



## (12) 发明专利

(10) 授权公告号 CN 101494566 B

(45) 授权公告日 2011.09.14

(21) 申请号 200810065629.9

(22) 申请日 2008.01.23

(73) 专利权人 华为技术有限公司

地址 518129 广东省深圳市龙岗区坂田华为  
总部办公楼Churn in Peer-to-Peer Networks.《Proceedings  
of the 6th ACM SIGCOMM conference on  
Internet measurement》.2006, 189–202.

审查员 张玉洁

(72) 发明人 施广宇 龚皓

(51) Int. Cl.

H04L 12/26(2006.01)

H04L 12/24(2006.01)

H04L 25/03(2006.01)

(56) 对比文件

CN 1972206 A, 2007.05.30,

US 2007/0061232 A1, 2007.03.15,

CN 101043525 A, 2007.09.26,

Daniel Stutzbach et al. Understanding

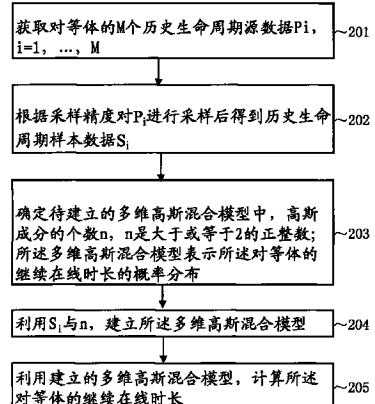
权利要求书 3 页 说明书 8 页 附图 4 页

(54) 发明名称

预测对等网络中对等体的继续在线时长的方  
法及装置

(57) 摘要

本发明实施例提供一种预测对等网络中对等体的继续在线时长的方法，包括：获取对等体的M个历史生命周期样本数据 $S_i$ ,  $i = 1, \dots, M$ ；确定待建立的多维高斯混合模型中，高斯成分的个数n, n是大于或等于2的正整数；所述多维高斯混合模型表示所述对等体的继续在线时长的概率分布；利用 $S_i$ 与n, 建立所述多维高斯混合模型；利用被建立的多维高斯混合模型，预测所述对等体的继续在线时长。本发明实施例还提供预测对等网络中对等体的继续在线时长的装置。本发明实施例提供的技术方案，基于历史生命周期以及多维高斯混合模型，能够预测出较逼近对等体节点的实际继续在线时长的预测结果。



1. 一种预测对等网络中对等体的继续在线时长的方法,其特征在于,包括:

获取对等体的 M 个历史生命周期样本数据  $S_i, i = 1, \dots, M$ ;

确定待建立的多维高斯混合模型中,高斯成分的个数 n,n 是大于或等于 2 的正整数,所述多维高斯混合模型表示所述对等体的继续在线时长的概率分布;

根据已知量  $s_i$  和 n,利用贝叶斯定理,计算多维高斯混合模型中各多维高斯成分的高斯分布参数以及每个高斯成分所占的权重,建立所述多维高斯混合模型,其中  $s_i$  表示各  $S_i$  出现的概率;

利用建立的多维高斯混合模型,计算所述对等体生命周期的出现概率密度,根据所述对等体生命周期的出现概率密度计算所述对等体继续存活预设变化时长的概率,根据所述概率确定所述对等体的继续在线时长。

2. 根据权利要求 1 所述的方法,其特征在于,所述的历史生命周期样本数据  $S_i$  的内容包括:对等体节点的历史在线时长信息和对等体节点的历史上线的时刻信息。

3. 根据权利要求 1 所述的方法,其特征在于,所述的历史生命周期样本数据  $S_i$  的内容包括:对等体节点的历史在线时长信息,对等体节点的历史上线的时刻信息,和对等体节点的历史上线的区间信息。

4. 根据权利要求 1 所述的方法,其特征在于,所述获取对等体的 M 个历史生命周期样本数据  $S_i$ ,包括:

获取对等体的 M 个历史生命周期源数据  $P_i, i = 1, \dots, M$ ;

对源数据  $P_i$  按采样精度进行采样得到历史生命周期样本数据  $S_i$ 。

5. 根据权利要求 1 所述的方法,其特征在于,预测所述对等体的继续在线时长后,该方法进一步包括:

根据确定出的所述继续在线时长,以及所述对等体已在线时长,算出所述对等体本次存活的生命周期或一继续在线时长可能出现的概率。

6. 根据权利要求 4 所述的方法,其特征在于,所述历史生命周期样本数据或历史生命周期源数据由所述对等体自身记录并保存,或者,通过让所述对等体节点发送历史生命周期样本数据或历史生命周期源数据到中心服务器的方法来集中获取。

7. 一种预测对等网络中对等体的继续在线时长的装置,其特征在于,包括:

获取单元,获取对等体的 M 个历史生命周期样本数据  $S_i, i = 1, \dots, M$ ;

模型建立单元,利用获取单元获取的  $S_i$ ,以及高斯成分的个数 n,n 是大于或等于 2 的正整数,根据已知量  $s_i$  和 n,利用贝叶斯定理,计算多维高斯混合模型中各多维高斯模型的高斯分布参数以及每个高斯成分所占的权重,建立多维高斯混合模型,所述高斯混合模型表示所述对等体的继续在线时长的概率分布,其中  $s_i$  表示各  $S_i$  出现的概率;

预测单元,利用模型建立单元建立的所述多维高斯混合模型,计算所述对等体生命周期的出现概率密度,根据所述对等体生命周期的出现概率密度计算所述对等体继续存活预设变化时长的概率,根据所述概率确定所述对等体的继续在线时长。

8. 根据权利要求 7 所述的装置,其特征在于,所述获取单元包括:

第一单元,获取对等体的 M 个历史生命周期源数据  $P_i, i = 1, \dots, M$ ;

第二单元,根据采样精度,对历史生命周期源数据  $P_i$  进行采样得到  $S_i$ ;

9. 根据权利要求 7 所述的装置,其特征在于,所述的历史生命周期样本数据  $S_i$  的内容

包括：对等体节点的历史在线时长信息和对等体节点的历史上线的时刻信息。

10. 根据权利要求 7 所述的装置，其特征在于，所述的历史生命周期样本数据  $S_i$  的内容包括：对等体节点的历史在线时长信息，对等体节点的历史上线的时刻信息，和对等体节点的历史上线的区间信息。

11. 根据权利要求 7 所述的装置，其特征在于，所述模型建立单元包括：

参数计算单元，计算所述多维高斯混合模型的混合模型参数，

所述混合模型参数包括：

每个高斯成分对应的多维高斯分布的分布参数，以及所述多维高斯混合模型中，每个高斯成分所占权重。

12. 根据权利要求 11 所述的装置，其特征在于，所述参数计算单元包括：

概率计算单元，根据采样得到的 M 个  $S_i$ ，算出各  $S_i$  的出现概率  $s_i$ ；

函数构建单元，建立包括所述多维混合模型参数的似然函数；

估算单元，利用所述概率计算单元算出的  $s_i$ ，计算所述函数构建单元构建的所述似然函数取最大值时，各多维混合模型参数的估算值。

13. 根据权利要求 7 所述的装置，其特征在于，预测单元包括：

概率密度计算单元，计算所述对等体生命周期的出现概率密度；

存活概率计算单元，利用概率密度计算单元算出的所述出现概率密度，计算所述对等体继续存活预设变化时长的概率；

时长计算单元，利用所述存活概率计算单元算出的概率，求取所述对等体的继续在线时长。

14. 根据权利要求 7 所述的装置，其特征在于，所述装置进一步包括：生命周期预测单元，根据所述预测单元预测出的所述继续在线时长，以及所述对等体已在线时长，算出所述对等体本次存活的生命周期或一继续在线时长可能出现的概率。

15. 根据权利要求 7 所述的装置，其特征在于，所述装置进一步包括：接收单元，接收待建立的多维高斯混合模型中，高斯成分的个数 n，n 是大于或等于 2 的正整数。

16. 一种建立对等网络中对等体的生命周期模型的方法，其特征在于，包括：

获取对等体的 M 个历史生命周期样本数据  $S_i$ ,  $i = 1, \dots, M$ ；

确定待建立的多维高斯混合模型中，高斯成分的个数 n，n 是大于或等于 2 的正整数，所述多维高斯混合模型表示所述对等体的继续在线时长的概率分布；

根据已知量  $s_i$  和 n，利用贝叶斯定理，计算多维高斯混合模型中各多维高斯模型的高斯分布参数以及每个高斯成分所占的权重，建立所述多维高斯混合模型，其中  $s_i$  表示各  $S_i$  出现的概率；

将所述高斯混合模型的描述信息发送。

17. 根据权利要求 16 所述的方法，其特征在于，所述的历史生命周期样本数据  $S_i$  的内容包括：对等体节点的历史在线时长信息和对等体节点的历史上线的时刻信息。

18. 根据权利要求 16 所述的方法，其特征在于，所述的历史生命周期样本数据  $S_i$  的内容包括：对等体节点的历史在线时长信息，对等体节点的历史上线的时刻信息，和对等体节点的历史上线的区间信息。

19. 一种建立对等网络中对等体的生命周期模型的装置，其特征在于，包括：

获取单元, 获取对等体的 M 个历史生命周期样本数据  $S_i, i = 1, \dots, M$ ;

模型建立单元, 利用获取单元获取的  $S_i$ , 以及高斯成分的个数 n, 根据已知量  $s_i$  和 n, 利用贝叶斯定理, 计算多维高斯混合模型中各多维高斯模型的高斯分布参数以及每个高斯成分所占的权重, 建立多维高斯混合模型, n 是大于或等于 2 的正整数, 所述高斯混合模型表示所述对等体的继续在线时长的概率分布, 其中  $s_i$  表示各  $S_i$  出现的概率;

发送单元, 将模型建立单元建立的所述高斯混合模型的描述信息发送。

20. 根据权利要求 19 所述的装置, 其特征在于, 所述的历史生命周期样本数据  $S_i$  的内容包括: 对等体节点的历史在线时长信息和对等体节点的历史上线的时刻信息。

21. 根据权利要求 19 所述的装置, 其特征在于, 所述的历史生命周期样本数据  $S_i$  的内容包括: 对等体节点的历史在线时长信息, 对等体节点的历史上线的时刻信息, 和对等体节点的历史上线的区间信息。

## 预测对等网络中对等体的继续在线时长的方法及装置

### 技术领域

[0001] 本发明涉及对等网 (P2P, Peer-to-Peer) 技术领域，尤其涉及一种预测对等网络中对等体的继续在线时长的方法及装置。

### 背景技术

[0002] 与传统的客户机 / 服务器模式不同，P2P 网络中不存在中心服务器节点，其中，每个节点既可用作服务器为其他节点提供服务，同时，又可以享受其他节点用作服务器时所提供的服务。因此，P2P 网络中，每个 Peer 节点处于对等地位，称每个节点为一个对等体，或一个 Peer。

[0003] P2P 网络是一种自组织形态的网络系统，该网络中，每个 Peer 加入网络或从网络中推出的行为均是随机性的。由于 P2P 网络中，每个 Peer 均作为一个为其他 Peer 提供服务的服务器，因此，Peer 加入或退出系统的随机性，会对节点间的数据传输造成扰动，如另一 Peer 在该 Peer 下线之前，连接到该 Peer，准备从该 Peer 下载数据，但因该 Peer 的突然下线，一方面使得该 Peer 不能够再作为服务器为另一 Peer 提供服务，另一方面，另一 Peer 需要重新变更路由，到其他 Peer 上获取相关数据。Peer 上线行为的随机性，会影响网络系统的正常运行，并导致整个系统性能的下降。称因 Peer 上线行为的随机性给 P2P 网络系统造成的影响为扰动 (Churn) 现象。

[0004] 需要采取相应措施，以尽量避免 Churn 现象给系统造成的不良影响，以提高 P2P 网络的抗干扰 (Churn Resistant) 能力。

### 发明内容

[0005] 本发明的实施例提供一种预测对等网络中对等体的继续在线时长的方法及装置，能够预测出的继续在线的时长。

[0006] 本发明的实施例提供一种预测对等网络中对等体的继续在线时长的方法，包括：

[0007] 获取对等体的 M 个历史生命周期样本数据  $S_i, i = 1, \dots, M$ ；

[0008] 确定待建立的多维高斯混合模型中，高斯成分的个数 n，n 是大于或等于 2 的正整数；所述多维高斯混合模型表示所述对等体的继续在线时长的概率分布；

[0009] 利用  $S_i$  与 n，建立所述多维高斯混合模型；

[0010] 利用被建立的多维高斯混合模型，预测所述对等体的继续在线时长。

[0011] 本发明的实施例提供一种预测对等网络中对等体的继续在线时长的装置，包括：

[0012] 获取单元，获取对等体的 M 个历史生命周期样本数据  $S_i, i = 1, \dots, M$ ；

[0013] 模型建立单元，利用采样单元获取的  $S_i$ ，以及高斯成分的个数 n，建立多维高斯混合模型，n 是大于或等于 2 的正整数，所述高斯混合模型表示所述对等体的继续在线时长的概率分布；

[0014] 预测单元，利用模型建立单元建立的所述多维高斯混合模型，预测所述对等体的继续在线时长。

- [0015] 本发明的实施例提供一种建立对等网络中对等体的生命周期模型的方法，包括：
- [0016] 获取对等体的 M 个历史生命周期样本数据  $S_i, i = 1, \dots, M$ ；
- [0017] 确定待建立的多维高斯混合模型中，高斯成分的个数 n，n 是大于或等于 2 的正整数；所述多维高斯混合模型表示所述对等体的继续在线时长的概率分布；
- [0018] 利用  $S_i$  与 n，建立所述多维高斯混合模型；
- [0019] 将所述高斯混合模型的描述信息发送。
- [0020] 本发明的实施例提供一种建立对等网络中对等体的生命周期模型的装置，包括：
- [0021] 获取单元，获取对等体的 M 个历史生命周期样本数据  $S_i, i = 1, \dots, M$ ；
- [0022] 模型建立单元，利用采样单元获取的  $S_i$ ，以及高斯成分的个数 n，n 是大于或等于 2 的正整数，建立多维高斯混合模型，所述高斯混合模型表示所述对等体的继续在线时长的概率分布；
- [0023] 发送单元，将模型建立单元建立的所述高斯混合模型的描述信息发送。
- [0024] 本发明实施例提供的预测对等网络中对等体的继续在线时长的方法及装置，利用历史生命周期样本数据，建立能够表示所述对等体的继续在线时长的概率分布的多维高斯混合模型，并基于这样的多维高斯混合模型预测 Peer 的继续在线时长。

## 附图说明

- [0025] 图 1 是一种高斯混合模型的示意图；
- [0026] 图 2 是一种三维高斯混合模型的示意图；
- [0027] 图 3 是本发明实施例中预测 Peer 生命周期的方法流程图；
- [0028] 图 4 是本发明实施例中多维高斯混合模型的算法示意图；
- [0029] 图 5 是本发明实施例中多维高斯混合模型计算过程示意图；
- [0030] 图 6 是本发明实施例中预测对等网络中对等体的继续在线时长的装置。
- [0031] 具体实施例：
- [0032] 下面将结合附图对本发明实施例的技术方案作进一步详细描述。
- [0033] 现有的 Peer 生命周期预测结果之所以难以体现 Peer 的实际生命周期，是因为现有技术采用幂率分布模型预测生命周期时，只考虑了 Peer 在线时的  $\Delta t_{alive}$  和  $\Delta t_{since}$  对预测结果的影响，即体现 Peer 当前的在线状态对预测结果的影响，而在线状态未必就是影响 Peer 生命周期的重要因素。实际上，由于用户上网行为通常呈现出一定的用户习惯，简单举例，如用户通常上午在线时间集中在九点到 10 点之间，晚上的上网时间通常集中在 20 点到 22 点之间，因此，Peer 的历史在线时间应可作为影响其生命周期的预测结果的重要影响因素，用于预测 Peer 的生命周期。
- [0034] 进一步说明，现有技术中通过当前已在线时长这一单一因素所遵循的概率分布，预测 Peer 的继续在线时长，欠缺对实际影响预测结果的准确率的各种可能因素的综合影响的考虑，因此，预测结果与实际结果偏离较大。由于真实的历史生命周期是在各种可能的因素的影响下所产生的，因此，本发明实施例中，根据历史生命周期的样本的在线数据（例如：上线时刻，在线时长等），从而统计出 Peer 的生命周期规律，并根据 Peer 的生命周期规律，预测出当前 Peer 的生命周期概率，以及可能继续存活的时长。
- [0035] 本发明实施例中，利用多维高斯混合模型，来描述 Peer 生命周期的概率分布，基

于建立的多维高斯混合模型,推算出 Peer 的继续在线时长,进而结合 Peer 已在线时长,推算出 Peer 的生命周期。

[0036] 先对高斯混合模型作简要说明。高斯混合模型是基于多个遵循高斯分布的高斯成分,以及每个高斯成分对应的权值,对多种高斯分布进行合成的概率分布模型。参见图 1,图 1 是一种高斯混合模型的示意图,该模型中的高斯成分有五个,每个高斯成分遵循对应的高斯分布,每个高斯分布由对应的高斯曲线所标示。高斯曲线对应的高斯分布函数如公式 (3) :

$$[0037] p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \text{ 公式 (3)}$$

[0038] 通常,方便起见,用  $N(\mu, \sigma^2)$  表示一种高斯分布,其中,  $\mu$  为  $x$  的均值,  $\sigma^2$  为  $x$  与均值的差方,对于一个高斯模型的建立来讲,若  $\mu$  和  $\sigma^2$  已知了,则该高斯模型也就被建立了。图 1 所示模型中,五个高斯分布分别是  $N(0, 0.1)$ 、 $N(2, 1)$ 、 $N(3, 0.5)$ 、 $N(3.5, 0.1)$  和  $N(4, 1)$ 。在对五个高斯分布进行合成时,考虑各高斯成分对合成结果的影响所占权重的不同,将五个拟合成图 1 所示的一条混合高斯曲线 Mixture。

[0039] 二维高斯曲线对应的高斯分布函数如公式 (4) :

$$[0040] f(x) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \exp\left\{-\frac{1}{2(1-r^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2r\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\} \text{ 公 式 (4)}$$

[0041] 在函数中,由  $N$  个多元数组来表示  $(\mu_x, \sigma_x, \mu_y, \sigma_y, r)$ ,其中  $(j = 1, 2, \dots, N)$ ,  $r$  表示二维高斯成分之间的相关性系数,  $|r| < 1$ 。在本发明中,令二元高斯分布随机变量  $\xi$ 、 $\eta$ ,其均值为

$$[0042] E\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} = \mu$$

[0043] 协方差矩阵为

$$[0044] B = \begin{pmatrix} \sigma_x^2 & r\sigma_x\sigma_y \\ r\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

[0045] 二维高斯分布可以表示为  $N(\mu, B)$ 。

[0046] 同理,  $d$  维高斯分布函数如公式 (5) 所示:

$$[0047] f(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\left[(x-\mu)^T \Sigma^{-1} (x-\mu)\right]\right\} \text{ 公式 (5)}$$

[0048] 最终由  $M$  个节点生命周期历史记录所构造形成的三维高斯混合模型如图 2 所示,在一定的统计时间内将会出现多个峰值相互叠加的情况。

[0049] 本发明实施例中,通过建立多维高斯混合模型来预测 Peer 的继续在线时长。

[0050] 参加图 3,图 3 是本发明实施例中预测 Peer 生命周期的方法流程图,该流程可包括以下步骤:

[0051] 步骤 201、获取对等体的  $M$  个历史生命周期源数据  $P_i, i = 1, \dots, M$ ,该历史生命周期样本数据可以包括对等体节点的历史上线的时刻和历史在线的时长。

[0052] 可以先获取对等体的 M 个历史生命周期源数据  $P_i$ ,  $i = 1, \dots, M$ 。该源数据  $P_i$  可以包括对等体节点每次的上线时刻和在线时长(如节点 A, 1 月 4 号 20 点整上线, 在线时间 2 小时; 1 月 5 日 10 点 20 上线, 在线时间 3 小时等)。

[0053] 本发明实施例中,为保证用户的隐私,历史生命周期样本数据可由 Peer 自身记录并保存。同时,也可以通过让 Peer 节点发送历史生命周期样本数据到某一中心服务器的方法来集中获取。

[0054] 步骤 202、根据采样精度对  $P_i$  进行采样后得到历史生命周期样本数据  $S_i$ , 该  $S_i$  可以包括对等体节点的历史在线的时刻和历史在线的时长,其中,对等体节点的历史在线时长数据包括,节点某一次上线时间点到下线时间点所经历的总时间长度。历史在线时长的单位可以为分钟、秒或者小时等,单位越小,数据越精确。

[0055] 另外,也可以采用步骤 20 替换步骤 201 和 202,步骤 20,直接对 Peer 节点进行采样,获取对等体的样本数据  $S_i$ 。

[0056] 本发明实施例中,为保证用户的隐私,历史生命周期源数据可由 Peer 自身记录并保存。也可以通过让 Peer 节点发送历史生命周期源数据到某一中心服务器的方法来集中获取。

[0057] 步骤 203、确定待建立的多维高斯混合模型中,高斯成分的个数 n, n 是大于或等于 2 的正整数;所述多维高斯混合模型表示所述对等体的继续在线时长的概率分布。

[0058] 实际应用中,可综合考虑所建成的多维高斯混合模型与生命周期实际概率分布的逼近程度,以及建立多维高斯混合模型这一过程的计算量,来确定 n 的取值。通常,n 越大,则建立模型时的运算量相对越大,但建立出来的多维高斯混合模型与实际概率分布较逼近。

[0059] 所述步骤 203 与步骤 201 和 202 之间没有顺序的先后,可以先做步骤 201 和 202,再做步骤 203,也可以先做步骤 203,再做步骤 201 和 202。

[0060] 步骤 204、利用  $S_i$  与 n,建立所述多维高斯混合模型。

[0061] 该步骤中,建立多维高斯混合模型的过程,即计算每个高斯成分对应的高斯分布参数  $\mu$ 、 $B$ ,以及多维高斯混合模型中,每个高斯成分所占权重的过程。本发明实施例中,取多维高斯分布参数为  $\mu$  和  $B$ 。

[0062] 将多维高斯混合模型建好后,可以发送出去,例如:可以由服务器建立模型,然后将模型发送给 Peer 应用,或者由 Peer 建立模型,然后将模型发送给服务器应用。将多维高斯混合模型发送时,可以将所述高斯混合模型的描述信息发出,其中,所谓高斯混合模型的描述信息也即上述混合模型参数,如  $w$ ,  $\mu$ ,  $B$ 。

[0063] 步骤 205、利用建立的多维高斯混合模型,计算所述对等体的继续在线时长。

[0064] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成,所述的程序可以存储于一计算机可读取存储介质中,所述的存储介质,如:ROM/RAM、磁碟、光盘等。

[0065] 基于算出的 Peer 的继续在线时长,基于当前该 Peer 已在线时长,预测出 Peer 本次存活的生命周期。进一步,P2P 网络系统可基于 Peer 在线时间的预测值,提前做好 Peer 下线准备,如可提前通知其他关联邻居节点刷新所维护的节点信息,有效避免其他关联邻居节点在搜索、路由过程中指向该节点的时刻刚好该节点离开的现象出现,从而,避免 Peer

下线随机性给网络系统造成的扰动,提高网络系统的抗扰动能力。系统通知其他关联邻居节点如,chord 网络中,指针表中包含该节点的节点标识 (ID) ;pastry 网络中路由表中包含该节点的 ID ;kademlia 网络中 K 桶中包含该节点的 ID ;等等。

[0066] 下面主要对上述步骤 204 中,如何建立高斯混合模型作进一步说明。

[0067] 参见图 4,图 4 是本发明实施例中高斯混合模型的算法示意图。参见图 5,图 5 是本发明实施例中高斯混合模型计算过程示意图。图 4 与图 5 中, S 表示历史生命周期样本数据序列, M 表示样本数据空间中有 M 个样本数据,  $\beta$  表示 n 个高斯分布, w 表示高斯成分的权重,Z 服从参数为  $\beta$  的高斯分布。其中,

[0068]  $S = (S_1, S_2, \dots, S_i, \dots, S_M)$ ;

[0069]  $\beta_j = N(\mu_j, B_j)$ ,  $j = 1, \dots, n$ ;

[0070]  $z_i \sim \text{Multinomial}(w)$ 。

[0071] 基于采集到的 M 个  $S_i$ ,可知道每个  $S_i$  在 M 中的出现次数,进而能够算出各  $S_i$  的出现概率  $s_i$ 。设  $s_i$  对应的概率分布为  $p(s_i | z_i, \beta)$ ,该概率分布即表示的是与实际概率分布相对应的需要建立的高斯混合模型。

[0072] 具体计算时,根据已知量  $s_i$  和 n,利用现有贝叶斯推理,计算多维高斯混合模型中,各多维高斯模型的高斯分布参数,即  $\beta_j = N(\mu_j, B_j)$ ,以及对应的 w。称要求解的  $\beta_j = N(\mu_j, B_j)$  与对应的 w 为混合模型参数。基于已知的  $s_i$  和 n,使用极大似然法估计 w 和  $\beta$ ,过程如下:设 w 和  $\beta$  的初始值为 0,

[0073] 第一步:建立包括有多维混合模型参数的似然函数 (likelihood function)

[0074]  $L(s; \theta), L(s; \theta) = \prod_{i=1}^M p(s_i | \theta), \theta = (w_j, \mu_j, B_j);$

[0075] 第二步:根据  $s_i$  求出  $L(s; \theta)$  达到极值时,混合模型参数的估计值。其中,因为似然函数  $L(s; \theta)$  与似然函数的对数  $\ln L(s; \theta)$ ,在同一参数  $\theta$  处获得最大值,为计算简便,通常对似然函数求对数来进行估计:

[0076]  $\ln L(s; \theta) = \sum_{i=1}^M \ln \left( \sum_{j=1}^N w_j N_{s_i}(\mu_j, B_j) \right)$ ,式一

[0077] 其中  $w_{ij} = p(w_i = j | \theta)$ ,且  $\sum_{j=1}^N p(w_i = j | \theta) = 1$ 。

[0078] 对式一采用 EM 估计:

[0079] E 步:利用从上一次 M 步估计获得的估计量  $\theta^{(k)} = (w^{(k)}, \mu^{(k)}, B^{(k)})$ ,可以求出参数 w 的后验概率。推导如下:

[0080]  $h_{ij}^{(k)} = p(w_i = j | s_i, \theta^{(k)}) = \frac{p(w_i = j, s_i | \theta^{(k)})}{p(s_i | \theta^{(k)})}$

[0081] 因为:  $p(s | w = j, \theta) = N(\mu_j, B_j)$ ,所以上式可写为:

[0082]  $h_{ij}^{(k)} = \frac{p(s_i | w_i = j, \theta^{(k)}) p(w_i = j | \theta^{(k)})}{\sum_{l=1}^N p(s_i | w_i = l, \theta^{(k)}) p(w_i = l | \theta^{(k)})} = \frac{w_{ij}^{(k)} N_{s_i}(\mu_j^{(k)}, B_j^{(k)})}{\sum_{l=1}^N w_{il}^{(k)} N_{s_i}(\mu_l^{(k)}, B_l^{(k)})}$

[0083] M 步:利用期望最大化,写出期望函数,求出使期望函数取得最大值的参数  $\theta$ 。求

最大值可以利用对似然函数求导取 0, 算出  $\theta$ 。即  $\frac{\partial L(\theta)}{\partial \theta} = 0$ 。在 M 步对似然函数求导的过程中, 可以通过一次对参数  $\theta = (\mu, B, p(w=j))$  求导。

[0084] 通过对  $\frac{\partial L(\theta)}{\partial \mu_j} = 0$  求解, 可以得出  $\mu_j^{(k+1)} = \frac{\sum_{i=1}^M h_{ij}^{(k)} s_i}{\sum_{i=1}^M h_{ij}^{(k)}}$ 。

[0085] 通过对  $\frac{\partial L(\theta)}{\partial B_j} = 0$  求解, 可以得出  $B_j^{(k+1)} = \frac{\sum_{i=1}^M h_{ij}^{(k)} (s_i - \mu_j^{(k)}) (s_i - \mu_j^{(k)})^T}{\sum_{i=1}^M h_{ij}^{(k)}}$ 。

[0086] 通过对  $\frac{\partial L(\theta)}{\partial p(w_i=j|\theta)} = 0$  求解, 可以得出  $w_{ij}^{(k+1)} = \frac{1}{M} \sum_{i=1}^M h_{ij}^{(k)}$ 。

[0087] 通过上述计算过程计算出  $\theta$  参数后, 混合模型参数也就确定, 相应地, 高斯混合模型也就确定下来。上述对混合模型参数的估算基于 EM 算法进行, 实际应用中, 也可采用变分法估算混合模型参数。

[0088] 之后, 保存算出的混合模型参数, Peer 可以利用建立的高斯混合模型, 预测该 Peer 的继续在线时长, 即估计  $S_{M+1}$ , 过程如下:

[0089] a、根据保存的各混合模型参数, 可算出 Peer 的生命周期的出现概率密度为:

[0090]  $p(s_i) = \sum_{j=1}^N w_{ij} N(\mu_j, B_j)$ 。

[0091] b、记 Peer 的继续存活  $y$  时长的概率为  $Q(y)$ , 通过  $Q(y) = p(s > (t+y) | s > t, \delta)$ , 则,

[0092]  $Q(y) = p(s > (t+y) | s > t, \delta) = \frac{p(s > (t+y) | \delta)}{p(s > t | \delta)} = \frac{p(s > (t+y) \cap \delta)}{p(s > t \cap \delta)} = \frac{\int_{t+y}^{\infty} p(s) ds}{\int_{t}^{\infty} p(s) ds}$  式

二

[0093] 其中,  $t = \Delta t_{alive}$  为给定值,  $y = \Delta t_{since}$  为变量,  $\delta$  为当前的上线时间区间值。分母为某一定值, 分子为  $y$  的表达式。

[0094] c、求 Peer 的继续在线时长的期望值  $E[y]$ , 则,  $E(y) = \int y Q(y) dy$ 。

[0095] d、基于预测出的继续在线时长, 预测出 Peer 的生命周期  $T$ :

[0096]  $T = t + E[y] = t + \int y Q(y) dy$ 。

[0097] 至此, 基于建立的多维高斯混合模型, 预测出 Peer 的生命周期。

[0098] 在  $s_i$  的内容中可以包括区间信息, 区间是表示取样的时间段, 区间的长度, 可以为一个小时, 或者一天, 或者半天, 或者一周, 或者一个月。可以分别以周一, 周二, 周三, 周四, 周五, 周六, 周日作为区间, 也可以将周一至周五作为一个区间, 将周六至周日作为一个区间等, 区间可以分为多种, 将属于同一个区间的多个样本数据进行统计, 可以得出在该区间上的上线规律, 并作为高斯模型的一个输入维度加入到模型参数计算中。

[0099] 另外, 可以针对每种区间分别建立高斯模型。在一个星期内, 可以分别以周一, 周二, 周三, 周四... 周日, 作为一个区间, 这样需要建 7 个高斯模型分别统计每个区间内的上

线规律。在一个星期内,也可以以周一至周五作为一个区间,将周六至周日作为一个区间,这样需要建 2 个高斯模型分别统计每个区间内的上线规律。

[0100] 每个区间内可以按照采样精度划分刻度,可以每 15 分钟划分为一个刻度,或者每半个小时划分为一个刻度等等,通过统计在每个刻度上的上线信息(例如:是否上线,上线时长),从而得出在该区间上的上线规律。

[0101]  $S_i$  的内容可以包括:在线时长信息,上线的时刻信息,这样,建立的混合高斯模型,就有三个维度,包括:在线时长,上线的时刻,和概率分布。如果将多个区间的信息汇集在一起,就会多一个维度,就是多个区间形成的维度,这样, $S_i$  的内容还可以包括:在线时长信息,上线的时刻信息,上线的区间信息。这样,建立的混合高斯模型,就有四个维度,包括:在线时长,上线的时刻,上线的区间,和概率分布。

[0102] 对于一周内的概率统计,可以将周一至周日作为一个区间, $S_i$  包括在线时长和上线的时刻两个维度。也可以将一周划分为几个区间(例如分别将周一,周二,周三,周四,周五...周日作为一个区间), $S_i$  包括在线时长,上线的时刻,以及上线的区间,这样,就有三个维度。

[0103]  $S_i$  的内容举例如下:

[0104] { 在线时长 (min), 上线时刻 (hour:minute), 上线区间 (week) }

[0105] = {120, 20:15, 5}, {240, 14:00, 6}.....{50, 21:30, 1}

[0106] 或者

[0107] { 在线时长 (min), 上线时刻 (hour:minute), 上线区间 (week) }

[0108] = {120, 20:15, 周末}, {240, 14:00, 周末}.....{50, 21:30, 工作日}

[0109] 或者

[0110] { 在线时长 (min), 上线时刻 (week:hour:minute) }

[0111] = {120, 5:20:15}, {240, 6:14:00}.....{50, 1:21:30}

[0112] 对应于上述本发明实施例中预测 Peer 的继续在线时长的方案,本发明实施例还提供一种预测对等网络中对等体的继续在线时长的装置,所述装置基于前面所述的方法实现,参见图 6,图 6 是该装置的结构示意图,该装置可设置于每个 Peer 上或者服务器上或者其他通信设备上,用于预测该 Peer 每次上线的继续在线时长,包括:获取单元、接收单元、模型建立单元和预测单元,其中,

[0113] 获取单元,获取对等体的 M 个历史生命周期样本数据  $S_i$ ,  $i = 1, \dots, M$ ;

[0114] 所述获取单元可以具体包括:

[0115] 第一单元,获取对等体的 M 个历史生命周期源数据  $P_i$ ,  $i = 1, \dots, M$ ;

[0116] 第二单元,根据采样精度,对历史生命周期源数据  $P_i$  进行采样得到  $S_i$ ;

[0117] 接收单元,接收待建立的多维高斯混合模型中,高斯成分的个数 n, n 是大于或等于 2 的正整数;所述多维高斯混合模型表示所述对等体的继续在线时长的概率分布。

[0118] 模型建立单元,利用采样单元获取的  $S_i$ ,以及高斯成分的个数 n,建立多维高斯混合模型,n 是大于或等于 2 的正整数;所述高斯混合模型表示所述对等体的继续在线时长的概率分布;

[0119] 预测单元,利用模型建立单元建立的所述多维高斯混合模型,预测所述对等体的继续在线时长。

- [0120] 另外,如果高斯成分的个数 n 保存在模型建立单元中,则不需要接收单元。
- [0121] 模型建立单元包括:
- [0122] 参数计算单元,计算所述多维高斯混合模型的混合模型参数;
- [0123] 所述多维混合模型参数包括:
- [0124] 每个高斯成分对应的高斯分布参数,以及所述多维高斯混合模型中,每个高斯成分所占权重。
- [0125] 参数计算单元包括:
- [0126] 概率计算单元,根据采样得到的 M 个  $S_i$ ,算出各  $S_i$  的出现概率  $s_i$ ;
- [0127] 函数构建单元,建立包括所述多维混合模型参数的似然函数;
- [0128] 估算单元,利用所述概率计算单元算出的  $s_i$ ,计算所述函数构建单元构建的所述似然函数取最大值时,各多维混合模型参数的估算值。
- [0129] 预测单元包括:
- [0130] 概率密度计算单元,计算所述对等体生命周期的出现概率密度;
- [0131] 存活概率计算单元,利用概率密度计算单元算出的所述出现概率密度,计算所述对等体继续存活预设变化时长的概率;
- [0132] 时长计算单元,利用所述概率密度计算单元算出的概率,求取所述对等体的继续在线时长。
- [0133] 该装置进一步包括:生命周期预测单元,根据所述预测单元预测出的所述继续在线时长,以及所述对等体已在线时长,算出所述对等体本次存活的生命周期或某一继续在线时长可能出现的概率。
- [0134] 本发明实施例还提供了一种建模装置,该装置可包括上述获取单元,接收单元和模型建立单元,模型建立单元可以是上述参数计算单元,包含上述函数构建单元和估算单元;该建模装置进一步包括:
- [0135] 发送单元,将模型建立单元建立的所述高斯混合模型的描述信息发送。
- [0136] 其中,所谓高斯混合模型的描述信息也即上述混合模型参数,如  $w$ ,  $\mu$ ,  $B$ 。
- [0137] 综上所述,本发明实施例提供的预测对等网络中对等体的继续在线时长的方法及装置,利用历史生命周期样本数据,建立能够表示所述对等体的继续在线时长的概率分布的多维高斯混合模型,该多维高斯混合模型中,不是基于受单一条件因素影响下 Peer 的继续在线时长所遵循的概率分布,预测 Peer 的继续在线时长,而是基于 Peer 的继续在线时长受多种条件因素影响,综合考虑多个条件因素影响下继续在线时长分别遵循的概率分布,合成最终的多维高斯混合模型,并基于这样的多维高斯混合模型预测 Peer 的继续在线时长,使预测结果逼近 Peer 的实际继续在线时长。

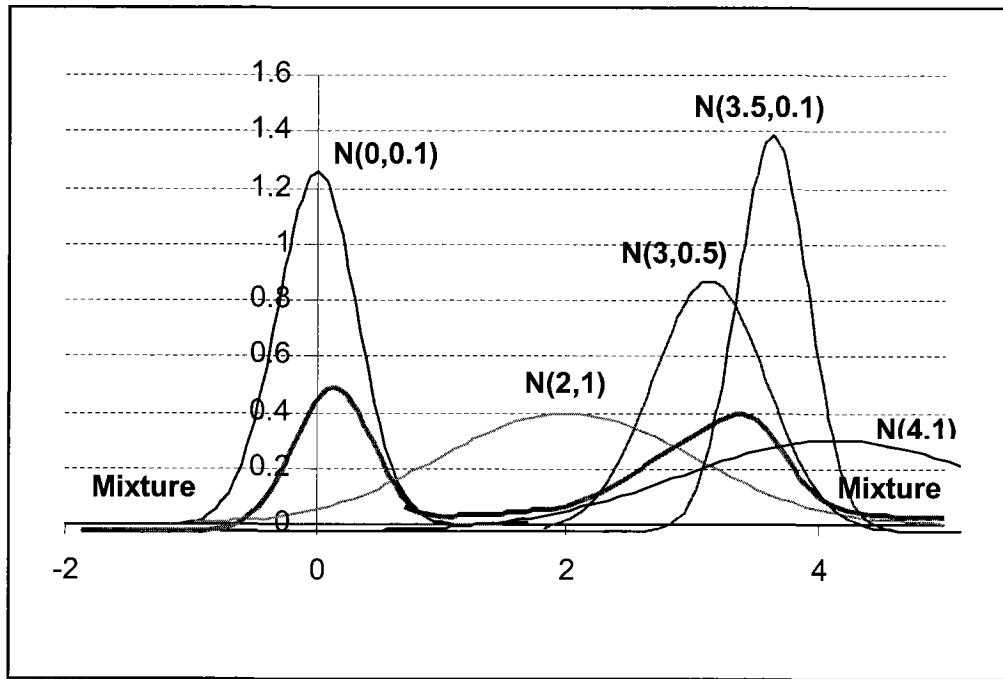


图 1

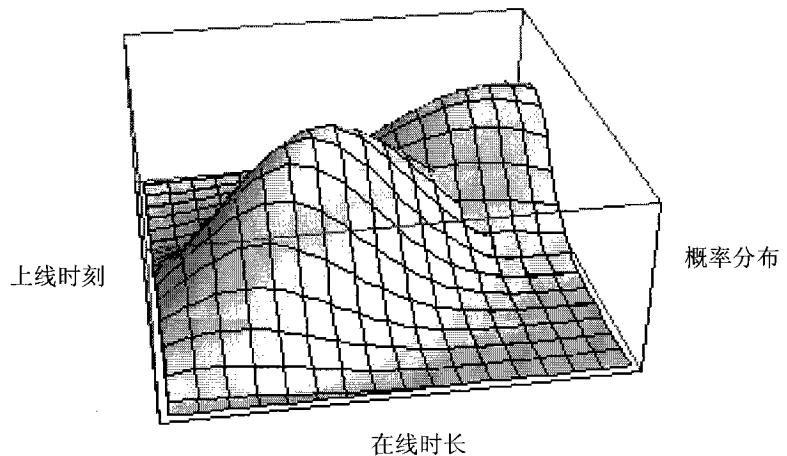


图 2

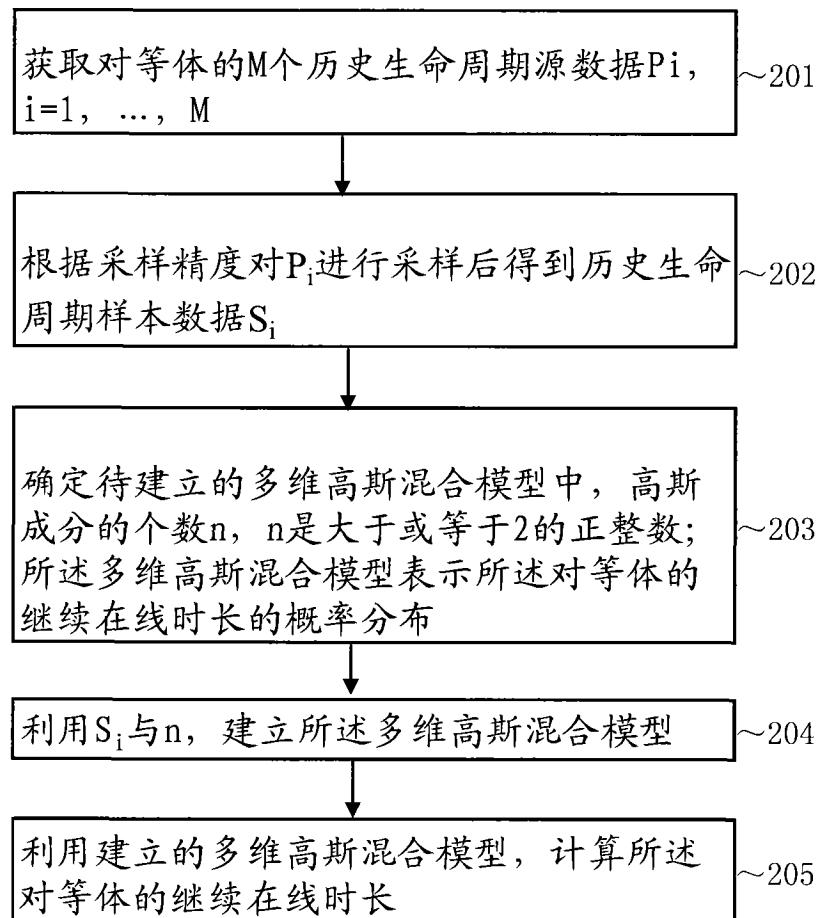


图 3

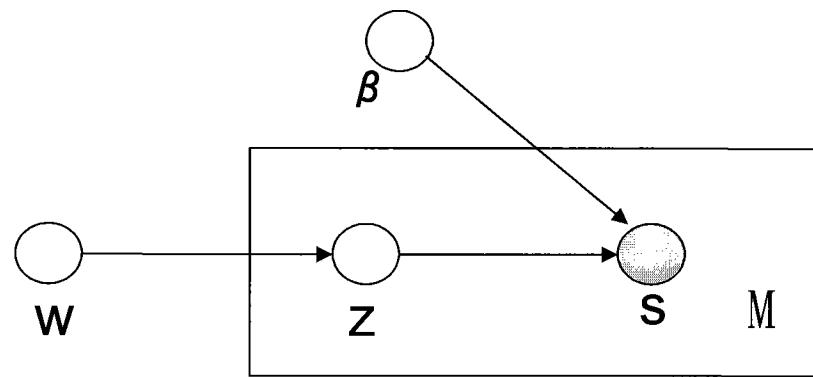


图 4

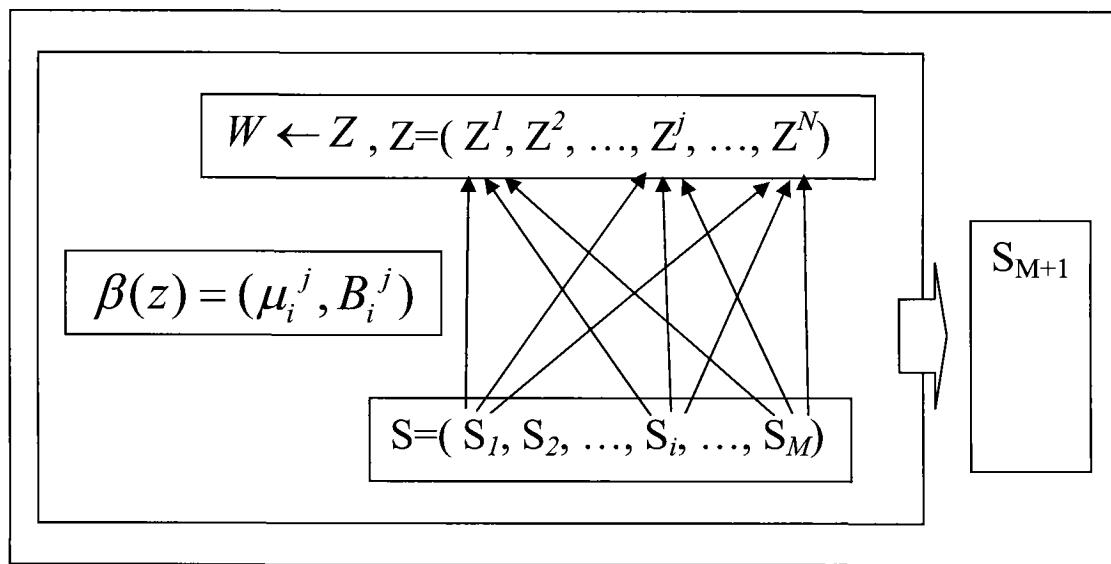


图 5

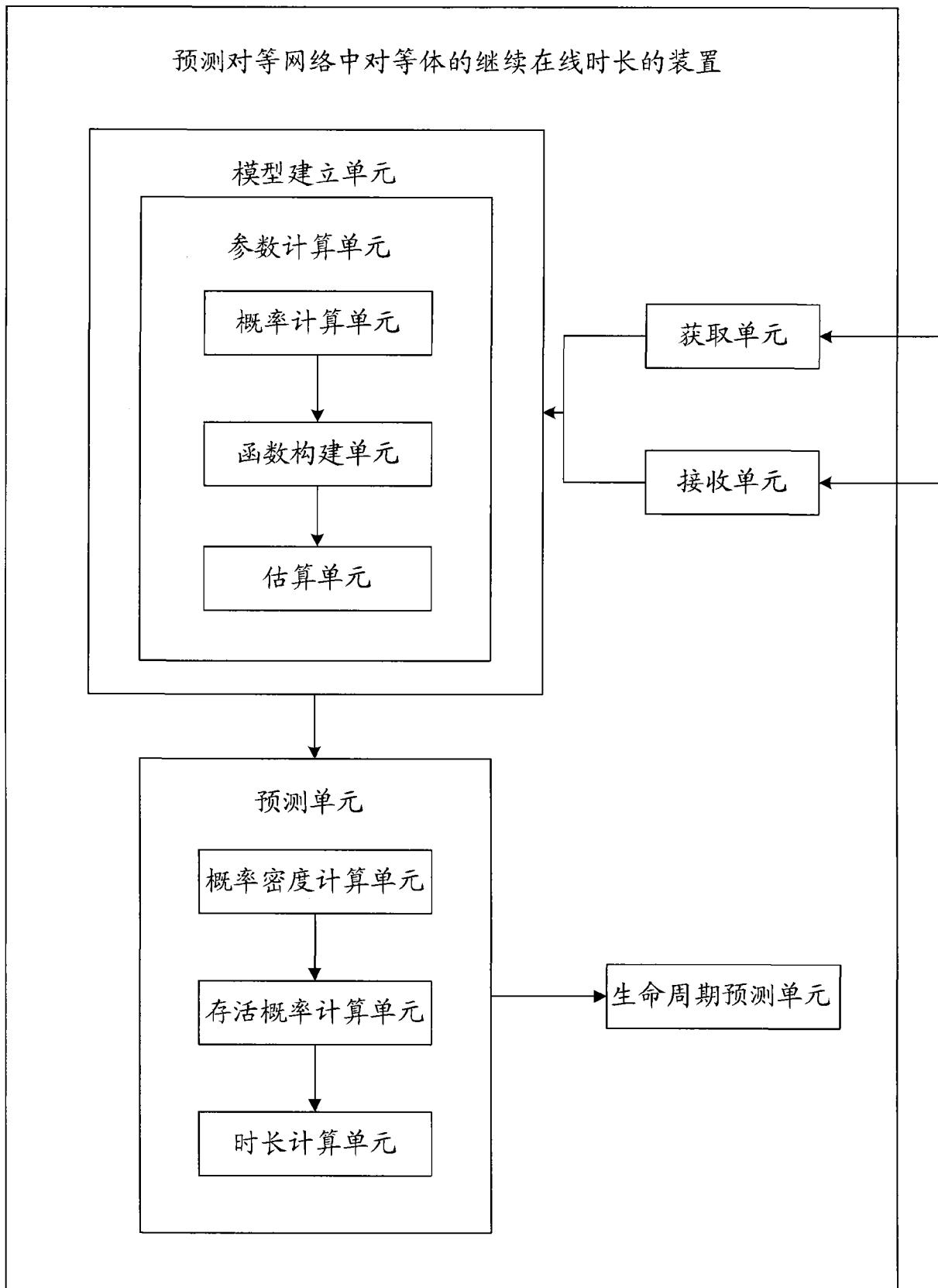


图 6