

文章编号: 1005-0523(2005)01-0107-04

# 遗传算法在集中管理式 Web 缓存系统的应用

郭家堂<sup>1</sup>, 郭厚<sup>1</sup>, 王 勇<sup>1</sup>, 吕 镇<sup>1</sup>, 卢伟海<sup>2</sup>

(华东交通大学 1. 机电工程学院; 2. 电气工程学院 江西 南昌, 330013)

**摘要:** 本文对集中管理式 Web 缓存系统进行了改进, 应用遗传算法寻找下载的最优路径, 既解决了传统的 Web 高速缓存 (Web Caching) 中存在的各缓存服务器之间因不断对话而牺牲不必要带宽的问题, 又增加了下载速度。

**关键词:** 遗传算法; 集中管理; 缓存系统

**中图分类号:** TP393.02

**文献标识码:** A

## 1 引言

随着 Internet 用户的迅速增多, WWW 的流行, 人们在工作、学习、生活等多方面享受着新技术带来的便利。但由于网络内容和网络用户急剧增加, 使网络负载不断加重。同时, 用户对网络的速度和效果愈加重视, 对网络的访问质量提出了更高的要求。

目前, 许多网络公司研究出对数据进行缓存 (Caching) 来节约成本和提高下载速度。然而, 计算临时数据存放的位置及其时间是一个相当复杂的问题。一个可能的办法是让缓存之间不断进行“对话”, 不过这种牺牲带宽的做法仍然会严重影响到网络效率的发挥。于是文献 [1] 提出了集中管理式 Web 缓存系统方案, 解决了缓存之间不断进行“对话”而牺牲带宽的问题。文献 [2][3][4] 等对 Cache 的结构和替换算法等提出了很好的方案, 但是均未对用户与 Cache 服务器间下载路径方式进行考虑。

本文在文献 [1] 提出的集中管理式 Web 缓存系统的基础上, 分析了用遗传算法对用户与 Cache 服务器之间下载路径进行优化的可行性, 并给了实现方法。

## 2 集中管理式 Web 缓存技术简介

集中管理式 Web 缓存技术是在 CERNET (China Education and Research Network) 基础上建立的, 其结构如图 1 所示, 其中不同于基于 ICP 的简单缓存系统之处是增加了一台集中管理器, 用户所有的请求必须先经这台集中管理器才能得到缓存服务器的服务, 负载均衡设备将用户的请求合理地分配给某一合适缓存服务器, 没命中的请求再交给父亲缓存服务器, 所以各兄弟缓存间不再进行对话。用户请求/对答过程是: 用户 <—> 某一合适的缓存服务器 <—> 父亲缓存服务器 <—> 源服务器。

集中管理式 Web 缓存技术解决了各兄弟缓存间不断进行对话而牺牲带宽的问题。但是当网络父亲缓存服务器和兄弟缓存服务器 X 同时存在客户所需要的资源时, 理想情况下, 客户可以有两种方式得到他所需要的资源: 方案一: 客户直接从兄弟缓存服务器 X 下载他要的资源; 方案二: 通过改进负载均衡设备, 能让客户借助另一兄弟缓存服务器 Y 从父亲缓存服务器得到资源。由于客户到不同兄弟缓存服务器间得到他所需要的资源花费 (费用 + 时间) 存在差异, 往往会出现方案二比方案一更加

收稿日期: 2004-10-12

作者简介: 郭家堂 (1982-), 男, 浙江龙泉人, 华东交通大学在读本科生。

理想的情况. 本文将用遗传算法对方案二进行可行性分析.

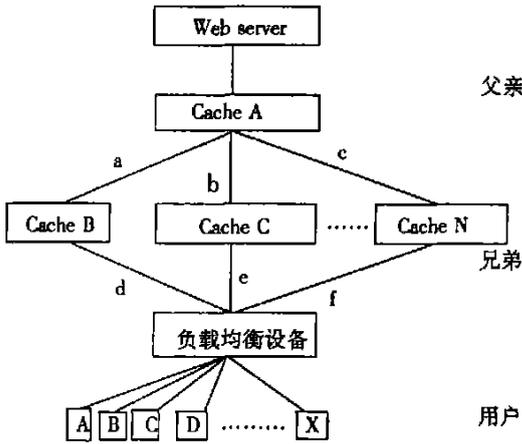


图1 集中管理式 Web 缓存系统结构图

使用保留最佳个体策略的遗传算法能搜索到最优解. 具体运算如图 2

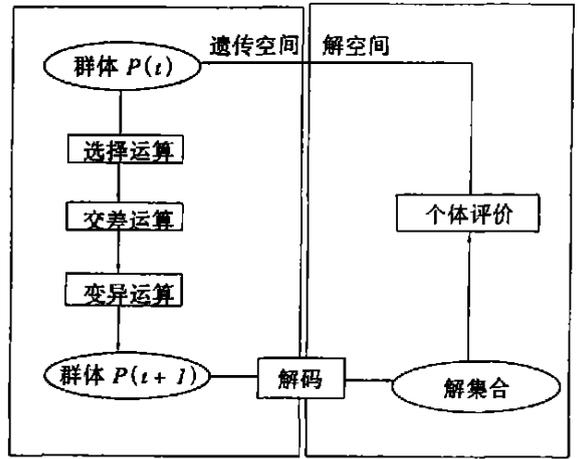


图2 遗传算法流程图

### 3 集中管理式 Web 缓存系统的改进

改进后的集中管理式 Web 缓存系统中负载均衡设备具备的功能: 1) 负载均衡设备对所有的缓存服务器进行所要的资源进行搜索; 2) 记下拥有所要资源的所有缓存服务器; 3) 对记下的每个缓存服务器下载路径进行遗传算法, 找出最优路径; 4) 比较各个最优路径, 找出总体最优路径, 此路径就是客户最终下载路径.

## 4 基于遗传算法的集中管理式 Web 缓存系统

### 4.1 遗传算法

遗传算法是模拟生物遗传进化机制而发展起来的一种算法, 是模仿生物界自然选择原理和遗传机制的随机搜索最优算法. 它把类似于遗传基因的一些行为, 如交叉重组、变异、选择和淘汰等引入到算法求解的改进过程中, 具有“生成+检测”的迭代搜索. 遗传算法的特点之一是, 它同时保留着若干局部最优解, 通过交叉重组或者解的变异来寻求更好的解.

基本的遗传算法可描述为一个齐次 Markov 链  $P_t = \{P(t), t \geq 0\}$ , 因为基本遗传算法的选择、交叉和变异操作都是独立随机进行, 新群体只与父代群体及遗传操作算子有关, 与父代之前的各群体无关, 并且各代群体之间的转换概率与时间起点无关. 在数学上已经证明了使用保留最佳个体策略的遗传算法能收敛于最优解的概率为 1, 这也说明了

### 4.2 模型的构件

我们可以用  $X = (C, D)$  代表整个 Web 缓存网络, 其中  $C = \{c_1, c_2, \dots, c_n\}$  代表各 Web 缓存服务器,  $D = \{d_1, d_2, \dots, d_n\}$  代表通信链路的下载费用和时间(折合为数值), 所以选择任意下载路径要消耗的网络花费  $H$  (费用+时间) 为:

$$H = \sum d_i \tag{1}$$

### 4.3 遗传算法的实现

#### 1) 个体的编码方法

我们对每一个 Web 缓存服务器进行  $m$  位二进制编码, 其中前  $t$  位是代表 Web 缓存服务器级别(如父亲级、兄弟 I 级、兄弟 II 级……)后  $k = m - t$  位是 Web 缓存服务器所在级内的编码, 这样编码的将有利于后面计算.

#### 2) 适应度函数

个体适应度评价函数直接影响到遗传算法的计算效率. 这里, 选择如下所示的个体适应度评价函数:

$$F = H$$

由于选择算子采用比例选择算子, 为了避免运算过程中出现“早熟”和收敛速度的降低, 本文采用文献[5]提供的自适应的比例变换方式, 具体的变换公式如下:

$$f'(x) = f(x) - q_{\min}$$

其中  $f(x)$  —— 变换前的适应度;  $f_{\min}$  —— 变换前适应度的最小值;  $f'(x)$  —— 变换后的适应度;  $a$  —— 自适应控制参数;  $a = 0.9 - (f_{\max} - f_{\min}) / f_{\max}$ ;

其中  $f_{\max}$  —— 变换前适应度的最大值;  $f_{\min}$  —— 变换前适应度的最小值.

### 3) 遗传算子的设计

(1) 选择算子. 采用比例选择算子, 使个体按照与适应度成正比的概率向下一代群体繁殖.

(2) 交叉算子. 采用单点交叉, 实现过程为随机选择 2 个已执行选择操作的染色体作为父本, 再随机选择一个交叉位置, 将这 2 个父本染色体依设定的交叉率  $P_c$ , 在其交叉点处相互交换 2 个个体部分染色体, 从而产生 2 个新个体.

(3) 变异算子. 用三种操作方式, 即从个体中随机撤除一个序号 (不包括起始点、终止点序号); 或从个体中随机选取一点, 插入一个新序号; 或在个体中随机选取一个序号, 用另一个随机产生的序号代替它. 显然, 这三种变异算子都有可能产生不连续的相邻等级缓存服务器点, 于是增加第四个算子.

(4) 插入算子. 把不连续的相邻等级缓存服务器点弥补, 使之连续. 可以用下述判别式来判断个体相邻等级是否连续:

$$\Delta = abs(t_{k+1} - t_k);$$

式中的  $t_{k+1}$ 、 $t_k$  分别是代表  $C_{k+1}$ 、 $C_k$  的 Web 缓存服务器级别 (如父亲级、兄弟 I 级、兄弟 II 级……). 若  $\Delta=1$ , 则  $C_{k+1}$  与  $C_k$  连续, 否则不连续.

不连续时, 按照中值法计算候补插入点:

$$t'_k = (t_{k+1} + t_k) / 2$$

计算出  $t'_k$  后, 找出一个所在级别编码为  $t'_k$  的 Web 缓存服务器插入. 上面过程反复执行, 直到个体变为连续的兄弟链为止.

(5) 删除算子. 当个体中存在两个或两个以上属于同一兄弟级 Web 缓存服务器时 (如图 1 中 Cache A ↔ Cache B ↔ Cache C ↔ 负载均衡设备这一链路中的 Cache B 与 Cache C), 将多余的删除, 只保留一个. 因为在集中管理式 Web 缓存系统中各兄弟缓存间不进行对话, 所以图 1 中 Cache A ↔ Cache B ↔ Cache C ↔ 负载均衡设备这一链路不通.

### 4.4 实例分析

本文采用图 1 的网络拓扑结构. 遗传算法的交叉概率为  $P_c = (0.9 - a)$ , 变异概率为  $P_m = 0.01 * (0.1 + a)$ ,  $P_m$  的取值范围为  $[0, 0.05]$ , 群体规模为 30, 遗传代数 100, 网中各边的费用和时间用数据由表 1 给出.

表 1 网络中各边的花费

组	边					
	a	b	c	d	e	f
1	7	2	6	8	4	16
2	6	5	9	5	7	9

不失一般性, 假设 Cache A 和 Cache N 同时存在客户需要的资源, 分别用表 1 的网络各边的花费数据, 经遗传算法运算后的到了如图 3 所示. 为进一步分析遗传算法的性能, 采用表 1 中的第一组数据, 在不同规模和遗传代数下用遗传算法重复做 100 次实验, 结果得到了 100% 的一致性. 由图 4 和计算结果可以看出, 在表 1 的第一组数据的情况下, 我们通过 Cache C 从 Cache A 下载数据的花费比直接从 Cache N 下载数据的花费要少; 在表 1 的第二组数据的情况下, 我们直接从 Cache N 下载数据的花费比通过 Cache C 从 Cache A 下载数据的花费要多. 在实际中表 1 中的第一种情况出现的可能性是非常大的.

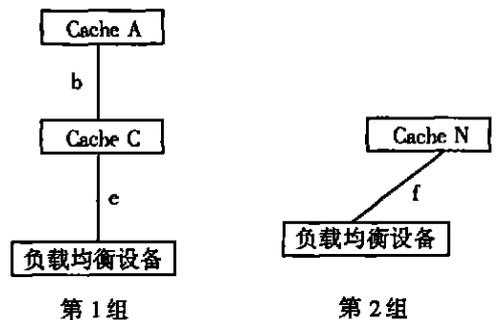


图 3 不同花费下的最优下载路径

## 4 结束语

本文在文献[1]提出的集中管理式 Web 缓存技术系统的基础上, 运用文献[5]的改进遗传算法, 结合实际问题, 对该系统提出了进一步的改进方案, 在避免传统的 Web 缓存技术中过大牺牲带宽同时提高了下载的速度, 降低了花费.

### 参考文献:

- [1] 姜彩萍, 李子木, 杨风杰. 集中管理式 Web 缓存系统及性能分析[J]. 小型微型计算机系统, 2004, 25(8): 1428—1431.
- [2] 贺琛, 陈肇雄, 黄河燕. Web 缓存技术综述[J]. 小型微型计算机系统, 2004, 25(5): 836—842.
- [3] 张兴军, 钱德沛, 刘轶, 朱利, 李越. 基于内容的 Web Caching 小型微型计算机系统 2004, 25(8): 1415—

- 1419.
- [4] 陈章龙. 嵌入式处理器的 Cache 结构研究[J]. 小型微型计算机系统, 2004, 25(7): 1204—1206.
- [5] 陈曦, 林涛, 唐贤瑛. 遗传算法的参数设计与性能研究[J]. 计算机工程与设计 2004, 25(8): 1309—1310, 1319.
- [6] 姜圳, 张宏科, 张礼勇. 基于遗传算法的流媒体组播路由选择方法[J]. 北京邮电大学学报, 2004, 27(2): 39—43.
- [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, 1999.

## Application of Genetic Algorithm in Web Caching System Based on Concentrated Management

GUO Jia-tang<sup>1</sup>, GUO Hou-kun<sup>1</sup>, WANG Yong<sup>1</sup>, LV Zhen<sup>1</sup>, LU Wei-hai<sup>2</sup>

(1. School of Mechanical & Electrical Engineering; 2. School of Electric & Electronic Engineering, East China Jiaotong University, Nanchang 330013, China)

**Abstract:** In this paper, the web caching system based on concentrated management is improved, and genetic algorithm is applied in finding the best route of downloading, which both resolves the problem of unnecessary waste of bandwidth in the constant conversations among all cache servers in the process of traditional web caching and increases the speed of downloading.

**Key words:** genetic algorithm; concentrated management; caching system.

(上接第 91 页)

## The Design of Hydrological Forecast System Based on VISUAL C++6.0

GE De-sheng, ZHANG Zhong-hui

(Nanchang University, Nanchang 330029, China)

**Abstract:** For precisely forecasting the reservoir inflow in time, According to the hydrological characteristic of HongMen reservoir basins. Using the Xinanjiang watershed model, the flood forecasting system is studied based on the system with Visual C++6.0 software package. The general project of this system was presented, and the basic framework and the main function was analyzed in this paper.

**Key words:** Visual C++6.0, Xinanjiang watershed model, flood forecasting, the general project