

融合网络异构 P2P 流媒体系统抗扰动性研究*

苏杭^{1,2}, 王劲林², 尤佳莉²

(1. 中国科学院研究生院, 北京 100049; 2. 中国科学院声学研究所 国家网络新媒体工程技术研究中心, 北京 100190)

摘要: 融合网络中的异构 P2P 流媒体系统中具有多种终端、应用及节目类型, 使得用户行为模式复杂, 为系统稳定性带来了挑战。针对以上问题提出融合网络异构 P2P 流媒体系统稳定性增强策略。通过分析流媒体系统用户日志, 确定用户行为影响因素, 在此基础上对不同类型的用户行为进行收集及拟合, 进行基于可靠性分析的路由、资源信息维护及合作节点选择, 提高了系统资源信息维护准确性及节点间数据合作有效性。使用真实系统日志驱动的仿真实验验证了该策略的效果。

关键词: P2P 流媒体; 异构网络; 用户行为; 扰动; 抗扰动

中图分类号: TP37 **文献标志码:** A **文章编号:** 1001-3695(2012)11-4220-04

doi:10.3969/j.issn.1001-3695.2012.11.055

Anti-churn research of integrated heterogeneous P2P streaming system

SU Hang^{1,2}, WANG Jin-lin², YOU Jia-li²

(1. Graduate University of Chinese Academy of Sciences, Beijing 100049, China; 2. National Network New Media Engineering Research Center, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Heterogeneous P2P streaming system in integrated networks contains multiple types of terminals, applications and program contents, which brings in complex user behavior patterns and make challenge to system stability. After extracting user behavior affect factors through analysis of streaming system user trace, this paper proposed an anti-churn scheme for heterogeneous P2P streaming (ACSHP) to solve these problems. ACSHP collected and fitted heterogeneous user behavior data to improve maintain strategy of routing table, published resource information and cooperating peer selection, which improved the accuracy of resource information maintaining and the effectiveness of data cooperation between peers. Real system trace-driven simulation verifies the effectiveness of the strategy.

Key words: P2P streaming; heterogeneous network; user behavior; churn; anti-churn

0 引言

与 P2P 文件共享系统相比, P2P 流媒体系统中的节点会更加频繁地加入、离开系统甚至直接失效退出, 即 churn。为保证视频播放的流畅度, P2P 流媒体系统对数据传输带宽、实时性等方面有更高的要求。因此需要设计良好的抗扰动机制, 令 P2P 流媒体系统能够在更加剧烈的扰动下仍然保持相对稳定的数据传输与播放性能。

现有 churn 研究主要基于同构 churn 模型, 所有节点遵循相同的固定 churn 规律。通过分析流媒体系统日志发现, 流媒体系统中的节点 churn 行为不完全一致且具有时变性。基于本文对异构 churn 环境下的 P2P 流媒体系统抗扰动性进行了优化。首先动态感知系统中节点的 churn 模式, 通过可靠性理论分析路由维护、信息发布及搜索过程中的节点及信息的可用性, 依据当前 churn 参数来计算路由维护周期等维护参数, 保证系统可用性。在选择候选合作节点时, 基于节点的可用性程度及其网络性能, 选择质量更高的节点。

本文描述了 P2P 系统 churn 相关概念及研究现状, 分析三网融合环境下异构 P2P 流媒体系统的 churn 模式, 提出了融合网络异构 P2P 流媒体系统稳定性增强策略 ACSHP (anti-churn scheme for heterogeneous P2P streaming), 描述仿真实验并分析结果。

1 相关研究

研究 churn 首先要对相关系统进行度量, 为 churn 模式进行建模。Churn 的度量主要从节点在线时长、离线时长、会话时长、剩余会话时长等方面进行^[1]。Vu 等人^[2]对 PPLive 的测量结果表明: 与 P2P 文件共享系统相比, P2P 流媒体中的节点更加“不耐烦”, 且节点上下线波动程度更大。对 IPTV 系统的测量分析^[3,4]也证明了 IPTV 系统中的节点在线时间较短, 且用户行为有着较强的时间相关性。因此, P2P 流媒体系统需要具有更好的抗扰动性能, 以便在剧烈的 churn 下保证流畅的视频数据交互及播放。

当前存在不同 churn 模型来描述 P2P 系统中的节点行为。

收稿日期: 2012-03-17; **修回日期:** 2012-04-22 **基金项目:** 国家“863”计划基金资助项目(2011AA01A102); 中国科学院战略性先导科技专项子课题资助项目(XD06010301); 国家自然科学基金资助项目(60903218)

作者简介: 苏杭(1985-), 男, 河北秦皇岛人, 博士, 主要研究方向为 P2P 流媒体系统、基于覆盖网的媒体数据传输优化(suheaven@gmail.com); 王劲林, 男, 研究员, 博导, 主要研究方向为数字音/视频及网络多媒体技术、融合业务体系; 尤佳莉, 女, 副研究员, 主要研究方向为分布式系统、虚拟化技术。

P2P 文件共享系统中主要使用韦伯、幂律、帕累托^[1,5]等重尾分布来描述节点在线时长。而 P2P 流媒体系统中的节点在线时长特性较为复杂,既具有负指数分布、混合负指数分布,也存在帕累托、韦伯^[3,4,6,7]等重尾分布的描述。现有研究多使用同构 churn 模型,即所有节点服从相同的静态的 churn 模式。文献[4]虽然考虑了 churn 参数的时变性,但仍无法描述不同节点遵循不同 churn 规律的系统。

为增强 P2P 流媒体系统的抗扰动性,可以从数据调度层及覆盖网层两方面着手。在数据调度方面,数据分发结构从最早的单树型发展到多树型直至当前广泛应用的网状分发结构^[7],通过改变数据分发传输过程中父子节点间的依赖关系,健壮性依次增强。而在覆盖网层,可通过在路由维护、信息发布及搜索、数据副本分布等方面增加冗余度^[8],来提高系统在 churn 下的健壮性,但过高的冗余度同时会带来额外的维护开销。文献[8,9]提出根据系统 churn 程度动态调整路由表规模来提高覆盖网的可靠性。另外可通过在选择合作节点时考虑节点的稳定性因素^[6,10],进而降低交互过程中的 churn 发生概率,提高数据交互的质量。但现有方法没有考虑节点 churn 的异构性,基于同构 churn 模型来进行路由、搜索、节点选择等方面的稳定性计算与应用,限制了稳定性的提高。

2 融合网络异构 P2P 流媒体系统扰动分析

2.1 融合网络异构 P2P 流媒体系统结构分析

融合网络环境下的大规模异构 P2P 流媒体系统具有多种应用平台,如 PC 机客户端、使用电视机播放的机顶盒以及智能手机、平板电脑等移动终端。用户网络具有 ADSL、FTTX + LAN、WLAN、3G 等多种接入方式。系统中的节目可采用分层编码,进而使用不同网络及终端的节点可根据自身屏幕分辨率、网络接入等条件获取相应编码层,并保持不同类型节点间的数据互通。不同类型的终端具有不同的计算、网络传输能力,以及如计算机显示器上的前倾式、电视屏幕上的后仰式收看的不同用户体验。对不同类型流媒体系统的测量^[1,4-6,10,11]表明:在面向电视机的 IPTV 系统中,用户会话时长主要为指数型分布,而计算机上的 IPTV 用户会话时长多服从重尾分布。因此同构系统 churn 模型已不适用于此类系统,需要针对不同类型的用户分别考虑其在线时长分布以及节点行为的时变性,在此基础上进行系统维护及优化。

结论 1 流媒体系统中不同类型的节点具有异构的 churn 模式,如图 1 所示。

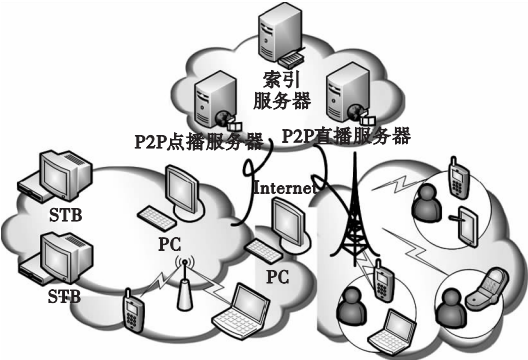


图1 异构融合P2P IPTV系统结构

2.2 融合网络异构用户行为分析

为了解流媒体系统 churn 模式,对上海嘉定有线 2010. 6.

15 ~ 2010. 6. 21 的用户日志进行分析,共涉及 94 个节目、1 481 个用户的 157 422 条收看记录。用户收看节目时存在快速浏览并切换频道的行为,频繁地快速切换频道会为 P2P 流媒体系统的信息维护及数据交互带来巨大扰动。将日志中 ≤ 30 s 的会话判定为快速节目浏览,进而定义平均快速浏览速率为:一个统计周期内的快速浏览总次数与该周期内在线节点数之比,用于描述快速浏览的对应的 churn 强度。以 10 min 为周期统计在线节点数、节点平均会话时长及平均快速浏览速率,其结果如图 2 所示。

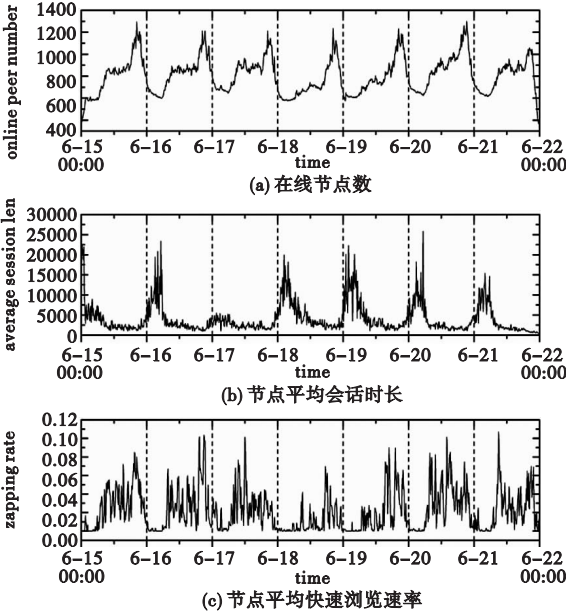


图2 2010.6.15—2010.6.21用户行为统计

图 2 中,各被测参数随时间有明显的规律性变化。在线节点数在夜晚黄金时间大幅上升,节点平均会话时长的变化规律恰与此相反。图 2(c)表明在线节点数量大时节点的频道切换强度也在增大。分析可知:在一天中的黄金时间,用户收看行为活跃,在线节点数上升,活跃的用户同时会进行频繁的切换频道等操作,进而拉低节点平均会话时间。而在后半夜等冷门时间,用户活跃度下降,切换频道频度与在线节点数同时降低,而会话时间相应增长。

为了定量分析各参量的关系,计算以上三被测量序列之间的 Spearman 相关系数,其结果如表 1 所示。在线节点数明显正相关于节点平均快速浏览速率,而与平均会话时长之间具有强度相当的反相关,与对图 2 的判断相吻合。

表 1 在线节点数、会话时长、快速浏览速率间相关系数

相关系数	在线节点	会话时长	浏览速率
在线节点	1	-0.76	0.76
会话时长	-0.76	1	-0.72
浏览速率	0.76	-0.72	1

结论 2 流媒体系统中的在线用户数、用户在线时长、churn 程度具有明显的时变性且相互关联。

分析前述用户日志中直播与点播节目节点整体会话时长分布,以及体育、新闻、影视类直播节目各自的用户会话时长 CDF 分布得到图 3。整体上点播节目的用户会话时长短于直播节目节点。而在直播节目内部,体育类、新闻类、影视类节目的会话时长依次变短。可见收看不同类型节目的节点的行为具有显著异构性。若对收看不同类型节目的节点采用同种覆盖网维护策略,难达到最优效果。

结论 3 流媒体系统中不同类型节目间的节点行为具有异构性。

3 异构 P2P 流媒体系统稳定性增强方法

P2P 流媒体节点结构的主要组成层次为:进行媒体数据传输合作的数据调度层,以及维护系统中节点与资源信息、为数据调度层提供候选合作节点的覆盖网层。本章中,覆盖网采用 Kademlia 协议进行组织,各节点将自身所拥有的媒体数据资源信息以〈key, value〉对的形式发布到 Kad 网络中。复杂的 churn 环境对覆盖网层的路由、信息维护检索性能都会造成影响,而覆盖网为数据调度层提供的候选合作节点的潜在稳定性将影响数据调度层的数据合作流畅度,影响播放性能。

为了应对 P2P 流媒体系统中异构及时变的复杂 churn 情况,提出异构 churn 环境下的抗扰动解决方法:该方法抛弃同构 churn 模型,首先对系统异构 churn 状况进行感知,获取当前 churn 参数。之后通过 churn 参数来计算节点、信息的可靠性,在覆盖网维护方面,通过 churn 参数计算得到的可靠性程度优化路由维护、信息发布与搜索的性能。最后在覆盖网为数据调度层提供候选节点时,使用基于可靠性分析的候选合作节点选择方法,提高合作节点稳定性,进而提高系统整体可靠性。

3.1 异构 churn 感知

由前述分析可知,节点 churn 模式与节点的终端类型、所收看节目类型及收看行为所处的时间有关,所以用终端类型与收看节目类型〈终端类型, 节目类型〉联合标志 churn 类型,并采取实时采集节点状态的方式来更新各类型 churn 的参数,捕获 churn 的时变动态性。每个节点在上线时,判别自己的终端类型并记录上线时间,通过心跳报文与捎带检查结合的方式,采样与自己类型相同、且收看同一类型节目的其他部分节点的存活状况。节点在回复存活探测消息时,告知自己的上线时间。节点间通过 Gossip 来传递各自的采样结果,令各节点都能够掌握自身类型节点的整体 churn 情况。根据以上采样情况,节点即可拟合出自身类型节点群的 churn 参数,并定期进行更新。对节点在线时长分布分别根据负指数分布、帕累托分布及韦伯分布进行极大似然估计,求得分布参数以及对数似然度,针对似然度最大的分布类型采用 Kolmogorov-Smirnov 法进行显著性水平为 0.05 的分布假设检验。若检验通过则使用拟合分布及其参数进行后续计算,否则直接根据在线时长值得经验累积分布,用于后续可靠度的数值计算。

节点 i 将节点 j 添加进自己的路由表时,记录 j 的终端类型和收看节目类型。此后采用捎带更新的方式,通过存活探测等消息来更新各节点的 churn 模型参数,当 i 接收到来自某节点的〈终端类型 u , 节目类型 v 〉类型的最新 churn 参数后,采用指数加权移动平均(exponential weighted moving average, EWMA)的方式,更新路由表中所有此类型节点的 churn 参数,来实现对时变 churn 参数的追踪。

3.2 基于可靠性分析的路由维护

DHT 通过定期检查路由表中节点的存活情况来保证路由的可靠性。Kad 采用并行度为 α 的迭代式路由方法,因此将路由过程视为 J 步的串行搜索过程,而其中每一步又由 α 个并行过程组成。设 $R^i(t) = P[T > t] = 1 - F(t)$ 为路由表中节点 i 的可靠性,则从路由表中任选 α 个节点的并行可靠性为

$$R_\alpha(t) = 1 - \prod_{i=1}^{\alpha} (1 - R^i(t)) \quad (1)$$

进而 J 步串行搜索的总可靠性为 $R_\alpha(t)^J$ 。设搜索可用性阈值为 θ_α ,对于路由表中的节点 i ,若 i 的加入时间为 t_0 ,最近判定存活时间为 t_1 ,则其下一次 ping 时间 t_2 应满足:

$$R^i(t_2 - t_0 | t_1 - t_0) \geq 1 - \sqrt[\alpha]{1 - \sqrt[\alpha]{\theta_\alpha}} \quad (2)$$

进而得到对节点 i 的存活检查时间

$$\tau^i(\theta_\alpha) = \sup\{t > 0; R^i(t - t_0 | t_1 - t_0) \geq 1 - \sqrt[\alpha]{1 - \sqrt[\alpha]{\theta_\alpha}}\} \quad (3)$$

在节点数为 n 的 Kad 网络中,最坏情况下的搜索跳数 $J_{\max} = \log_2 n$ 。研究表明 DHT 中的平均搜索跳数远低于最大搜索跳数^[12],在原始 Kad 协议中,平均路由跳数为 $E[J] = 1 + \frac{\log_2 n - 7.73}{6.98}$,改进的 Kad 协议平均路由跳数将进一步缩小。

由式(2)可知, $R^i(t)$ 是 J 的增函数,过高的 J 值将导致每一步 $R^i(t)$ 值过高,超出了平均路由可靠性需求,并导致维护周期缩短。因此在计算检查周期时以 $E[J]$ 替代式(2)中的 J ,来降低检查开销。

进而对于一个 K 桶(K-Bucket)内的 K 个节点,设置该 K 桶的存活检查周期为

$$\tau_K(\theta_\alpha) = \min\{\tau^i(\theta_\alpha)\}, \text{Peer}_i \in \text{Bucket}_K \quad (4)$$

传统 Kad 协议假定网络中节点行为大多服从重尾分布, $R_i(t) > R(t)$,已在线节点稳定性高于新加入的节点,因使用懒惰的路由更新策略。而异构 P2P 流媒体中节点行为的异构特性将导致以上策略的基础失效。因此使用基于节点稳定性的路由更新策略:发现新节点时,计算新节点的剩余在线时长 $R^{\text{new}}(t)$ 与其对应的路由表项中可靠性最低的节点的 $R^{\text{low}}(t)$ 值,将该值较高的节点放入路由表。

3.3 基于可靠性分析的信息发布与搜索

在 DHT 网络中,节点下线后将导致该节点所维护的其他节点发布到本节点上的资源信息丢失,流媒体系统中节点高动态性将加剧该信息丢失现象。一个节点所发布信息丢失将导致系统中其他节点无法搜索到该节点的资源,进而导致该节点资源无法被有效利用,系统带宽利用率下降。在 ACSHP 中设置基于信息维护节点可靠度的信息可检索性检查:信息发布者存储维护着自身所发布的资源信息的节点的〈终端类型, 节目类型〉及加入、存活信息,据此计算信息维护节点的可靠度,设置信息可检索性检查周期。当发现信息副本数少于一定阈值时,进行资源信息的重新发布。对于资源信息副本数为 β 的系统,可靠度 θ_β 下的资源信息可检索性的检查周期为

$$\tau^i(\theta_\beta) = \sup\{t > 0; R^i(t - t_0 | t_1 - t_0) \geq 1 - \sqrt[\beta]{1 - \theta_\beta}\} \quad (5)$$

系统中节点的频繁下线、切换节目的行为将导致该节点在覆盖网中发布的自身拥有的视频数据资源信息快速失效,令覆盖网层为数据调度层提供候选合作节点时的有效信息比例下降,降低数据合作成功概率。因此进行信息可靠性检查。与式(4)同样形式,设置可靠性阈值 θ_j 来计算检查周期。

3.4 基于可靠性分析的候选合作节点选择

数据调度层选择合作节点的目标位进行有效的数据传输,该有效性体现在一定带宽条件下的可靠传输时间。在候选节点已经拥有所需数据的情况下,仅依据节点带宽或剩余在线时长因素选择合作节点均忽略了数据传输的另一面。因此在 ACSHP 中,根据节点的带宽、可靠性分布情况,计算对方节点为本节点进行数据传输的能力,以此作为候选合作节点选择标准。

设节点 j 可为 i 提供的传输带宽为 B_j ,则在未来 t 时长内 j

对 i 而言的可用性价值 $U_i^j(t)$ 为: $U_i^j(t) = B_{ij} \int_0^t v R^i(v) R^j(v) dv$ 。当从本节点视角选择其他各节点时, $R^i(v)$ 所造成的影响一致, 可以省略。而当本节点自身的在线概率低于一定程度时, 继续考虑其余节点的可靠性意义不大, 因此节点 i 设置自身可靠性阈值下限 θ_{self} , 使用式(7)表征节点 j 对 i 的可用性价值, 并在选择候选合作节点时, 优先选择 U_i^j 值更大的节点。

$$t_\theta = \sup \{t > 0: R(t) \geq \theta_{self}\} \quad (6)$$

$$U_i^j(t) = B_{ij} \int_0^{t_\theta} v R^i(v) dv \quad (7)$$

4 仿真实验及分析

4.1 实验设置

本文基于 P2P Strmsim^[13] 仿真器实现了异构融合流媒体系统采用的覆盖网层, 加入多直播、点播节目的支持和覆盖网的构建模块。仿真实验设置直播频道 11 个, 分为体育、新闻、影视三大类, 设置点播影片 10 个。共设置节点 1 241 个, 节点终端类型分为机顶盒与 PC 两种, 采用上海嘉定有线系统日志驱动直播节目的节点行为, PC 终端节点采用文献[14]所述日志驱动, 与嘉定有线日志的点播部分共同驱动点播节目节点行为(随机各占 50%)。

节点上下行带宽设置如表 2 所示。各组实验采用数据调度算法为随机调度, VRNS 算法^[15] 通过选择带宽满足一定条件的节点, 对 Gridmedia 的合作节点选择方式进行了优化。本文与以上两种流媒体系统的节点选择方式进行对比, 路由及信息维护性能维护方式与 Kademlia 及 A-Kademlia^[9] 作对比分析。实验开始为 4 h 的预热启动阶段, 之后为 24 h 的测量仿真阶段。

表 2 仿真实验中节点带宽参数设置

类型	上行带宽	下行带宽	节点比例
1	1 Mbps	3 Mbps	0.1
2	384 kbps	1.5 Mbps	0.4
3	128 kbps	768 kbps	0.5

4.2 实验结果及分析

图 4 统计了覆盖网所维护的其他节点发布的资源信息的准确性。在原始 Kademlia 及 A-Kademlia 中, 由于没有针对信息的有效性进行维护, 流媒体系统中节点的快速离线、切换节目行为导致节点所发布的自身资源信息快速失效, ACSHP 的信息维护策略有效提高了信息的准确性, 且抑制了节点动态性的时变特征带来的影响。

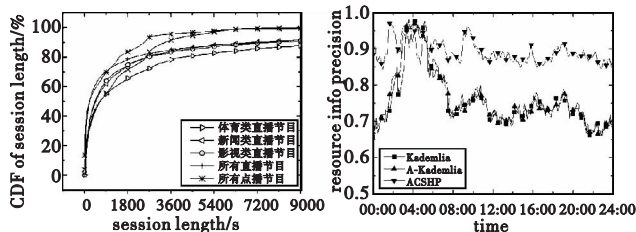


图 3 不同类型节目中节点在线时长分布对比

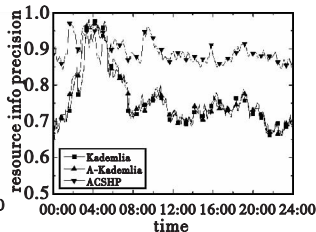


图 4 覆盖网所维护资源信息准确性

实验统计了仿真实验中数据调度层向覆盖网获取候选合作节点的频率分布。如图 5 所示, 由于 Gridmedia 采用随机节点选择方法, 其合作节点的稳定性及数据传输性能均难以保证, 因此其候选合作节点搜索频率最高; VRNS 一定程度上改善了合作节点的带宽条件, 但与 Gridmedia 类似, 由于底层覆盖网提供的拥有视频资源的节点信息的准确性波动大, 调度层从

覆盖网获取了较多数量的无效节点, 因此需要进行较多次的候选合作节点搜索与鉴别。ACSHP 一方面提高了覆盖网层为数据调度层提供信息的准确性, 另一方面对候选合作节点的数据传输能力进行了优化, 从两方面共同降低了候选合作节点搜索速率。

定义节点的播放质量为该节点获取有效视频数据流的速率与视频码率的比值, 则播放质量值越接近于 1 表明节点播放越流畅。节点加入系统后先进行缓冲, 在节点加入 30 s 后开始统计其播放质量。实验中以 10 min 为周期统计当前在线节点的平均播放质量。如图 6 所示, Gridmedia 及 VRNS 策略下的播放质量波动较大, 对比图 1 可知, 在系统中节点动态性高时, Gridmedia 及 VRNS 下的播放质量降低。由于 Gridmedia 随机选择合作节点, 在节点动态性高时受影响较大, VRNS 对合作节点的带宽条件进行了优化选择, 其播放质量优于 Gridmedia, 但由于没有考察节点的在线时长特性, 对高动态环境适应性也较差。与此相比, ACSHP 策略选择合作节点时综合了节点的带宽及在线稳定性因素, 以成功传输数据的能力为选择依据, 因此其对动态环境适应性较强。

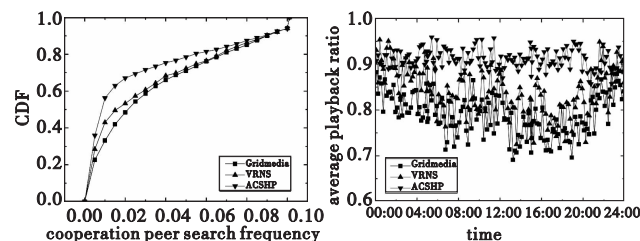


图 5 候选合作节点搜索速率累积概率分布

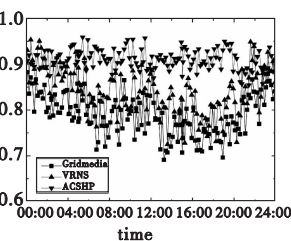


图 6 在线节点平均播放质量

5 结束语

通过对真实流媒体系统用户日志及相关文献进行分析, 三网融合环境下的 P2P 流媒体系统具有与时间、节目类型、终端类型相关的异构、时变的用户行为, 为系统带来更加复杂的扰动性。为了应对此扰动性, 提出 ACSHP 策略。在对异构扰动性进行感知的基础上, 在进行路由维护、覆盖网资源信息维护及合作节点选择时, 根据节点的扰动模式, 分析相关信息的可靠性, 并据此决定维护及选择目标。真实系统驱动的仿真实验证明了 ACSHP 策略的有效性。本文仅分析了广电网与互联网中流媒体系统数据, 没有涉及移动终端用户行为及网络、设备的数据分析, 这一方向值得继续深入研究。

参考文献:

- [1] 张宇翔, 杨冬, 张宏科. P2P 网络中 Churn 问题研究[J]. 软件学报, 2009, 20(5): 1362-1376.
- [2] VU L, GUPTA I, NAHRSTEDT K, et al. Understanding overlay characteristics of a large-scale peer-to-peer IPTV system [J]. ACM Trans on Multimedia Computing, Communications, and Applications, 2010, 6(4): 1-24.
- [3] GENG Y, WESTHOLM T, KIHLM M, et al. Analysis and characterization of IPTV user behavior [C]//Proc of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting. 2009: 13-15.
- [4] QIU Tong-qing, GE Zi-hui, LEE S, et al. Modeling user activities in a large IPTV system [C]//Proc of the 9th ACM SIGCOMM Conference on Internet Measurement Conference. Chicago: ACM, 2009: 430-441.

如表 3 所示,是使用离散化的朴素贝叶斯分类器,抽取前 10 个关键词时的效果。其中 Bayes 代表离散化的朴素贝叶斯, Trees 代表装袋决策树,其他同表 2。朴素贝叶斯假设各特征间相互独立,而使用朴素贝叶斯分类器时,主题特征与常用特征结合不能取得更好效果。逆向推测可知,主题特征与基本特征间应有较紧密的关系。其实,本文使用的几个特征之间都应有较大关系,在很多情况下,关键词在具有较高 TF-IDF 值的同时,也会出现在文章开头/结尾,其节点度也可能较高,同时还可能有较高的主题特征值。

表 3 使用不同分类器的对比效果

分类器	特征集	P/%	R/%	F/%
Bayes	Base	16.00	21.97	18.51
	Base + TF1	15.90	21.52	18.29
Trees	Base	17.90	26.16	21.26
	Base + TF1	19.60	28.17	23.12

5 结束语

本文在使用常用特征构建一个关键词抽取方案的基础上,基于 LDA 模型中的主题下词的分布,提出一种主题特征计算方法,并将其应用到关键词抽取中。实验表明,新的特征给系统的性能带来了显著提高,但性能稳定性还有很大改进空间。

主题模型为计算词的主题相关特征提供了很多方便,在今后的工作中,还可以尝试在 LDA 主题模型的基础上寻找更加稳定有效的主题特征计算方法。虽然本文的主题特征是语言无关的,但主题特征的表现形式可能有差异。所以,在中文语料上验证主题特征的有效性,也是今后的工作之一。也可以考虑将主题特征应用到自然语言处理的其他方面。

参考文献:

- [1] TURNEY P D. Learning to extract key phrases from text, NRC Technical Report ERB-1057 [R]. Canada: National Research Council, 1999.
- [2] WITTEN I H, PAYNTER G W, FRANK E, *et al.* Kea: practical automatic keyphrase extraction [C]//Proc of ACM Conference on Digital Libraries. New York: ACM Press, 1999: 254-255.
- [3] MEDELYAN O, FRANK E, WITTEN I H. Human-competitive tagging using automatic keyphrase extraction [C]//Proc of Conference on Empirical Methods in Natural Language Processing, 2009: 1318-1327.
- [4] MEDELYAN O, WITTEN I H, MILNE D. Topic indexing with Wikipedia [C]//Proc of Wikipedia and AI workshop at the AAAI-08 Conference (WikiAI08). 2008.
- [5] PASQUIER C. Task 5: single document keyphrase extraction using sentence clustering and latent dirichlet allocation [C]//Proc of ACL Workshop on Semantic Evaluation, 2010: 154-157.
- [6] DAVID M B, ANDREW Y N, MICHAEL I J. Latent dirichlet allocation [J]. *Journal of Machine Learning Research*, 2003, 3: 993-1022.
- [7] 耿焕同, 蔡庆生, 于琨, 等. 一种基于词共现图的文档主题词自动抽取方法 [J]. *南京大学学报: 自然科学版*, 2006, 42 (2): 156-162.
- [8] 蒋昌金, 彭宏, 陈建超, 等. 基于组合词和同义词集的关键词提取算法 [J]. *计算机应用研究*, 2010, 27 (8): 2853-2856.
- [9] KIM S N, MEDELYAN O, KAN Min-yen, *et al.* SemEval-2010 task 5: automatic keyphrase extraction from scientific articles [C]//Proc of ACL Workshop on Semantic Evaluation, 2010: 21-26.
- [10] STEYVERS M, GRIFFITHS T. Probabilistic topic models [M]//Latent semantic analysis: a road to meaning. (In Press).
- [11] McCALLUN A K. MALLET: a machine learning for language toolkit [EB/OL]. (2002). <http://mallet.cs.umass.edu>.
- [12] PORTER M F. An algorithm for suffix stripping [J]. *Program*, 1980, 14 (3): 130-137.
- [13] MEDELYAN O, WITTEN I H. Domain-independent automatic keyphrase indexing with small training sets [J]. *Journal of the American Society for Information Science and Technology*, 2008, 59 (7): 1026-1040.
- [14] HULTH A. Combining machine learning and natural language processing for automatic keyword extraction [D]. Swedish: Stockholm University, 2004.
- [15] MEDELYAN O. Human-competitive automatic topic indexing [D]. New Zealand: University of Waikato, 2009.
- [5] STEINER M, EN-NAJJARY T, BIRSACK E W. Long term study of peer behavior in the KAD DHT [J]. *IEEE/ACM Trans on Networking*, 2009, 17 (5): 1371-1384.
- [6] VASSILAKIS C, STAVRAKAKIS I. Minimizing node churn in peer-to-peer streaming [J]. *Computer Communications*, 2010, 33 (14): 1598-1614.
- [7] LIU Yong, GUO Yang, LIANG Chao. A survey on peer-to-peer video streaming systems [J]. *Peer-to-Peer Networking and Applications*, 2008, 1 (1): 18-28.
- [8] JINYANG L, STRIBLING J, MORRIS R, *et al.* A performance vs. cost framework for evaluating DHT design tradeoffs under churn [C]//Proc of the 24th IEEE International Conference on Computer Communications. [S. l.]: IEEE, 2005: 225-236.
- [9] 徐强, 孙乐昌, 张旻, 等. Kademlia 协议中的路由表自适应调节机制 [J]. *应用科学学报*, 2011, 9 (1): 66-72.
- [10] WANG Feng, LIU Jiang-chuan, XIONG yong-qiang. Stable peers: existence, importance, and application in peer-to-peer live video streaming [C]//Proc of the 27th IEEE Conference on Computer Communications. [S. l.]: IEEE, 2008: 1364-1372.
- [11] WANG Xiao-ming, YAO Zhong-mei, DOGUINOV L. Residual-based estimation of peer and link lifetimes in P2P networks [J]. *IEEE/ACM Trans on Networking*, 2009, 17 (3): 726-739.
- [12] STUTZBACH D, REJAIE R. Improving lookup performance over a widely-deployed DHT [C]//Proc of the 25th IEEE International Conference on Computer Communications. [S. l.]: IEEE, 2006: 1-12.
- [13] ZHANG Meng, ZHANG Qiao, SUN Li-feng, *et al.* Understanding the power of pull-based streaming protocol: can we do better? [J]. *IEEE Journal on Selected Areas in Communications*, 2007, 25 (9): 1678-1694.
- [14] YU Hong-liang, ZHENG Dong-dong, ZHAO B Y, *et al.* Understanding user behavior in large-scale video-on-demand systems [J]. *SIGOPS Oper Syst Rev*, 2006, 40 (4): 333-344.
- [15] GHOSHA J, MIAO W, LISONG X, *et al.* Variable neighbor selection in live peer-to-peer multimedia streaming networks [C]//Proc of the 5th International Conference on Broadband Communications, Networks and Systems. New York: ACM, 2008: 8-11.

(上接第 4223 页)