

自适应累加型失效检测模型研究

石磊^{1*}, 陈文远¹, 陶永才¹, 卫琳²

(1. 郑州大学 信息工程学院, 郑州 450001; 2. 郑州大学 软件技术学院, 郑州 450002)

(* 通信作者电子邮箱 shilei@zzu.edu.cn)

摘要:传统失效检测输出二值信息分别代表信任或怀疑,然而该机制存在灵活性不足的问题。针对这种问题,累加型失效检测以怀疑级别为输出,能适应同时运行的不同进程的 QoS 需求。在分析和研究已有失效检测模型和累加型失效检测算法的基础上,提出一种新的累加型失效检测模型——EXP-ACC-FD。该模型利用幂律计算出心跳间隔的加权平均值,将该均值和距上次心跳到达的时间代入指数分布函数,从而计算出被监测进程的怀疑级别。实验分析表明,在相同的检测时间内,EXP-ACC-FD 准确性高于 NFD-E 失效检测模型和 PHI 失效检测模型。

关键词:失效检测; 自适应; 怀疑级别; 指数分布; 幂律

中图分类号: TP301.6; TP393.0 **文献标志码:** A

Adaptive accrual failure detection model

SHI Lei^{1*}, CHEN Wen-yuan¹, TAO Yong-cai¹, WEI Lin²

(1. School of Information Engineering, Zhengzhou University, Zhengzhou Henan 450001, China;

2. School of Software Technology, Zhengzhou University, Zhengzhou Henan 450002, China)

Abstract: The output binary information of traditional failure detection represents trust or suspicion respectively. However, the mechanism is lack of flexibility. For this problem, cumulative-type failure detection, taking the level of suspicion as output, can adapt to the QoS requirements of different processes running simultaneously. An accrual failure detection model named EXP-ACC-FD was proposed on the basis of analyzing and researching the existing failure detection models and accrual failure detection algorithms. It calculates the weighted mean of heartbeat inter-arrivals with power law and substitutes the weighted mean and the time between last heartbeat coming and current time into exponential distribution to get suspicion level of monitored process. The simulation analyses show that the accuracy of EXP-ACC-FD is higher than NFD-E and PHI within the same detection time.

Key words: failure detection; adaptive; suspicion level; exponential distribution; power law

0 引言

失效检测指的是检测者利用失效检测模式及失效检测算法对被检测对象进行检测,从而判断被检测者是否失效的过程。分布式系统中,失效检测器是失效恢复、动态重启、可靠性通信、集群管理等功能的基础^[1],在无线网络、云计算等领域也有重要的应用^[2-4]。出现失效现象之后,需要有效的失效检测以及数据修复技术来延缓系统的平均数据丢失时间。然而,实际系统中很难在系统发生失效后立即进行数据修复,简单的做法是系统中的存储节点定期互相发送心跳信息以确定是否出现设备失效^[5-6]。

传统的失效检测(如 Chen^[7])输出一个二进制信息,用 1 和 0 分别代表信任和怀疑。在较为复杂的应用条件下,这种输出机制存在灵活性不足的问题^[8]。针对这种缺点,Defago 等^[9]提出一种新的失效检测模型,即累加型失效检测模型。这种失效检测对每一个进程都设置一个对应的值作为输出,这个值并不简单地判定信任或怀疑,而是代表对被检测进程的怀疑级别。监测进程可以根据不同的怀疑级别采取不同的措施,而不是等进程失效后再做出决定,从而提高了系统的灵活性。此外,传统的失效检测器将失效检测和失效后采取的措施做成一个应用,而累加型失效检测器利用其特殊的输出

机制,使失效检测可以成为一种服务,不同的进程可以对怀疑级别做出不同的解读。

Hayashibara 等^[8]提出的 φ 失效检测模型假设心跳到达间隔服从正态分布,根据正态分布函数计算出时间 t 之前心跳到达的概率,作为 t 时刻的怀疑级别,然后与设定的阈值比较来判定进程是否失效。然而,正态分布并不是对心跳间隔分布的理想近似。指数分布是一种连续概率分布,可以用来表示事件发生的时间间隔,比如旅客进机场的时间间隔等。研究表明指数分布在网络运算中具有广泛的应用^[10-12]。本文提出一种基于指数分布的自适应的累加型失效检测模型,称之为 EXP-ACC-FD,采用的与心跳间隔时间对应的指数分布函数如式(1)所示。

$$F(x) = \begin{cases} 1 - \exp\left(-\frac{x}{\mu}\right), & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (1)$$

由于较近一段时间的心跳间隔更能真实地反映当前网络状况,因此在计算均值 μ 时较近一段时间的心跳间隔应赋予较高的权值。研究表明,许多网络特征都服从幂律分布^[13]。因此,本文提出的 EXP-ACC-FD 中,窗口中的心跳间隔权值的分配采用了幂律计算加权平均值。

收稿日期: 2011-08-29; 修回日期: 2011-11-15。 基金项目: 河南省自然科学基金资助项目(2011B520035)。

作者简介: 石磊(1967-),男,河南郑州人,教授,博士,CCF 会员,主要研究方向: 高性能计算、Web 挖掘; 陈文远(1986-),男,河南开封人,硕士研究生,主要研究方向: 网络信息; 陶永才(1976-),男,河南武陟人,讲师,博士,主要研究方向: 网络计算; 卫琳(1968-),女,河南郑州人,副教授,主要研究方向: Web 挖掘。

1 系统模型

失效检测模型可以描述为一个四元组 $\{P, T, a, O\}$, 其中:

P 表示进程的集合 $\{p_1, p_2, \dots, p_n\}$;

T 表示全局时钟, 该时钟为虚拟设备^[14];

$a: P \times T \rightarrow O$ 表示由失效检测算法决定的映射;

O 表示输出的集合, 令 $t \in T, o_t \in O$, 有 $o_t = \{o_{t-p_1}, o_{t-p_2}, \dots, o_{t-p_n}\}$, 其中 o_{t-p_i} 表示时刻 t , 进程 p_i 的输出。

1.1 进程通信与心跳采样

异步分布式系统由一系列进程 $\{p_1, p_2, \dots, p_n\}$ 构成, 每个进程上运行着一个失效检测模块, 任意进程之间可以通过网络通信, 并周期性地发送心跳信息。每个失效检测模块设立一张系统进程表, 每个表项包括一个进程号和一个滑动窗口。心跳间隔信息就储存在滑动窗口中。

最新的心跳到达时, 计算出当前时刻与之前一个心跳到达时刻的时间间隔, 并将这个时间间隔存入滑动窗口。如果窗口已被填满, 则会同时将窗口中最旧的心跳间隔移出窗口。随后, 采取加权平均的方法计算出窗口中心跳间隔的均值。在分配权值时, 窗口填入新数据的时间与当前时间越近, 所赋权值越高。

权值分配方法如下: 设滑动窗口中储存的最后一个心跳间隔, 即最新计算出的心跳间隔对应的权值为 P_1 , 之前计算出的心跳间隔对应的权值为 P_2 , 再之前计算出的心跳间隔对应的权值为 P_3 , 以此类推, 且有 $\sum_{i=1}^w P_i = 1$, 其中 w 为窗口中心跳间隔的数量。根据幂律, 存在以下关系:

$$P_i = K \cdot i^{-m}$$

已有的研究表明, 参数 m 接近于 1, 为方便起见, 令 $m = 1$, 则 $P_i = K/i$ 。由于

$$\sum_{i=1}^w K/i = K \sum_{i=1}^w 1/i \approx K(\ln w + r) = 1$$

故 $K \approx 1/(\ln w + r)$

其中 r 为欧拉常数。据此, 窗口中每个心跳间隔对应的权值可被计算出来。

1.2 算法描述

EXP-ACC-FD 的监测进程为每一个进程设立一个怀疑级别 sl 值。假设心跳间隔服从指数分布, sl 可由 t 时刻心跳到达的概率表示。如果在 t 时刻心跳未到达, 那么此时对应的到达概率越高, 怀疑级别就会越大。 sl 可由式(2)得出:

$$sl(t) = 1 - \exp\left(-\frac{t-t_0}{\mu}\right) \quad (2)$$

其中: t 为当前时间, t_0 为上次心跳到达时间, μ 为滑动窗口中心跳间隔的加权平均。假定窗口中心跳间隔和之后到来的心跳间隔服从同一概率分布, 即参数(均值)相同, 因此 μ 在计算怀疑级别时作为参数使用。 $sl(t)$ 表示对进程的怀疑级别, $sl(t)$ 越大怀疑级别越高。当 $sl(t) = 0$ 时, 表示对应进程正常; 当 $sl(t) = 1$ 时, 表示对应进程已经失效。

对于两个进程 p 和 q , 如果进程 q 想知道进程 p 是否失效, 就向自己的检测模块发送一个信号; 检测模块在收到信号后计算 p 的怀疑级别 sl , 并将 sl 的值返回给进程 q ; 进程 q 将 sl 的值与阈值比较, 判定 p 是否失效或应采取哪些措施。

1.3 失效检测服务

失效检测服务需要具备 3 个特点: 可扩展性、灵活性和自适应性^[15]。EXP-ACC-FD 能够作为一种失效检测服务来使用。

可扩展性体现在新的进程加入系统时, 对系统性能影响较小。新加入的进程向需要通信的进程发送心跳, 接收到心跳的进程与新进程将对方的信息添加到自己的进程列表中。新进程对系统性能的影响主要依赖于当前的网络状况, 而非系统本身。

针对不同的 QoS 需求, 不同的进程对失效检测的要求不尽相同。有些进程要求失效检测模型能够快速检测出失效, 对检测的准确性要求较低; 而有些进程要求失效检测模型具有尽可能高的准确性, 不要求其能快速检测出失效。对于这些不同的服务需求, 可以对不同的进程设定不同的阈值。由式(2)可知, 设置较大的阈值意味着检测时间较长且准确率较高, 而设置较小的阈值意味着检测时间较短且准确率较低。因此, 通过为不同的进程设置合适的阈值, EXP-ACC-FD 能满足同时运行的不同进程的 QoS 需求。

EXP-ACC-FD 是一种自适应的失效检测模型, 能够适应变化的网络环境。随着网络状况的变化, 窗口中的心跳间隔随之改变, 使得均值 μ 不断调整, 达到设定的怀疑级别所需的时间也与之前不同。 μ 能够对网络状况的变化及时做出反应。当心跳间隔变大时, μ 值也随之增加, 这样达到怀疑级别所需的时间也会增加, 使得失效检测模型不会在较长时间没有心跳到达的情况下判定进程失效。

1.4 算法失效级别证明

传统的失效检测模型根据限制错误的程度可以分为 8 类。一般认为, 实际应用的失效检测模型应达到 $\diamond P$ 类的要求, 即具有强完整性和最终强准确性的特点。Défago 等给出了关于累加型失效检测模型特有的分类。例如 $\diamond P_{AC}$ 表示具有累加性和上界性的累加型失效检测模型。其中累加性是指所有正确进程对所有失效进程的怀疑级别最终会以一个正速率单调递增, 上界性是指所有正确的进程对其他正确进程的怀疑级别存在上界。Défago 等已经给出证明, $\diamond P_{AC}$ 和 $\diamond P$ 具有同样的效力^[9]。欲证明 EXP-ACC-FD 满足 $\diamond P_{AC}$ 类, 只需证明其满足性质 1 和性质 2。

性质 1 累加性。如果进程 p 失效, 那么其他监测进程关于 p 的怀疑级别最终会以一个正速率单调递增。

证明 设最后一次收到 p 的心跳的时刻为 0, 此后无心跳到达。用 $sl(t)$ 表示监测进程在 t 时刻对进程 p 的怀疑级别。

设有两时刻 t_1, t_2 , 其中 $t_2 > t_1$, 则:

$$sl(t_2) - sl(t_1) = \left(1 - \exp\left(-\frac{t_2}{\mu}\right)\right) - \left(1 - \exp\left(-\frac{t_1}{\mu}\right)\right) = \exp\left(-\frac{t_2}{\mu}\right) \left(\exp\left(-\frac{t_1 - t_2}{\mu}\right) - 1\right)$$

由于 $t_1 - t_2 < 0$, 则:

$$\exp\left(-\frac{t_1 - t_2}{\mu}\right) > 1$$

$$\exp\left(-\frac{t_1 - t_2}{\mu}\right) - 1 > 0$$

而 $\exp\left(-\frac{t_2}{\mu}\right) > 0$, 故 $sl(t_2) - sl(t_1) > 0$ 。

因 $t_2 > t_1$, 可知 $sl(t)$ 为单调递增函数。证毕。

性质 2 上界性。如果进程 p 正常, 则其他监测进程关于 p 的怀疑级别有上界。

证明 如果进程 p 正常, 则 p 会持续向其他进程发送心跳信息, 其他进程会在有限的时间内收到 p 的心跳信息。设上一次心跳到达时刻为 0, 则必然存在时刻 t_a , 在该时刻 p 的心跳信息到达监测进程。用 $sl(t)$ 表示监测进程在 t 时刻对进程 p

的怀疑级别, $sl(t)$ 为单调递增函数。在任意时刻 $t \in (0, t_a)$, 必有 $sl(t) < sl(t_a)$ 。故 $sl(t_a)$ 为监测进程关于 p 的怀疑级别的上界。证毕。

由于 EXP-ACC-FD 满足性质 1 和性质 2, 故 EXP-ACC-FD 满足 $\diamond P_{AC}$ 类。

2 实验结果与分析

实验测量了 EXP-ACC-FD 阈值变化, 窗口变化对检测时间及准确性的影响。之后, EXP-ACC-FD 与 φ 失效检测模型及 Chen 的 NFD-E 失效检测模型在同等条件下进行性能对比。实验中 EXP-ACC-FD 的相关结果均为判定进程彻底失效时的数据, 不考虑采取其他措施的情况。

2.1 阈值变化对 EXP-ACC-FD 的影响

通过测试阈值改变时错误率、查询准确率及检测时间的变化, 可以得到阈值变化对本文所讨论 EXP-ACC-FD 的影响。在窗口大小为 1000 时, 阈值对 EXP-ACC-FD 错误率、查询准确率及检测时间的影响如图 1~3 所示。

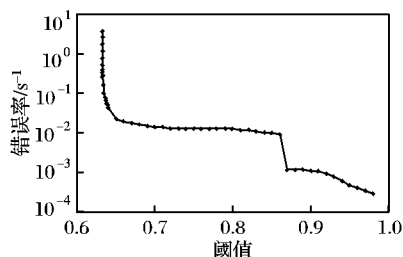


图1 阈值与错误率的关系

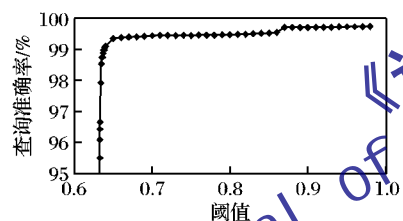


图2 阈值与查询准确率的关系

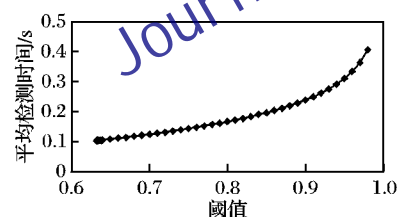


图3 阈值与检测时间的关系

从图1~3可知, 随着阈值的增加, 错误率逐渐降低, 查询准确率逐渐升高, 且检测时间越来越长, 结果与理论分析相同。阈值在0.63附近的变化对错误率及查询准确率的影响非常大, 而此时对应的检测时间恰巧在心跳间隔均值附近。当阈值在0.87附近时, 错误率和查询准确率变化较为剧烈, 而检测时间变化相对平稳。这是由于此时的阈值对应的检测时间在两倍均值左右, 即使出现一次丢包, 也会在要求的检测时间内收到心跳信息。

2.2 窗口大小对 EXP-ACC-FD 的影响

实验测试了窗口大小对 EXP-ACC-FD 错误率、查询准确率以及检测时间的影响。窗口的大小从 110 到 10000 进行变化, 选取阈值 0.64 做实验。实验结果如图 4~6 所示。

可以看出, EXP-ACC-FD 的错误率、查询准确率和检测时间变化随窗口大小改变并不明显。由于与当前时间较近的心跳间隔记录被赋予了较高的权值, 因此窗口大小对 EXP-

ACC-FD 性能影响不大。

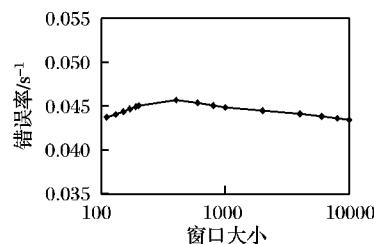


图4 窗口大小与错误率的关系

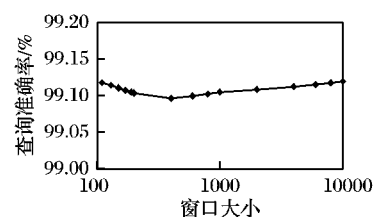


图5 窗口大小与查询准确率的关系

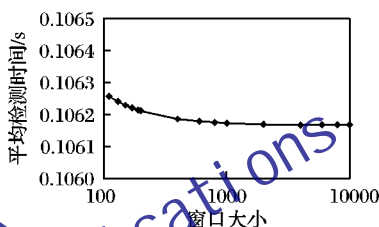


图6 窗口大小与检测时间的关系

2.3 EXP-ACC-FD 与其他失效检测模型对比

EXP-ACC-FD 分别与 φ 失效检测模型和 Chen 的 EFD-E 失效检测模型进行对比。根据测试所得到的数据, 可以描绘出 3 种失效检测模型在窗口大小 1000 的情况下错误率, 查询准确率随检测时间变化的曲线。实验结果如图 7~8 所示。

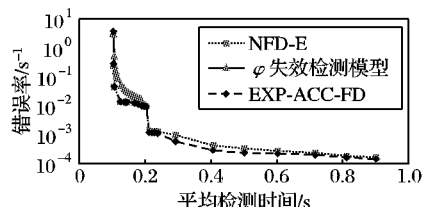


图7 错误率对比

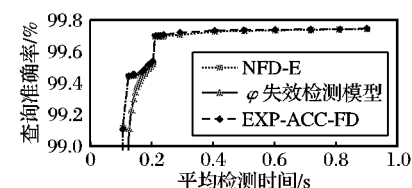


图8 查询准确率对比

从图7~8中可以看出, EXP-ACC-FD 在同样的检测时间内, 无论在错误率和查询准确率方面均优于 Chen 的 NFD-E 失效检测模型和 φ 失效检测模型。在检测时间在 0.21 s 左右时, 几种失效检测模型的准确性有跳跃式的提升, 这是由于此时检测时间大约等于两倍均值, 即使丢包一次也能在要求的检测时间内收到心跳信息。而与 Chen 的 NFD-E 或其他传统失效检测模型相比, 作为一种累加型失效检测模型, EXP-ACC-FD 还具有较高的灵活性。

3 结语

基于指数分布和幂率计算心跳间隔权值的方法, 本文提出了一种累加型失效检测模型 EXP-ACC-FD, 实验与分析表

(下转第 624 页)

3 结语

在具有数千节点的云存储系统中,如何将数据对象均衡、高效地放置到系统中,并能保证数据的高可靠性是一个重大挑战。本文提出的 RPGCH 算法摒弃了记录数据对象映射信息的做法,通过改进的一致性哈希算法获得副本与存储节点之间的映射信息。该算法支持存储节点分组,在每个分组内只存放同一数据对象的一个副本,很大程度上提高了数据的可靠性。此外,该算法分配到每个节点的副本数量与该节点的服务能力成正比,并且能自适应节点数量变化。因此,该算法很适合节点数据庞大的云存储系统,是云存储系统副本布局的一个良好选择。

参考文献:

- [1] GRAY J. What next? A few remaining problems in information technology[EB/OL]. [2011-05-10]. http://research.microsoft.com/~gray/talks/Gray_Turing_FCRC.pdf.
- [2] 云存储[EB/OL]. [2011-05-10]. <http://baike.baidu.com/view/2044736.htm>.
- [3] GHEMAWAT S, GOBIOFF H. The Google file system [C]// Proceedings of the 19th ACM Symposium on Operating Systems Principles. New York: ACM Press, 2003: 19–22.
- [4] SHVACHKO K, KUANG H, RADIA S. The Hadoop distributed file system [C]// IEEE 26th Symposium on Storage Systems and Technology. Piscataway, NJ: IEEE Press, 2010: 1–10.
- [5] WEIL S A, BRANDT S A, MILLER E L, *et al.* Ceph: A scalable,

high-performance distributed file system [EB/OL]. [2010-05-10]. <http://www.ssrc.ucsc.edu/Papers/weil-osdi06.pdf>.

- [6] KARGER D, LEHMAN E, LEIGHTON T, *et al.* Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the World Wide Web [C]// STOC'97: Proceedings of the 29th Annual ACM Symposium on Theory of Computing. New York: ACM Press, 1997: 654–663.
- [7] BRINKMANN A, EFFERT S, auf der HEIDE F M. Dynamic and redundant data placement [C]// ICDCS'07: Proceedings of 27th International Conference on Distributed Computing Systems. Piscataway, NJ: IEEE Press, 2007: 29.
- [8] HONICKY R J, MILLER E L. Replication under scalable hashing: a family of algorithms for scalable decentralized data distribution [C]// Proceedings of the 18th International Parallel and Distributed Processing Symposium. Piscataway, NJ: IEEE Press, 2004: 96.
- [9] WEIL S A, BRANDT S A, MILLER E L, *et al.* CRUSH: Controlled, scalable and decentralized placement of replicated data [C]// Proceedings of the 2006 ACM/IEEE Conference on Supercomputing. New York: ACM Press, 2006: 31–42.
- [10] XIAO N, CHEN T. RAEDP: An effective hybrid data placement algorithm for large-scale storage systems [J]. *Journal of Supercomputing*, 2011, 55(1): 103–122.
- [11] DeCANDIA G, HASTORUN D. Dynamo: Amazon's highly available key-value store [C]// Proceedings of the 21st ACM SIGOPS Symposium on Operating Systems Principles. New York: ACM Press, 2007: 14–17.

(上接第616页)

明,在同样的检测时间内准确性优于 Chen 的 NFD-E 失效检测模型和 φ 失效检测模型,并且具有受窗口大小影响较小的特性。该模型能够适应变化的网络环境,可作为失效检测服务使用。

参考文献:

- [1] 陈宁江,魏峻,杨波,等. Web 应用服务器的适应性失效检测[J]. *软件学报*, 2005, 16(11): 1929–1938.
- [2] LINDHORST T, LUKAS G, NETT E, *et al.* Data-mining-based link failure detection for wireless mesh networks [C]// Proceedings of the 29th IEEE International Symposium on Reliable Distributed Systems. Piscataway, NJ: IEEE Press, 2010: 353–357.
- [3] TSAI W, SHAO Q, SUN X, ELSTON J. Real-time service-oriented cloud computing [C]// Proceedings of the 6th World Congress on Services (SERVICES-1). Piscataway, NJ: IEEE Press, 2010: 473–478.
- [4] GREVE F, SENS P, ARANTES L, *et al.* A failure detector for wireless networks with unknown membership [C]// Proceedings of the 17th International Conference on Parallel Processing. Berlin: Springer-Verlag, 2011, II: 27–38.
- [5] 穆飞,薛巍,舒继武,等. 一种面向大规模副本存储系统的可靠性模型[J]. *计算机研究与发展*, 2009, 46(5): 756–761.
- [6] DING X, HOU Y, GU Z, *et al.* A failure detection model based on message delay prediction [C]// GCC'09: Proceedings of the 2009 Eighth International Conference on Grid and Cooperative Computing. Washington, DC: IEEE Computer Society, 2009: 24–30.
- [7] CHEN W, TOUEG S, AGUILERA M K. On the quality of service of failure detectors[J]. *IEEE Transactions on Computers*, 2002, 51(5): 561–580.

- [8] HAYASHIBARA N, DEFAGO X, YARED R, *et al.* The φ accrual failure detector [C]// SRDS'04: Proceedings of the 23rd IEEE International Symposium on Reliable Distributed Systems. Washington, DC: IEEE Computer Society, 2004: 66–78.
- [9] DÉFAGO X, URBA P, HAYASHIBARA N, *et al.* Definition and specification of accrual failure detectors [C]// DSN'05: Proceedings of the 2005 International Conference on Dependable Systems and Networks. Washington, DC: IEEE Computer Society, 2005: 206–215.
- [10] BHOLE Y, POPESCU A. Measurement and analysis of HTTP traffic[J]. *Journal of Network and Systems Management*, 2005, 13(4): 357–370.
- [11] GOLMIE N, REBALA O. Bluetooth adaptive techniques to mitigate interference [C]// Proceedings of the Global Telecommunications Conference. Piscataway, NJ: IEEE Press, 2003: 405–409.
- [12] TENG W, CHANG C, CHEN M. Integrating Web caching and Web prefetching in client-side proxies[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2005, 16(5): 444–454.
- [13] 张世武,吴月华,杨杰,等. 基于信息寻觅智能体的网络用户浏览模式研究[J]. *计算机研究与发展*, 2004, 41(11): 1966–1973.
- [14] TUSHAR D C, SAM T. Unreliable failure detectors for reliable distributed systems [J]. *Journal of the ACM*, 1996, 43(2): 225–267.
- [15] KALEWSK M, KOBUSINSKA A, KOBUSINSKI J. Fast failure detection service for large scale distributed systems [C]// Proceedings of the 2009 17th Euromicro International Conference on Parallel, Distributed and Network-based Processing. Piscataway, NJ: IEEE Press, 2009: 229–236.