策略部署到CDN系统中提供依据[17],减少后期对内容副本的调整,减轻I/O负担,降低访问延迟,提高CDN系统服务质量。

1 数据的获取与预处理

起点中文网隶属于国内最大的数字内容综合平台——阅文集团,是国内最大文学阅读与写作平台之一,也是目前国内领先的原创文学门户网站,树立了行业领导地位,具有很高的影响力。

起点中文网包含大量拥有庞大阅读基群的优质网络小说,又囊括了众多处在成长期的新生网络小说,个例样本鲜明,整体样本题材丰富,使数据更全面充分,因此本文选择起点中文网作为数据来源。

1.1 数据来源

本文选择起点中文网作为网络小说数据获取源网站,采用八爪鱼数据采集器作为数据采集工具。起点中文网网络小说页面的数据主要分为两种,一种是不进行周期清零,从网络小说创作开始,数据值随着每天的增长而不断的累积,如总点击量、总推荐量等特征。另一种是积累一定周期后清零,新周期内重新统计的数据。如月票数以月为周期,月统计数据在月末清零,周打赏人数、周会员点击、周推荐量等特征则是以周为周期,周统计数据在每周末清零。

针对起点中文网的这一规律,本文选择2017年

6 月 30 日作为采集数据的时间节点对网络小说页面数据进行抓取，旨在得到六月份网络小说月票的月统计数，同时该时间节点恰好作为六月份最后一周的周末，从而得到周打赏人数、周会员点击、周推荐等特征的周统计数据。当一部网络小说进入成熟期时，粉丝群体相对稳定，周期数据增长量应当保持相对稳定，而能够保持稳定增长的网络小说热度更持久，在网络分发内容策略中需要被替换的概率更低。由于本文对网络小说热度的预测是一个状态预测，因此选用结合历史累积的数据特征和能反映常态的周期统计数据特征作为网络小说抓取的对象，因为不考虑特定时间段而抓取的数据更能真实反映日常网络小说的热度情况。对起点中文网原创风云榜的 501 部网络小说排名信息及网页数据进行抓取，作为网络小说热度预测的初始知识库。同时抓取 5 649 部有人气排名但缺乏热度评价的网络小说作为热度预测数据库。

1.2　数据特征

网络小说作为文学作品，具有文学价值但难以衡量转化为数值信息的特征，因此很难通过网络小说本身的内容分析而获得量化信息。但通过读者对网络小说点击量、推荐量、打赏、评论等特征以及作者创作网络小说的相关信息则可以从侧面反映网络小说的受欢迎程度[18]，即网络小说的热度。本文根据先验知识与相关文献参考，针对影响网络小说热度的特征在各个维度上进行选择[19]，具体变量选择及数据描述如表 1 所示。

表 1　变量定义与数据描述

变量	最小值	最大值	平均值	标准差
作品总字数	1.02	8 475	189.173	257.773
作者创作天数	1	4 412	869.605	868.180
累计创作字数	0.8	3 779.27	447.305	556.500
创作作品数	1	18	2.767	2.344
作者号召力	0	1	0.112	0.316
签约状态	0	1	0.935	0.246
连载状态	0	1	0.718	0.450
总点击量	598	150 502 700	1 266 659.230	4 987 495.590
总推荐数	1	13 565 100	207 537.954	717 913.661
作品会员周点击量	0	233 700	844.047	9 937.182
月票数量	0	55 209	283.356	1 539.234
周推荐数	0	254 900	1 114.379	8 019.463
周打赏人数	0	1 004	2.430	15.786
评论数	20	872 185	9 294.709	36 208.460
都市题材	0	1	0.248	0.432
二次元题材	0	1	0.099	0.299
军事题材	0	1	0.024	0.153
科幻题材	0	1	0.124	0.329
历史题材	0	1	0.102	0.303
灵异题材	0	1	0.026	0.159
奇幻题材	0	1	0.036	0.186
体育题材	0	1	0.022	0.147

续表

变量	最小值	最大值	平均值	标准差
武侠题材	0	1	0.018	0.134
仙侠题材	0	1	0.087	0.283
现实题材	0	1	0.010	0.102
玄幻题材	0	1	0.130	0.336
游戏题材	0	1	0.073	0.260

2 热度定义

热度是一个虚拟概念,也是一个综合性的评价指标,用来衡量作品的受欢迎程度或销售情况等。热度概念在电影、电视剧的预测问题研究中较为普遍。电影热度一般以票房作为表征进行分析预测,电视剧热度一般以单集电视剧点播量作为表征进行分析预测。电影更倾向于一次性消费,通过设定票价与销售票数的积累来获得收益,电视剧则更倾向于一段时间内的持续消费,周期性播放的电视剧吸引的流量表现在点播量的积累上,从而获得相应收益。比较之下,网络小说则是综合了电影、电视剧的双重特点,同时具有区别于电影和电视剧的特殊特征。

一方面,当网络小说达到上架要求,从章节免费阅读升级为章节VIP阅读后,将会对网络小说每一章节进行定价销售,通过点击量和单章节的定价来获得当天网络小说更新章节的销售收益。因此,某一天某一章节的故事情节决定了当天章节销售收益的高低,这一点与电影票房的概念相似。另一方面,网络小说的章节更新周期是以日为单位,且一本网络小说的完本一般需要持续更新至少一年以上,是一个持续性的消费,每天的点击量积累形成总点击量,大量的点击代表网络小说吸引的人气和阅读基础,这一点与电视剧的点播量概念相似。最后,网络小说拥有显著区别于电影、电视剧的打赏投票机制。读者可以根据个人意愿,以打赏、投月票、投推荐票的方式表达个人对网络小说的喜爱和支持,打赏与月票的收益是与章节销售收益独立区分的收益。值得一提的是,虽然属于个别现象,但不可否认个别网络小说存在刷票、刷点击的行为,造成诸如点击量数据极高、推荐票等其他变量数据极低的畸形现象,如果单一从点击量或其他某个单一变量来反映热度概念,将不可避免受到人为或其他外部因素的干扰。

综合上述分析,本文认为用单一维度来衡量网络小说热度缺乏足够的信服力,具有片面性。因此本文结合相关文献以及网络小说本身的特点进行了变量选择,综合定义网络小说热度,具体如下:

(1) 阅读基群维度。由总点击量、总推荐量、周会员点击量构成。点击量和推荐量能够直观反映网络小说的读者总基群,周会员点击量则反映一周内选择VIP阅读的读者基群。

(2) 阅读收益维度。由月票、周打赏人数和周推荐票数构成。起点中文网采用周清和月清两种方式更新网络小说数据,月票每月月末统计清零,周打赏人数和周推荐票数每周周末统计清零。由于打赏和月票收益与章节销售独立区别,阅读收益反映的是读者在购阅章节之后对网络小说的额外支持度。

(3) 阅读讨论维度。阅读讨论数的多少反映的是读者在阅读网络小说后的感受反馈,也影响着新读者选择阅读的意向。阅读讨论维度体现了网络小说的话题讨论参与热度。

2.1 综合热度评分

本文根据原创风云榜501部排名网络小说建立初始知识库,采用1～4分评分制对每一部网络小说(P)在选择的维度(N、S、D)上进行热度评分,加和得到每一部小说的综合热度评分 H_P。使用符号来标注信息:

(1) 阅读基群(N)、总点击量(N_1)、总推荐量(N_2)、周会员点击量(N_3),i 取值范围为1,2,3。网络小说(P)的阅读基群热度评分,如式(1)所示。

$$\sum_{i=1}^{3} H(N_i^p) \tag{1}$$

(2) 阅读收益(S)、月票(S_1)、周打赏人数(S_2)、周推荐票(S_3),j 取值范围为1,2,3。网络小说(P)的阅读收益热度评分,如式(2)所示。

$$\sum_{j=1}^{3} H(S_j^p) \tag{2}$$

(3) 阅读讨论(D)。阅读讨论热度评分,如式(3)所示。

$$H(D^P) \tag{3}$$

(4) 热度评分(H)。一部网络小说的综合热度评分由三个维度的评分加和获得，如式(4)所示。

$$H_P = \sum_{i=1}^{3} H(N_i^p) + \sum_{j=1}^{3} H(S_j^p) + H(D^p) \tag{4}$$

热度评分数量级如表 2 所示。

表 2　热度评分数量级

阅读基群(N)			阅读收益(S)			阅读讨论(D)	热度评分(H)
总点击量(N_1)($\times 10^3$)	总推荐量(N_2)($\times 10^3$)	周会员点击量(N_3)($\times 10^3$)	月票(S_1)($\times 10^3$)	周打赏人数(S_2)	周推荐票(S_3)($\times 10^3$)	评论数(D)($\times 10^3$)	
≤50	≤50	≤1	≤1	≤10	≤1	≤1	1
[50,500]	[50,500]	[1,5]	[1,5]	[10,50]	[1,5]	[1,10]	2
[500,5000]	[500,5000]	[5,10]	[5,10]	[50,100]	[5,10]	[10,100]	3
≥5000	≥5000	≥10	≥10	≥100	≥10	≥100	4

2.2　数据拟合

经过数据预处理及统计分析发现，初始知识库中原创风云榜 501 部网络小说的综合热度评分 H_P 取值范围为[7,28]。由于缺乏综合热度评分与热度等级之间的对应关系，无法确定网络小说热度等级的取值规律。受到电影影片热度通常符合 $Zipf$ 分布的启发[20]，本文将 501 部网络小说的人气排名与综合热度评分分别采用傅里叶函数、有理函数、幂律分布、样条插值平滑进行数据拟合。数据拟合分布图如图 1～4 所示。

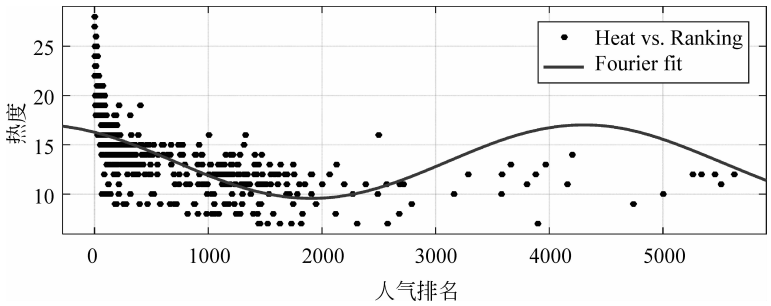

图 1　傅里叶数据拟合图像

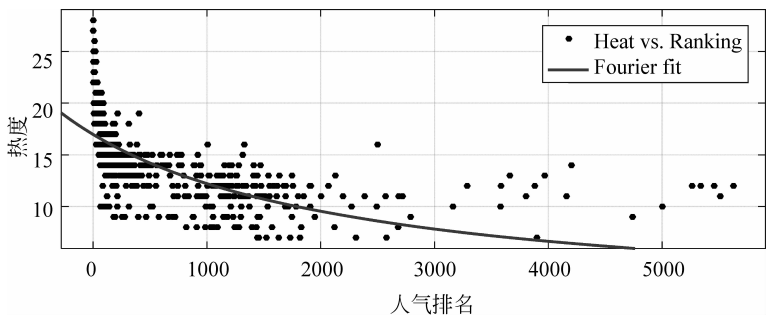

图 2　有理函数数据拟合图像

幂律分布函数为，如式(5)所示。

$$Y = cX^{-r} \tag{5}$$

其中 X,Y 是正的随机变量，c,r 均为大于零的常数。

对比四种数据拟合图像，样条插值平滑数据拟合存在过拟合现象，偏差较大，不予考虑。比较有理函数数据拟合和傅里叶数据拟合，幂律分布数据拟合效果更优。同时幂律分布数据拟合中确定系数为 0.692 2(确定系数 R-square，该值越接近 1 代表拟合程度越好)，傅里叶数据拟合确定系数为 0.405，有理函数数据拟合确定系数为 0.426 6。

根据数据拟合显示，网络小说综合热度评分

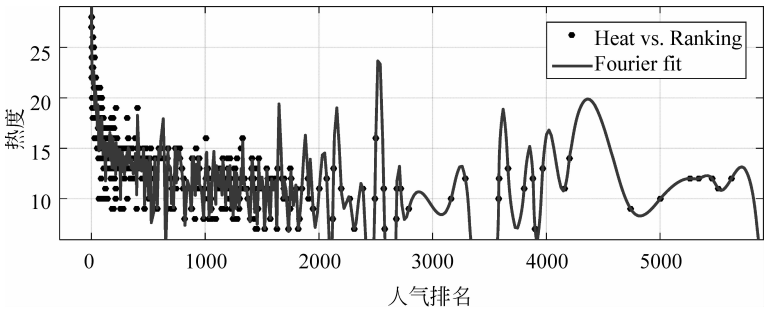

图3 样条插值平滑数据拟合图像

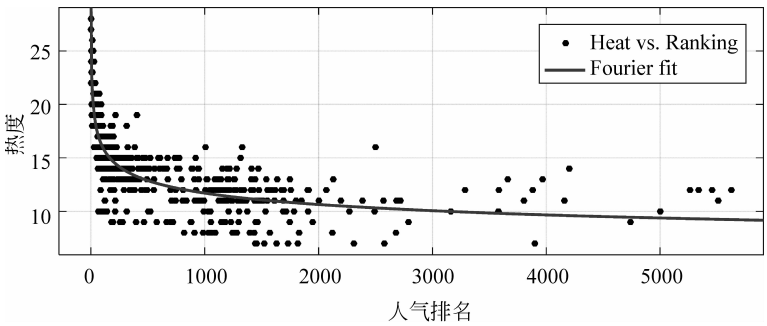

图4 幂律分布数据拟合图像

与人气排名的数据分布符合幂律分布，满足帕累托定则，说明网络小说80%的人气集中在20%的热度网络小说之上，这一点符合当前网络小说的现实认知。

2.3 热度等级划分

根据数据拟合挖掘出的数据分布规律，按照二八定律划分网络小说综合热度评分 H_P，将综合热度评分与冷、一般、热、非常热四级热度等级对应。按照热度评价标准，将抓取到5 649部只有人气排名而没有热度评价的网络小说进行分类。具体的热度评价标准及网络小说作品数的分类情况如表3所示。

表3 网络小说热度评价标准

综合热度评分(H_P)	热度等级	作品数
[7,9]	1(冷)	2 989
[10,11]	2(一般)	1 618
[12,14]	3(热)	877
[15,28]	4(非常热)	165
Min7,Max28		5 649

2.4 特征选择

为了去除冗余特征，根据网络小说的数据类型，使用对数据分布条件要求更宽松的Spearman秩相关系数来验证热度等级特征与其余特征之间的相关关系，Spearman秩相关系数公式，如式(6)所示。

$$R = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{6}$$

其中 N 为样本数，$d_i = X_i - Y_i$。

将秩相关系数 R 的绝对值与Spearman秩相关系数统计表中的临界值 W_p 进行比较，当 $|R| > W_p$ 时则表明变化趋势有显著意义，当 $|R| <= W_p$ 则表明变化趋势没有显著意义。本文样本数 N 为5649，数值较大，统计表中没有给出具体 W_P，但是在同一显著水平下，随着样本数的增大，临界值减少。当 $n=30$ 时，$\alpha = 0.05$ 的置信水平上，查表得：$R = 0.306$。因此判断，当 R 值高于0.306时，认为相关关系显著。具体网络小说热度等级特征与其他特征秩相关系数见表4。

表4 热度相关性斯皮尔曼秩相关系数表

特征	斯皮尔曼		与热度相关性
	R	P	
作者号召力	0.375 289 246	2.02E−188	显著相关
评论数	0.808 858 221	0	显著相关
签约状态	0.183 559 544	5.39E−44	不显著相关

续表

特征	Spearman		与热度相关性
	R	*P*	
累计创作字数	0.544 623 067	0	显著相关
连载状态	−0.315 292 332	1.36E−130	显著相关
月票数量	0.457 951 882	5.62E−291	显著相关
作者创作作品数	0.362 588 809	4.21E−175	显著相关
周推荐数	0.424 821 317	2.20E−246	显著相关
总点击量	0.839 806 545	0	显著相关
总推荐数	0.740 359 329	0	显著相关
总字数	0.433 163 016	3.79E−257	显著相关
周打赏人数	0.346 233 343	8.01E−159	显著相关
会员周点击量	0.475 226 759	2.895 080 8E−316	显著相关
创作天数	NaN	NaN	不显著相关
都市题材	0.039 335 654	7.10E−04	不显著相关
二次元题材	−0.126 629 923	1.16E−27	不显著相关
军事题材	−0.029 103 487	0.012 247 971	不显著相关
科幻题材	−0.003 997 93	0.730 778 337	不显著相关
历史题材	0.057 106 583	8.87E−07	不显著相关
灵异题材	−0.053 079 117	4.91E−06	不显著相关
奇幻题材	−0.011 678 461	0.314 826 465	不显著相关
体育题材	0.012 926 209	0.265 908 562	不显著相关
武侠题材	−0.046 320 482	6.70E−05	不显著相关
仙侠题材	0.054 170 738	3.12E−06	不显著相关
现实题材	0.035 305 135	0.002 376 155	不显著相关
玄幻题材	0.028 777 608	0.013 253 486	不显著相关
游戏题材	−0.016 370 453	0.158 837 075	不显著相关

根据表 4 特征判断结果，去除相关关系不显著的特征，剩余与网络小说热度等级特征有显著相关的特征共有 12 个，据此建立预测网络小说热度等级的数学模型。

2.5　热度预测意义

（1）单独依靠热度评价标准判断热度的滞后性。由于内容分发网络的分发策略是以预分发在服务器上的内容来命中用户对内容的请求，减少用户因无法从边缘网络获得内容而需要请求源服务器的情况，需要提前预见用户可能访问的内容并命中，同时进行热度内容的推送。而在数据生成后的热度评价将很难对预分发内容的部署提供参考和指导，同时也无法根据内容的热度变化情况提前预见并及时调整分发策略。

（2）通过预测机制的应用。首先，可以降低热度评价标准中可能存在的人为因素影响。虽然人为刷票的行为在网站监督和个人自觉的情况下被禁止，但这种行为仍然不可避免。由于无法从数据中判断热度评价标准中选取的七个特征变量是否存在刷票行为，因此借助作者创作字数、连载状态等更多维度的综合衡量，可以有效识别网络小说真实热度等级。其次，本文构建热度评价标准时对网络小说热度等级的分值对照是依照初始知识库中 501 部网络小说的数据拟合得到的，这个标准对 5 649 部网络小说的适用情况是需要通过机器学习来进一步更新特征权重和规律来获得更准确的预测模型。

因此，热度评价标准是作为预测机制应用下的基础构建，是为了实现对网络小说热度预测，从而为预分发内容的判断进行的必要过程，对网络小说内容分发策略的部署提供有效参考和指导。

3　数据挖掘

鉴于以上分析，当新获取网络小说信息时，根据相关性检验，选取作者号召力、评论数、累计创作字数、连载状态、月票数量、作者创作作品数、周推荐数、总点击量、总推荐数、总字数、总打赏人数、会员周点击数共 12 个特征对网络小说热度等级建立预测模型。

由于本文是针对网络小说热度四个等级进行预测，属于分类预测，因此选择贝叶斯网络、逻辑回归、随机森林共三种典型分类算法。在 WEKA 数据挖掘平台进行十折交叉验证预测对比研究，旨在选择更适用 CDN 的算法。[21]

3.1　模型建立

贝叶斯网络通过学习寻找最佳树结构，可以用来表示和推理不确定条件，同时贝叶斯网络在概括朴素贝叶斯分类器的概率分布效果很好，能清晰地反映独立性，作为机器学习工具有很好的分类优势[22]。基于贝叶斯网络建立的网络小说热度预测模型如图 5 所示。

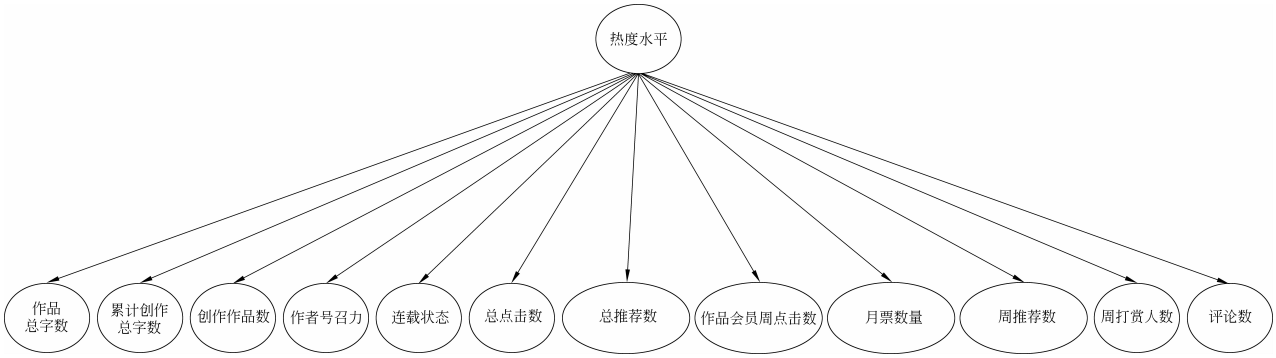

图5 贝叶斯网络

在根据票房划分成功电影的预测中，逻辑回归有很好的应用[23]，因此在这里选择了逻辑回归作为对比算法建立模型，逻辑回归是一种广义线性回归，通过函数 L 将 $w*x+b$ 对应一个隐状态 p，$p=L(w*x+b)$，然后根据 p 与 $1-p$ 的大小决定因变量的值。其本质是为了改变取值区间的矛盾和因变量、自变量之间关系而进行对数变换。实践表明，变换后的因变量与自变量之间一般呈线性关系，从根本上是为解决因变量不是连续变量的约束问题。基于逻辑回归建立的网络小说热度预测模型如表5所示。

表5 逻辑回归模型

变量	Class =1		Class=2		Class=3	
	Coefficients	Odds Ratios	Coefficients	Odds Ratios	Coefficients	Odds Ratios
作品总字数	−10.78	0	−13.25	0	−4.804 8	0.008 2
累计创作字数	−3.36	0.034 6	−2.47	0.084	−2.407 8	0.09
创作作品数	−1.23	0.291 9	−0.23	0.790 4	1.066 5	2.905 1
作品号召力	0.22	1.246 6	−0.13	0.869 8	0.035 9	1.036 5
连载状态	−0.08	0.921	−0.12	0.886 4	−0.476 9	0.620 7
总点击量	−927.30	0	−54.88	0	2.129 5	8.410 3
总推荐数	−90.91	0	−16.03	0	−5.969 8	0.002 6
会员周点击量	−418.65	0	−161.23	0	−4.876 3	0.007 6
月票数量	−267.8	0	−151.31	0	−65.489 2	0
周推荐数	−320.22	0	−100.11	0	−56.661 1	0
周打赏人数	−125.13	0	−124.61	0	−2.95	0.052 3
评论数	−452.87	0	−98.84	0	−6.158 4	0.002 1
二次元题材	14.245 3		10.06		5.942 2	

逻辑回归模型中，Coefficients 代表在 Class 分类下的变量对 administration=1 的系数。Odds Ratios 代表在 Class 分类下的变量对 administration=1 的胜率。系数 b 与 Odds Ratios 的关系式，如式(7)所示。

$$\text{Odds Ratios} = e^{\text{Coefficients}} \tag{7}$$

Class=1 代表热度等级为冷，Class=2 代表热度等级为一般，Class=3 代表热度等级为热，其余样本认为热度等级为非常热。以 Total clicks 变量 Class=3 的 Odds Ratios 的值为 8.410 3 为例解释逻辑回归模型：在其余变量都相同的条件下，总点击量每提高一个单位，网络小说热度分类为热的几率提高 841.03%。说明在其他条件都相同的情况下，点击量越多，网络小说热度等级越高，符合实际情况。通过对逻辑回归模型的观察可以发现，要判断 Class=3，即网络小说热度等级为热时，影响热度等级判断的变量数比判断 Class=1 和 Class=2 的

变量数更多，说明判断网络小说热度等级越高，需要考虑的变量越多，这与本文对热度是一个综合性指标的判断相吻合。

随机森林是结合 Bagging 方法和决策树方法建立的多功能机器学习算法[24]，在随机森林中，不同于 CART 模型只生成唯一的树，而是生成很多决策树，当基于某种属性对对象进行分类判别时，随机森林中的每一棵决策树都会做出自我的分类选择，进行“投票”，输出结果取决于投票结果，分类选项的票数多者胜出，输出该分类选项。

随机森林对变量（列）和数据（行）的随机化使用可以避免过拟合现象，拥有较强的抗噪声能力，无须对数据集进行规范化，可以大量处理高维数据，针对本文网络小说拥有 12 个特征类型数据有着很好的降维效果，同时输出相关属性的重要程度。本文通过随机森林算法建立了 100 棵决策树模型，每一棵树带有四个随机特征，oob 错误率为 0.029，由于随机森林无法显示全部决策树，在此不以展示。

3.2　实验结果

由于在这些特征中包括数值特征与布尔型特征，总点击量和总推荐量等数值特征数值过大，会影响布尔型特征在模型中的权重比例，因此对数值较大的特征进行 MathExpression-E（A-MIN）/（MAX-MIN）数据预处理，将数值转化到 0 至 1 之间。通过贝叶斯网络、Logistic 回归、随机森林三种算法对抓取到的 5 649 部网络小说数据的热度预测结果显示如表 6 所示。

表 6　网络小说热度预测结果对比

算法	分类正确率/%	分类错误率/%	一致性检验	平均绝均差	均方根误差	相对误差绝对值	相对误差平方根
贝叶斯网络	82.970 4（4 687）	17.029 6（962）	0.726 7	0.098 5	0.252 2	32.117 3	64.421 0
逻辑回归	84.687 6（4 784）	15.312 4（865）	0.745 6	0.109 8	0.234 6	35.816 4	59.913 3
随机森林	97.079 1（5 484）	2.9209（165）	0.952 2	0.039 1	0.112 8	12.758 2	28.819 7

对比预测结果可以看出，随机森林算法在分类正确率及错误率上明显优于贝叶斯网络与逻辑回归，其中 Kappa 检验是评价一致性的测量值，其大小是用一个由渐进及标准误差构成的 t 统计量决定，当 Kappa＞0.75 表示好的一致性（Kappa 最大值为 1），随机森林算法的 Kappa 值达到 0.952 2，说明两次判断的一致性很好。在其余误差检验中，随机森林算法都有着较好的显示效果。

这样的预测结果虽然让人欣喜，但这个结果是否令人足够信服？本文从随机森林算法原理对预测结果进行分析判断，认为这样一个结果是科学可信的，依据有以下几点。

（1）随机森林算法通过自助法（bootstrap）重采样技术，使用决策树作为弱学习器。从节点上所有的 N 个样本特征中有放回地随机选择节点上的一部分样本特征，这个数字小于 N，假设为 N_{sub}，生成多个决策树组成随机森林。这种有放回的随机性选择样本的方法提高了模型的泛化能力，很好地降低了模型的方差。

（2）随机森林的模型输出采用投票法，对每一棵决策树的分类结果进行统计，得到最多票数的类别或类别之一作为最终模型输出，由于每一棵决策树的左右子树划分都是根据最优特征划分，因此投票法输出的结果更优，且在训练后，可以给出各个特征对输出的重要性。

（3）由于在建立每一棵决策树的过程中，训练样本的采集采用了有放回的随机性采集，保证了随机性的需求，因此就算没有进行剪枝，也不会出现过拟合情况。

综上所述，可以认为随机森林算法对网络小说的热度预测结果是科学有效的。随机森林算法对网络小说热度的预测结果相比 Logistic 回归和贝叶斯网络算法更优。

3.3　实验分析

根据实验结果，本文选择分类正确率达到 97.079 1%的随机森林算法作为网络小说热度预测及探寻网络小说在 CDN 中的分布方法。如表 7 所示，为随机森林预测参数，其中 TP Rate 是真正率，代表被预测模型预测为正的正样本，FP Rate 是假正率，代表被预测模型预测为正的负样本。分类器的分类效果越好，TP 值越高，FP 值越低。在四种

热度预测中，TP值均远大于FP值，分类效果较好。将系统检索到的相关文档数为A，系统检索到不相关文档为B，相关但系统没有检索到的文档为C，精度(Precision)$=A/(A+B)$，召回率(Recall)$=A/(A+C)$，F值(F-Measure)为精度与召回率的调和平均数。精度、召回率与F值是对分类器分类效果的度量值，值越大，代表结果质量越好，最高为1。

从表7中可以看出，三种衡量指标值均很高，说明采用随机森林算法对网络小说热度进行预测结果较好。ROC曲线又称为感受性曲线，它是一个综合性指标，用于反映连续变量的敏感性和特异性，通过构图显示敏感性与特异性的相互关系。曲线下面积越大，代表准确性越高，在表7中，四种热度预测的ROC曲线面积反映了建立模型的分类器的分类准确性较高。

表7 随机森林分类参数

等级	TP Rate	FP Rate	准确率	召回率	F-值	ROC曲线
1(冷)	0.993	0.013	0.989	0.993	0.991	1
2(一般)	0.968	0.017	0.959	0.968	0.964	0.998
3(热)	0.94	0.012	0.934	0.94	0.937	0.997
4(非常热)	0.752	0.001	0.954	0.752	0.841	0.998
平均权重	0.971	0.013	0.971	0.971	0.97	0.999

通过混淆矩阵可以更直观地看出随机森林分类器对5 649部网络小说的热度分类情况，对角线代表分类正确的样本，分类越集中在对角线，代表分类效果越好，具体的显示结果如表8所示。

表8 网络小说热度预测混淆矩阵

a(冷)	b(一般)	c(热)	d(非常热)	分级
2 969	20	0	0	a(冷)
34	1 567	17	0	b(一般)
0	47	824	6	c(热)
0	0	41	124	d(非常热)

预测模型对样本的预测与样本真实值匹配度高，分类效果好，具有很好的应用推广性。

4 结论

在CDN系统推(Push)策略中存在由于缺乏网络小说热度判断科学标准，主要依靠管理员的主观经验判断而存在预分发内容频繁替换的现象。因此造成内容边缘命中率低、用户启动延迟长、内容分发网络服务器负担重而严重影响服务质量的问题。基于上述问题，考虑到预分发内容的热度是直接决定边缘命中率和用户启动延迟的因素，同时考虑到流媒体服务器的缓存策略并不适用于网络小说服务器的前提下，将研究重点放在对网络小说热度的准确预测和判断的问题上，为预分发内容的策略部署提供有效的参考和指导。通过建立网络小说热度评价标准，采用分类算法对网络小说进行热度预测。对比贝叶斯网络、Logistic回归与随机森林三种分类方法对网络小说热度预测结果，显示随机森林算法预测准确率达到97.079%，有很好的预测效果，更适用于CDN系统对网络小说热度的预测。结果表明，表征网络小说热度的变量选择有很好的解释性，借助本文建立的网络小说热度评价标准，采用随机森林算法对网络小说进行热度预测，可以为管理员对预分发内容热度的判断提供科学有效的判断依据，提前预判并及时调整最优部署策略，提高CDN系统服务质量与运作效率。

参考文献

[1] Saroiu S, Gummadi K P, Dunn R J, et al. An analysis of internet content delivery systems [C]//Proceedings of the 5th Symposium on Operating Systems Design and Implementation. NewYork: ACM, 2002, 36: 315-327.

[2] A Charya S, Smith B, Pames P, Characterizing user access to Videos on World Wide Web[C]//Proceedings of the Multimedia Computing and Networking 2000, SPIE 2000.

[3] M Yang, Z Fei. A model for replica placement in content distribution networks for multimedia applications [J]. IEEE International Conference on Communications, 2003, 1(1): 557-561.

[4] Tim Wauters, Jan Coppens, et al. Replica placement in

ring based content delivery networks[J]. Computer Communications,2006,29(16): 3313-3316.
[5] 黄灿,杨鹏,顾梁. 播存网络中一种融合信任机制的协同过滤推荐算法[J]. 小型微型计算机系统,2016,37(11): 2504-2508.
[6] 何腾蛟. 基于 CDN 的视频流媒体内容分发策略的研究[D]. 深圳: 深圳大学硕士学位论文,2017.
[7] 王洁,汤小春. 基于社区网络内容的个性化推荐算法研究[J]. 计算机应用研究,2011,28(04): 1248-1250.
[8] 孙立奋,潘达儒. 一种基于兴趣挖掘的机会网络内容分发策略[J]. 华南师范大学学报(自然科学版),2017,49(05): 108-114.
[9] 覃梦河,晋佑顺,邱远棋. 基于内容分析的微博用户关系推荐机制研究[J]. 图书馆论坛,2013,33(04): 104-108.
[10] 王刚,蒲国林,邱玉辉. 一个基于社会网络的内容推荐模型研究[J]. 计算机应用与软件,2012,29(12): 47-50.
[11] 芮兰兰,彭昊,黄豪球,等. 基于内容流行度和节点中心度匹配的信息中心网络缓存策略[J]. 电子与信息学报,2016,38(02): 325-331.
[12] 熊庆昌. 媒体内容分发网络的内容部署策略及性能分析[D]. 合肥: 中国科学技术大学硕士论文,2009.
[13] 杨传栋,余镇危,王行刚,等. 一种流媒体 CDN 的内容部分推送策略[J]. 计算机工程与应用,2007(25): 162-164,185.
[14] 杨磊. 内容网络中内容调度技术研究[D]. 重庆: 重庆大学博士学位论文,2015.
[15] 王晓耘,袁媛,史玲玲. 基于微博的电影首映周票房预测建模[J]. 现代图书情报技术,2016(04): 31-39.
[16] 王炼,贾建民. 基于网络搜索的票房预测模型——来自中国电影市场的证据[J]. 系统工程理论与实践,2014(12): 3079-3090.
[17] W Fu, N Xiao, X Lu. A quantitative survey on qos-aware replica placement[C]//Grid and Cooperative Computing, 2008. GCC'08. Seventh International Conference on. IEEE, Shenzhen, 2008,281-286.
[18] 徐晓枫,等. 融合社交与搜索数据的电视剧点播排名预测方法[J]. 计算机工程,2015,41(8): 6-12,17.
[19] 王铮,许敏. 电影票房的影响因素分析——基于 Logit 模型的研究[J]. 经济问题探索,2013(11): 96-102. [2017-08-29].
[20] M Hefeeda, O Saleh. Traffic modeling and proportional partial caching for peer-to-peer systems[J]. IEEE/ACM Transactions on Networking, 2008, 16(6): 1447-1460.
[21] Markov Z, Russell I. An introduction to the WEKA data mining system[C]//Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education. New York: ACM, 2006(38): 367-368.
[22] Friedman N, Geiger D, Goldszmidt M. Bayesian nerworks classifiers[J]. Machine Learning, 1997, 29(2/3): 131-163.
[23] Randy A Nelson, Robert Glotfelty. Movie Stars and box office revenues: An empirical analysis[J]. Journal of Cultural Economics, 2012, 36(2): 141-166.
[24] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

赵礼强(1975－),博士,教授,主要研究领域为电子商务环境下供应链优化。
E-mail: zhao_liqiang@163.com

靖可(1981－),博士,副教授,主要研究领域为应急物流与应急管理。
E-mail: chloe.jingke@gmail.com

姜崇(1992－),通信作者,硕士研究生,主要研究领域为数据挖掘、数据分析。
E-mail: m15840316785@163.com