

混合流媒体分发系统中多指标用户群分组算法

万明刚,李泽平,张 军

(贵州大学 计算机科学与技术学院,贵州 贵阳 550025)

摘要:针对混合流媒体分发系统中服务节点提供服务时出现额外的跨地域、跨网络开销以及搜索服务节点效率低下的问题,提出一种基于多指标的用户群分组算法,并应用于服务节点选择。依次分别利用节点位置和网络类型对用户群进行分组,将候选服务节点限制在同地域、同ISP范围内,以减少不必要的跨地域、跨网络开销;然后结合节点兴趣对用户群作进一步划分,将搜索服务节点的范围限制在兴趣组内,以减小搜索流量、提高搜索效率。仿真实验表明:提出的算法能有效将用户群分组,提高分发系统服务效率。

关键词:混合流媒体分发系统;节点位置;网络类型;节点兴趣;分组;服务节点选择

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2016)10-0036-05

doi:10.3969/j.issn.1673-629X.2016.10.008

A User-group Grouping Algorithm Based on Multiple Indicators in Hybrid Streaming System

WAN Ming-gang, LI Ze-ping, ZHANG Jun

(College of Computer Science & Technology, Guizhou University, Guiyang 550025, China)

Abstract: To address the problem of additional cross-regional and cross-network cost that appears when the service node provides streaming media services, and the problem of inefficiency in searching service node, a grouping algorithm based on multiple indicators is proposed and used in server selection. As the algorithm goes, the user-group would be grouped by the node location and network type successively, limited the candidate server nodes in the same area and the same ISP, consequently reduced additional cross-regional and cross-network cost. Then the user-group would be divided further based on nodes' interest, limited the search scope within an interest group, consequently reduced searching traffic and improved searching efficiency. Simulation demonstrates that the proposed algorithm can effectively divide the user-group, improving service efficiency.

Key words: hybrid streaming system; node location; Internet type; node interest; grouping; server node selection

1 概述

在CDN-P2P混合流媒体分发系统中,用户节点可以通过两种方式获取流媒体资源:一是由边缘服务器提供,二是由peer节点之间共享。在以peer节点之间共享的方式获取资源时,请求节点搜索拥有所请求的流媒体资源的peer节点,并从中选择一个或多个节点作为服务节点。而这些节点可能分布于不同的地域、接入不同的ISP。如果请求节点忽略位置较近的节点而选择位置较远的节点作为服务节点,提供服务时流媒体资源需要经过较多的物理链路才能到达请求节点,增加了网络负载、引起了不必要的跨地域开销;

如果请求节点忽略自己所在ISP内部的节点,而选择不同ISP的节点作为服务节点,提供服务时流媒体资源需要跨过网络边界才能到达请求节点,引起了不必要的跨网络开销。另外,在搜索服务节点时如果不对搜索范围加以限制而采用全局洪泛搜索的方式,将会导致搜索效率低下,并大幅增加骨干网的通信负载,导致网络运营商有更强烈的愿望来限制P2P网络流量。

针对上述问题,国内外众多研究机构、专家学者都积极展开了研究。

文献[1]基于节点之间的网络延迟提出一种分箱策略,测量网络节点到地标节点集中各地标节点的网

收稿日期:2015-08-05

修回日期:2015-12-16

网络出版时间:2016-09-18

基金项目:国家自然科学基金资助项目(61462014)

作者简介:万明刚(1989-),男,硕士研究生,研究方向为计算机网络与流媒体技术;李泽平,通讯作者,博士,教授,硕士研究生导师,研究方向为计算机网络与流媒体技术。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20160918.1707.010.html>

络延迟,据此将网络节点分到不同的箱中。文献[2]基于网络坐标计算节点之间的欧氏距离,结合遗传聚类算法和 $K - \text{means}$ 算法提出一种混合聚类算法对节点进行有效聚类。文献[3-4]对计算机网络上的距离预测技术进行了研究。文献[5]提出一种基于兴趣相关度的 P2P 网络搜索优化算法,介绍了兴趣向量及兴趣相关度的描述方法。文献[6]描述了节点兴趣以及兴趣相关度的表示方法,借助层次聚类法和 $K - \text{means}$ 聚类法探讨 P2P 社区的形成过程。文献[7]提出一种基于网络拓扑和节点兴趣偏好的 P2P 搜索机制。文献[8-10]对近年来提出的较有代表性的聚类算法进行分析概括,介绍了聚类分析的研究热点、难点、不足和有待解决的一些问题。

现有这些对分组方法的研究多是基于节点位置或者节点兴趣某一个指标,设计出的算法对用户群体划分不够细致,不能兼顾搜索效率以及网络开销的问题,

分组结果具有明显的缺陷。文中提出一种基于多指标的用户群分组算法,结合基于节点位置分组与基于节点兴趣分组各自的优势,引入网络类型这一指标,利用这三个指标对用户群进行层次划分,尽可能将候选服务节点限制在同地域、同网络类型以及兴趣相似的节点小组内,以减少不必要的跨地域、跨网络开销,提高搜索效率。

2 体系结构

目前主流的流媒体分发结构有基于 CDN^[11-13]、基于 P2P^[14-15] 和 CDN-P2P^[16] 混合结构。文中研究的分组算法是基于 CDN-P2P 混合结构。将 CDN 的思想引入 P2P 网络,在骨干网部署 CDN 系统,在接入网构建 P2P 区域化网络,融合了 CDN 高可靠性和 P2P 低成本、高可扩展性的优势,应用前景广泛。

混合流媒体分发系统体系结构如图 1 所示。

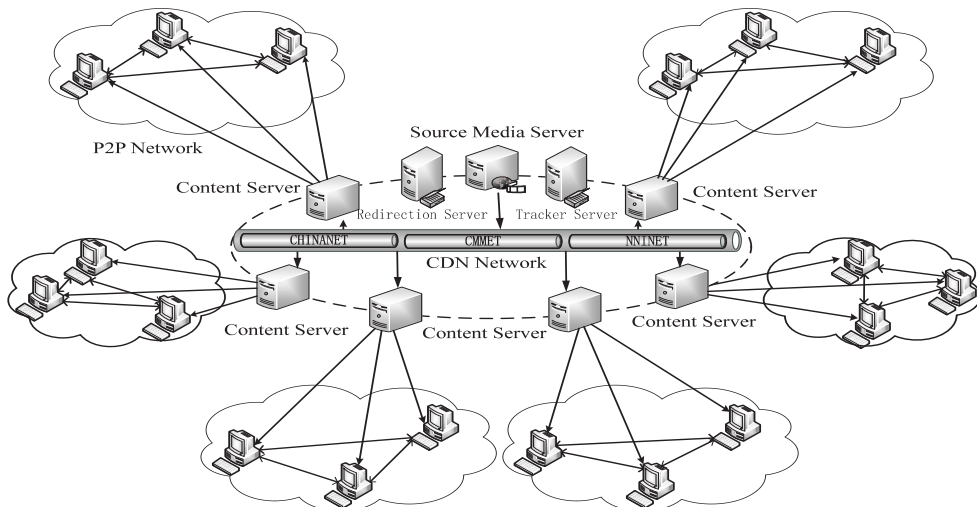


图 1 混合流媒体分发系统体系结构

3 分组算法

3.1 分组指标

文中选取节点位置、网络类型和节点兴趣三个指标设计混合流媒体分发系统中的用户群分组算法,尽可能将位置邻近、网络类型相同、兴趣近似的用户节点划分到同一个小组。

基于节点位置分组的基本思想是根据节点在网络中所处的位置直接计算节点间的网络距离或者估计节点间的网络邻近度,将网络距离较小或者相对邻近的节点划分到同一个位置组中,将候选服务节点限制在位置组内,以减少不必要的跨地域开销。其中,网络距离是一种抽象概念,通常用网络延迟来衡量,以 RTT 表示^[8]。

文中采用基于参考点的网络邻近度估计技术,选取混合流媒体分发系统中的边缘服务器节点作为地标

节点(Landmark),用节点到地标节点的网络延迟来定义它们之间的距离,根据每个节点到各地标节点的距离预估节点之间的邻近度,进而根据邻近度进行第一次基于节点位置的分组。

定义 1:节点位置。

给定地标节点集 $L = \{L_1, L_2, \dots, L_n\}$,用节点 n_i 到地标节点 $L_j, j = 1, 2, \dots, n$ 的网络延迟(RTT)来定义节点到地标节点的距离,进而估计节点在网络中的位置。

算法 1:基于节点位置分组算法。

对每个节点 n_i ,测量它到地标节点集 L 中所有地标节点的 RTT,并按升序排列,得到节点 n_i 到各个地标节点的 RTT 序列。将 RTT 序列相同的节点归到同一个位置组。

由于同一网络类型内的节点之间网络连接状况、通信质量比较有保证,而跨网络之间的通信质量比较差,因此选择服务节点时应尽量选择同一网络类型的

节点。基于网络类型分组的思想便是基于这一点。

定义 2: 网络类型。

根据网络运营商, 文中将网络类型分为电信、移动、联通三类。

算法 2: 基于网络类型分组算法。

将基于节点位置分组的结果按网络类型继续划分, 将每个位置组的用户群节点按其所接入的网络类型进一步划分为更小的网络类型组。

在流媒体分发系统中, 节点请求流媒体资源表现出一定的兴趣, 兴趣之间也表现出一定的相似程度。兴趣相似的节点之间共享资源的概率明显大于兴趣不相关的节点之间共享资源的概率。为此, 文中试图找出一种描述节点兴趣、度量兴趣相似程度的方法, 并据此将基于网络类型分组的结果按节点兴趣继续划分, 将候选服务节点限制在兴趣组内, 有效减小搜索流量、提高搜索效率。

定义 3: 节点兴趣。

节点请求流媒体资源时表现出一定的兴趣偏好, 可以用兴趣向量和兴趣相似度来度量。

定义 4: 兴趣向量。

将网络中的流媒体资源分类, 记录节点拥有各类资源的数量, 这样组成的一个向量即为该节点的兴趣向量。比如将网络中的流媒体资源分为 n 类, 节点 n_i 拥有各类资源的数量分别为 $\omega_{i1}, \omega_{i2}, \dots, \omega_{in}$, 则节点 n_i 的兴趣向量为 $\text{In}_i = (\omega_{i1}, \omega_{i2}, \dots, \omega_{in})$ 。

定义 5: 兴趣相似度。

假设节点 n_i 、 n_j 的兴趣向量分别为: $\text{In}_i = (\omega_{i1}, \omega_{i2}, \dots, \omega_{in})$ 、 $\text{In}_j = (\omega_{j1}, \omega_{j2}, \dots, \omega_{jn})$, 则节点 n_i 与 n_j 之间的兴趣相似度为:

$$\text{Sim}(n_i, n_j) = \frac{\text{In}_i \cdot \text{In}_j}{|\text{In}_i| \cdot |\text{In}_j|} = \frac{\omega_{i1} * \omega_{j1} + \omega_{i2} * \omega_{j2} + \dots + \omega_{in} * \omega_{jn}}{\sqrt{\omega_{i1}^2 + \omega_{i2}^2 + \dots + \omega_{in}^2} * \sqrt{\omega_{j1}^2 + \omega_{j2}^2 + \dots + \omega_{jn}^2}}$$

算法 3: 基于节点兴趣分组算法。

1) 层次聚类: 给定目标聚类数 k , 用层次聚类法确定 k 个初始聚类中心。

(1) 给定节点集合 $N = \{n_i, i = 1, 2, \dots, m\}$, 将集合 N 中每个节点看作一个具有单独成员的类。即:

$$N = C = \{C_i, i = 1, 2, \dots, m\}$$

其中, $C_i = \{n_i\}, i = 1, 2, \dots, m$ 。

(2) 计算 C 中每两个类之间的兴趣相似度 $\text{Sim}(C_i, C_j), i = 1, 2, \dots, m, j = 1, 2, \dots, m$ 。当 $\text{Sim}(C_i, C_j) = \text{Sim}(n_i, n_j) \geq \theta$ 时, 有:

$$\begin{cases} C_i = C_i \cup n_j \\ C_j = \text{删除数据} \end{cases}$$

这样得到一组新的 $C_i, i = 1, 2, \dots, m$ 。将 C_i 按模降序排列, 取前 k 个 C_i 对应的节点 n_i 作为初始聚类中心, 记为 $s = \{n_j, j = 1, 2, \dots, k\}$ 。

2) K -means 聚类。

依次计算 N 中每个节点 n_i 与每个种子节点 s_j 的兴趣相似度 $\text{Sim}(n_i, s_j)$, 找出与节点 n_i 兴趣相似度最大的种子节点 s_j , 将 n_i 归入以 s_j 为聚类中心的类。

3) 重复计算聚类中心 s 并重新聚类, 直至聚类结果稳定。

3.2 初始分组形成

算法 4: 初始分组形成算法。

Step1: 依次测量节点 $N = \{n_i, i = 1, 2, \dots, m\}$ 中每个节点 n_i 与地标节点集 L 中各地标节点的距离 $d(n_i, L_j)$, 得到 $\{d(n_i, L_j) | j = 1, 2, \dots, n\}$ 并按升序排列, 记录地标序。将节点 n_i 归入到与该地标序对应的分组 C_r 。

Step2: 依次对每一个位置组 C_r 根据节点 n_i 所属网络类型 (电信、移动、联通) 进一步划分为 $\{C_{r1}, C_{r2}, C_{r3}\}$, 将节点 n_i 归入相应的分组 $C_{rs}, s = 1, 2, 3$ 。

Step3: 依次对每一个网络类型组 C_{rs} 中的节点两两之间计算兴趣相似度 $\text{Sim}(n_i, n_j)$, 并根据兴趣相似度聚类, 将网络类型组 C_{rs} 进一步划分为 $\{C_{rs1}, C_{rs2}, \dots, C_{rsk}\}$, 将节点 n_i 归入相应的分组 C_{rst} , 得到最终分组 $\{C_{111}, C_{112}, \dots, C_{11k}, C_{121}, C_{122}, \dots, C_{12k}, \dots, C_{n31}, C_{n32}, \dots, C_{n3k}\}$, 将分组结果上报给 Tracker Server 并建立索引。

3.3 分组更新

考虑到网络环境的动态性, 节点之间的网络延迟、节点位置等都不是固定不变的, 因此设计的算法应该能随网络状况的变化而动态更新。

算法 5: 分组更新算法。

每经过一个固定的时间间隔 t , 根据当前网络状况重新计算分组。

4 服务节点选择

文中算法将对用户群分组的结果存储在追踪服务器 TR 上并建立索引, RS 是重定向服务器。

当一个节点需要获得某个流媒体资源时, 首先将服务请求同时发送给 RS 和 TR。RS 在查询了索引表之后, 根据某种重定向策略返回一个离请求节点距离较近且负载较轻的边缘服务器的地址信息。TR 在查询了索引表之后, 返回与请求节点同属于一个兴趣组的其他节点的地址信息。请求节点在收到返回的边缘服务器和同兴趣组节点的地址信息后, 同时向该兴趣组内其他节点发出流媒体服务请求, 在兴趣组内搜索

拥有所请求资源的 peer 节点作为候选服务节点。如果搜索到有节点存储有所请求的流媒体资源,则在其中选取一个或几个节点作为服务节点为请求节点提供服务;如果搜索到没有节点存储有所请求的流媒体资源,则将服务请求重定向到边缘服务器,由边缘服务器为其提供服务。

服务节点选择过程如图 2 所示。

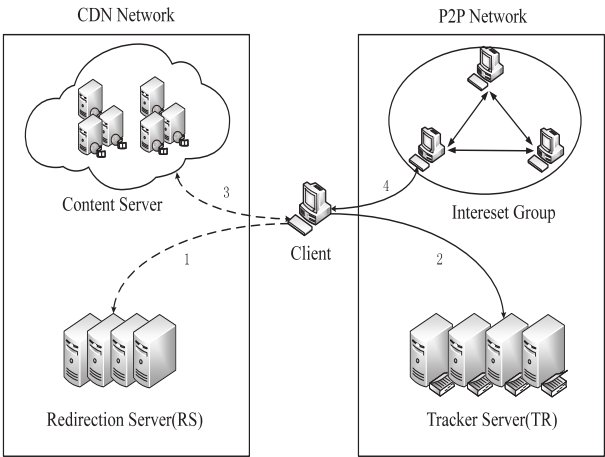


图 2 服务节点选择过程

5 仿真及结果分析

实验在一台 PC 机上完成,模拟具有 1 000 个节点的对等网络。每个节点用一个数据结构模拟,数据结构中定义了该节点到各地标节点的网络延迟、所属网络类型以及节点兴趣。实验数据采用人工模拟的方式获得,在设定的各个参数取值范围内随机赋值。为简化模型,假设:地标节点数为 3,每个节点到各地标节点的网络延迟在[0,100]区间内随机取整数值;节点网络类型取值为 0,1,2,分别对应电信、移动、联通;网络中流媒体资源分为 5 类,每个节点拥有每类资源的数量在[0,10]区间内随机取整数值。分组算法用 Java 语言编程实现,取 K-means 算法中 $k=3$,这样最终将 1 000 个节点分为 54 组。每次实验随机选取一个节点作为请求节点,对其最感兴趣的一类资源发出搜索请求。

在实验中对比较分析了全局洪泛法、基于节点位置分组算法、基于网络类型分组算法、基于节点兴趣分组算法和文中算法五种服务节点选择方案的性能差异。为方便比较,定义了四个主要性能指标:

跨地域服务节点占比=搜索到的节点中属于不同地域的节点数/搜索到的节点总数

跨网络类型服务节点占比=搜索到的节点中属于不同网络类型的节点数/搜索到的节点总数

搜索到的资源占比=搜索到的某类资源数/网络中该类资源总数

搜索范围占比=搜索范围内所有节点数/网络中总节点数

采用多次重复实验的方式,每次实验记录下各个性能指标的值,用 Excel 和 Matlab 对实验数据进行处理,结果如图 3~6 所示。

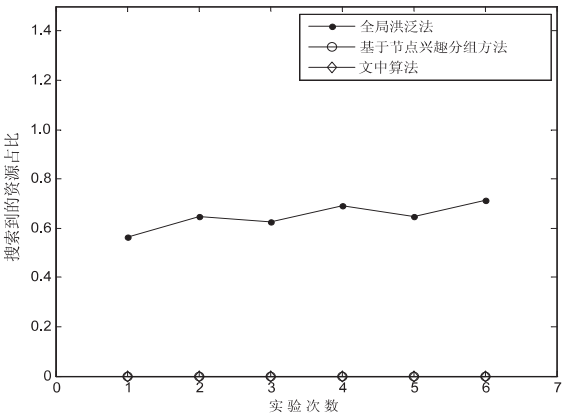


图 3 跨地域服务节点占比

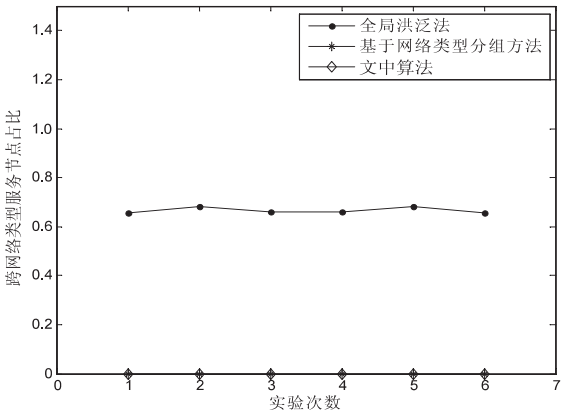


图 4 跨网络类型服务节点占比

图 3、图 4 表明:相比较全局洪泛法选择服务节点而言,文中所提用户群分组算法集合了基于节点位置分组与基于网络类型分组的优点,能有效将服务节点选择范围限制在同地域、同网络类型范围内,从而达到减少不必要的跨地域、跨网络开销的目的。

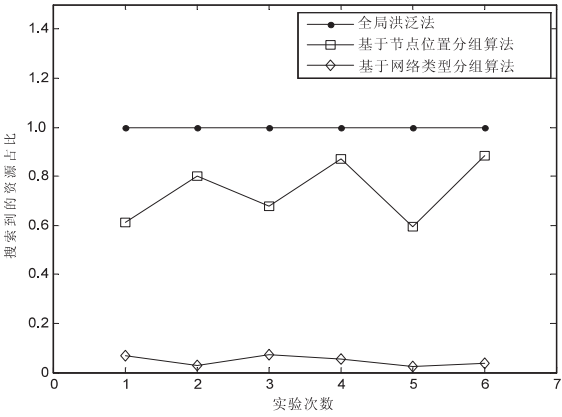


图 5 搜索到的资源占比

图 5、图 6 表明:文中所提用户群分组算法较其他几种常见分组算法而言,选择服务节点时搜索范围明

显更小,能有效降低搜索流量,同时也能保证一定的资源搜索量。

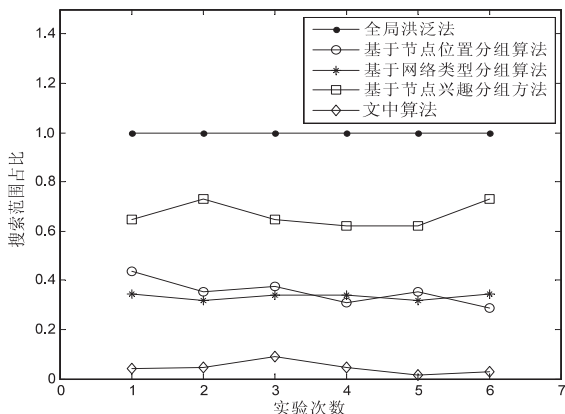


图 6 搜索范围占比

6 结束语

文中提出一种混合流媒体分发系统中的多指标用户群分组算法,并通过仿真实验验证了该算法的有效性,但对于其在真实网络环境中的应用价值尚有待验证。下一步的研究将集中于此,将该算法部署到流媒体分发系统中加以论证。

参考文献:

- [1] Ratnasamy S, Handley M, Karp R, et al. Topologically-aware overlay construction and server selection [C]//Proc of INFOCOM 2002. [s. l.]: IEEE, 2002: 1190–1199.
- [2] 周振朝, 费耀平, 李敏. 基于网络坐标的无结构 P2P 节点聚类算法[J]. 计算机工程, 2010, 36(11): 98–100.
- [3] 邢长友, 陈鸣. 网络距离预测技术[J]. 软件学报, 2009, 20(9): 2470–2482.
- [4] 王意洁, 李小勇. 网络距离预测技术研究[J]. 软件学报, 2009,

20(6): 1574–1590.

- [5] 吴思, 欧阳松. 基于兴趣相关度的 P2P 网络搜索优化算法[J]. 计算机工程, 2008, 34(11): 102–104.
- [6] 赵捧未, 马琳, 秦春秀. P2P 用户兴趣社区形成研究[J]. 现代图书情报技术, 2013(10): 53–58.
- [7] 梁卫芳, 黄建华. 基于网络拓扑和节点兴趣的 P2P 搜索机制[J]. 计算机工程与设计, 2008, 29(6): 1316–1318.
- [8] 郑力明, 李晓冬, 李小勇, 等. P2P 覆盖网中的聚类研究综述[J]. 计算机应用研究, 2010, 27(3): 806–810.
- [9] 杨博, 刘大有, LIU Jiming, 等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54–66.
- [10] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48–61.
- [11] Adhikari V K, Jain S, Chen Y, et al. Vivisecting YouTube: an active measurement study [C]//Proc of INFOCOM. Orlando, FL: IEEE, 2012: 2521–2525.
- [12] Adhikari V K, Guo Y, Hao F, et al. Unreeling netflix: understanding and improving multi-CDN movie delivery [C]//Proc of INFOCOM. Orlando, FL: IEEE, 2012: 1620–1628.
- [13] Adhikari V K, Guo Y, Hao F, et al. A tale of three CDNs: an active measurement study of Hulu and its CDNs [C]//Proc of IEEE conference on computer communications workshops. Orlando, FL: IEEE, 2012: 7–12.
- [14] Huang Yan, Fu T Z J, Chiu Dah-Ming, et al. Challenges, design and analysis of a large-scale p2p-vod system [C]//Proc of SIGCOMM. [s. l.]: ACM, 2008: 375–388.
- [15] Lei J, Shi L, Fu X. An experimental analysis of Joost peer-to-peer VoD service [J]. Peer-to-Peer Netw. Appl., 2010, 3(4): 351–362.
- [16] Zhang G, Liu W, Hei X, et al. Unreeling Xunlei Kankan: understanding hybrid CDN-P2P video-on-demand streaming [J]. IEEE Transactions on Multimedia, 2015, 17(2): 229–242.

(上接第 35 页)

- 2010, 42(9): 1428–1431.
- [7] 周明, 孙树栋. 遗传算法原理及应用[M]. 北京: 国防工业出版社, 2001.
- [8] 谢政, 李建平. 网络算法与复杂性理论[M]. 长沙: 国防科技大学出版社, 2003.
- [9] Alsultanny Y A, Aqel M M. Pattern recognition using multilayer neural-genetic algorithm [J]. Neurocomputing, 2003, 51: 237–247.
- [10] 康立山. 非数值并行算法(第一册): 模拟退火算法[M]. 北京: 科学出版社, 2000: 22–38.
- [11] 赵礼峰, 王小龙. 图的 Steiner 最小树问题的混合遗传算法

[J]. 计算机技术与发展, 2014, 24(10): 110–114.

- [12] 鲁建业, 李琦, 董蕴华, 等. 采用混合遗传-模拟退火算法对 DOE 的直接设计[J]. 光电子·激光, 2001, 12(4): 365–367.
- [13] Gao Erbao, Lai M. An improved genetic algorithm for the vehicle routing problem with simultaneous delivery and pickup [C]//Proc of the 6th Wuhan international conference on e-business – innovation management track. Wuhan: [s. n.], 2007: 2100–2106.
- [14] 王海英. MATLAB 遗传算法工具箱及应用[M]. 北京: 北京航空航天大学出版社, 2010.