

# 一种自适应数据变化规律的数据采集算法

庞希愚<sup>1</sup>, 姜波<sup>2</sup>, 仝春玲<sup>1</sup>, 王成<sup>1</sup>

(1. 山东交通学院 信息科学与电气工程学院, 山东 济南 250357;

2. 中兴通讯股份有限公司 手机事业部, 江苏 南京 210000)

**摘要:**传统的等时间间隔的数据采集算法,在占用网络带宽和拟和精确度上存在着缺陷。鉴于此,该文提出了一种新的自适应数据变化规律数据采集算法,在该算法中,根据对当前一段时间所采集的数据变化情况的评估,设置了一种判断数据变化平滑度的策略,根据网络性能参数在当前时刻的变化趋势,自我调整采集数据的时间间隔,当判断数据变化比较平缓时,增加数据采集时间间隔;当判断数据变化比较剧烈时,减小数据采集时间间隔。最后,通过理论和实验分析验证了该数据采集算法在较小占用网络带宽的情况下,能得到更精确的数据拟和曲线。

**关键词:**数据采集;数据变化规律;自适应;网络性能管理

中图分类号:TP393

文献标识码:A

文章编号:1673-629X(2013)02-0157-05

doi:10.3969/j.issn.1673-629X.2013.02.042

## A Kind of Data Acquisition Algorithm of Adaptive Data Change Rule

PANG Xi-yu<sup>1</sup>, JIANG Bo<sup>2</sup>, TONG Chun-ling<sup>1</sup>, WANG Cheng<sup>1</sup>

(1. College of Information Science and Electrical Engineering, Shandong Jiaotong University,  
Jinan 250357, China;

2. Mobile Phone Division, ZTE Corporation, Nanjing 210000, China)

**Abstract:** The traditional equal-time interval data acquisition algorithm has defects in occupancy network bandwidth and fitting accuracy. In view of this, present a new data acquisition algorithm of adaptive data change rule. In this algorithm, according to the collected data changing situation assessment during the current period of time, set up a strategy of judging data change smoothness. According to the change trend of network performance parameters in the present moment, adjust the data gathering time interval, when data change judged is gentle, add data acquisition time interval; when data change judged is severe, reduce data acquisition time interval. Finally, through theoretical and experimental analysis verify that the data acquisition algorithm in small occupancy network bandwidth, can get more accurate data fitting curve.

**Key words:** data acquisition; data change rule; adaptive; network performance management

## 0 引言

目前现有的大量网络管理系统软件中,一般采用的数据采集算法都是等时间间隔的数据采集算法<sup>[1,2]</sup>;由系统或者用户事先设定好采样时间间隔,周期性地对性能数据进行采集和分析,这种方法的优点就是比较简单,易于实现,但是它完全忽略了数据值变化的特点,所以效率不高。

在这类网络管理系统软件的实际应用中,针对上面所提到的问题的处理往往是由管理员人工来进行调

整,管理员根据性能管理的实际要求,手工地增加或者减小数据采集的时间间隔,但是这样调整的缺点也是显而易见的。当要提高采样仿真曲线在原数据变化剧烈区的精度,则提高数据采集的频率(即降低采样时间间隔),但是这样的话,在数据变化平缓区的采样频率也同时被提高了,而对于数据变化平缓区的仿真不需要如此高的采样频率(提高数据采样频率,对于提高数据变化平缓区的仿真精度的影响不大),反而带来了相反的效果——浪费了大量的系统带宽。当要降低网络性能管理对于系统带宽的影响,人为地降低数据采集频率,对于数据变化剧烈区的仿真精度又不能得到保障。因此,对于等时间间隔数据采集算法,若想提高数据仿真曲线的精度,会造成网络带宽资源的浪费;而若想减小系统管理对于资源的使用,又不能保障数据仿真曲线的精度。

收稿日期:2012-06-21;修回日期:2012-09-25

基金项目:国家自然科学基金资助项目(61103022);山东省高等学校科技计划项目(J10LG05)

作者简介:庞希愚(1981-),男,山东济南人,讲师,硕士,CCF会员,研究方向为网络通信安全及相关算法设计。

正是对于上述等时间间隔数据采集算法不足的分析,目前已经有人提出了几种不同的应对方法<sup>[3~5]</sup>,但是这些算法大多不具有扩展性,功能不灵活,由此文中提出了一种在数据变化剧烈区就自动提高数据采样频率,而在数据变化平缓区就自动降低数据采样频率的自适应数据变化规律的数据采集算法。

## 1 自适应数据变化规律的数据采集算法的设计

对于一个实用的网络性能管理系统,一般来说,其本身对于网络带宽的消耗应该控制在 1%~5% 之间,否则网络管理系统就不能实现对网络的有效管理,甚至出现网络性能管理系统启动后,整个网络的性能反而下降的现象。但是,对于网络性能管理系统来说,其最根本的要求就是对性能数据采集及分析的有效性、实时性,在分析模型和算法不变的前提下,若想得到更精确的网络状况,通常需要采集更多的数据,但是采集更多的数据,无疑会消耗管理站和管理设备以及网络的软硬件资源,这与网络性能管理系统的本身对网络带宽的影响尽可能小的要求是背道而驰的。

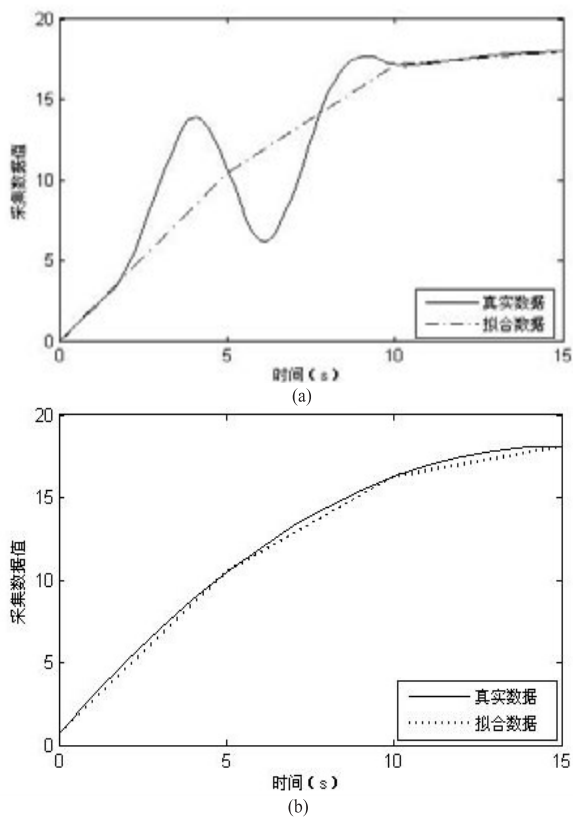


图 1 变化平缓程度对仿真精确度的影响

在实际的研究工作中,发现在网络性能管理系统中对于采集到的性能数据进行平滑仿真时,当性能数据的变化规律平缓程度不同时,同样的采样时间间隔得到数据的仿真精确度相差特别大,当性能数据值的

变化比较平缓时,仿真曲线所造成的失真比较小;反之,当性能数据值的变化比较剧烈时,仿真曲线所造成的失真则比较大,对于相同的性能指标数据进行采样,采用等时间间隔 15s,对于数据变化规律完全不同的两种情况,可得到如图 1 所示的仿真结果,(a) 图和 (b) 图分别是数据变化平缓和数据值变化剧烈时的不同仿真结果。从中可以清楚地看出,对于 (a) 图,当数据值的变化比较剧烈时,每 15s 采集一个数据值,来对实际性能状况进行仿真,明显造成了很大的失真;而对于 (b) 图,当数据值的变化比较平缓时,同样每 15s 采集一个数据值,对实际的性能状况进行仿真,造成的失真较小,这种情况下的仿真曲线精确度较高<sup>[6~9]</sup>。

由上面的分析可以得知,数据值的不同变化规律会对数据的仿真结果的失真造成完全不同的影响,下面文中先分析等时间间隔数据采集算法对于这种情况的处理方法的缺陷,然后具体设计一个自适应数据变化规律的数据采集算法。

### 1.1 数据变化平滑度的判断策略

根据等时间间隔数据采集算法的缺陷,文中设计了一种可以根据数据变化规律进行自我调整数据采集时间间隔的算法。简单来说,即此算法可以根据数据的变化情况,自我调整采集数据的时间间隔,当数据变化比较平缓的时候,自动加大数据采集的间隔时间,以节省带宽;当数据变化剧烈的时候,减小数据采集的间隔时间,以可以更加准确地分析被管网络的性能。

自适应数据变化规律的数据采集算法的主要思想是:根据已采集的性能指标的变化规律,动态地自我调整数据采集的时间间隔,当数据变化比较平缓时增大数据采集时间间隔;当数据变化比较剧烈时,则减小数据采集时间间隔。相应地,对于此算法设计应从如下两个方面着手:

①如何判断数据变化的剧烈程度,文中将其称为数据变化平滑度;

②用什么算法增减采样间隔时间。

由此,提出了将自适应数据变化规律的数据采集算法分为如下两个部分:

①数据变化平滑度的判断策略;

②自适应数据变化规律的时间间隔调整算法。

本算法判断数据变化平滑度的方法,是利用对当前一段时间内采集的数据的变化评估,来评价目前的数据变化平滑度。

由于网络数据的变化是随机的,很难找到一种函数或者曲线来具体模拟或者实现网络数据的图示,文中主要对采集数据的波动进行考察,对数据的波动大小进行量化评估,根据其波动大小的量化值的大小,从而得到了一种判断数据变化平滑度的方法。

首先,给出了如下的定义:

假设当前时刻  $t$  的性能数据  $D$  的采样值为:  $D_i$ , 其中  $i = 1, 2, 3, \dots$

定义1:将  $|D_i - D_{i-1}|$  称作当前时刻  $t$  数据的变化值, 其中  $i = 1, 2, 3, \dots$

定义2:取一时间段  $m (m > 0)$ , 单位为秒(s), 将其称为数据变化量化区间。

假设时间区间  $[t - m, t]$  内有  $k$  个采样点, 记为  $D_j, j = 1, 2, 3, \dots, k$ , 显然有  $D_k = D_t$ , 将  $\bar{D}_m = \frac{|D_k - D_{k-1}| + |D_{k-1} - D_{k-2}| + \dots + |D_2 - D_1|}{k}$  称为量化区间的数据平均变化值。

显然, 如果仅仅考虑此值的大小来判断一个数据的平滑度是不合理的, 因为此值的大小根据数据单位的不同会有很大的区别, 为此, 文中引入了利用平均值来判断数据段的平滑度。

定义3:取另一个时间段  $n (n \geq m > 0)$ , 其单位为秒(s), 将其称为数据变化平滑度分析区间, 假设时间区间  $[t - n, t]$  内有  $p$  个采样点, 记为  $D_j, j = 1, 2, 3, \dots, p$ , 显然有  $D_p = D_t, p \geq k$ , 将  $\bar{D}_n = \frac{D_1 + D_2 + \dots + D_p}{p}$

称为平滑度分析区间的数据平均值。

定义4:将

$$\text{flatDegree}(m, n) = \frac{\bar{D}_m}{\bar{D}_n} = \frac{p * (|D_k - D_{k-1}| + |D_{k-1} - D_{k-2}| + \dots + |D_2 - D_1|)}{k * (D_1 + D_2 + \dots + D_p)}$$

称为当前时刻  $t$  的采样点  $D_i$  在数据变化量化区间  $m$ , 数据变化平滑度分析区间  $n$  处的数据变化平滑度, 其中  $n > m > 0, i = 1, 2, 3, \dots$

数据变化平滑度 flatDegree 的具体意义表示是:当数据变化平滑度比较大时, 则表示数据变化比较剧烈; 反之, 当数据变化平滑度较小时, 则说明数据变化比较平缓。

不难看出, 数据变化平滑度是一个大于0, 通常情况下小于1的一个小数, 在文中提出的自适应数据变化规律的数据采集算法中, 用户可以根据数据的精度需求设定一个数据变化平滑度的临界阈值。当系统的一次采样结束后, 就计算此次采样点的数据变化平滑度, 如果此值大于事先设定的临界阈值, 就认为当前的采集间隔太大, 而数据变化比较剧烈, 应该减小数据采集的时间间隔, 提高采样的精度; 反之如果得到的值小于临界阈值, 那么则认为当前的数据变化比较平缓, 应该增加采样时间间隔, 从而可以节约更多的系统带宽。

此外系统可以根据随机数据的特点设定数据变化量化区间和平滑度分析区间的大小。数据变化量化区

间越大, 说明参与量化数据变化的点越多。在实际的网络监控中发现, 许多反映网络性能的数据会在一段时间的平滑之后, 出现一个数值的突变, 但是这个突变的时间很短, 一般在一个采样间隔的时间内就结束了, 如果因此而调整(减小)数据采集间隔, 则会造成控制信息的浪费, 同时也会造成网络流量的浪费。

图2是在对一台主机设备监控(采样间隔为4秒)下获取的一段网络流量的性能图, 可以看出在1250秒附近, 数据的值突然变大, 但是没有持续很长时间就结束了, 这种突变不代表网络性能的趋势, 如果因此而调整, 那么下一个采样点又要调整回原来的采样间隔, 可以看出这种性能突变给文中的算法带来了不必要的系统开销, 为了消除或者尽量减少这种突变数据对系统带来的负面影响, 可以增加数据变化量化区间的大小, 即增加参与量化数据变化的采样点, 从而减小了一个突变对整个系统的影响。但是, 这样的话, 必然会降低判断数据显著变化的灵敏程度。即是说, 对于数据变化量化区间的大小, 需要找到一个合乎系统的临界值, 太小不利于排除数据突变对系统的影响, 太大则会降低系统响应数据剧烈变化的灵敏度。

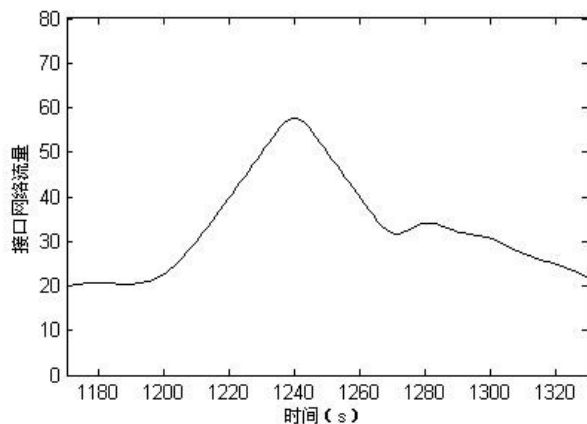


图2 一次短暂的数据值突变影响

## 1.2 自适应数据变化规律的数据采集时间间隔调整

自适应数据变化规律的数据采集时间间隔调整算法是指当系统根据上一小节中给出的策略判断出网络性能参数在当前时刻的变化趋势后, 系统根据这一结果来自动调整数据采集时间间隔的算法, 当判断数据变化比较平缓时, 增加数据采集时间间隔; 当判断数据变化比较剧烈时, 减小数据采集时间间隔。

为了更好地描述这一算法, 系统开始时, 将设定如下定义的参数。

数据采集基准时间间隔: 是指本算法中的最小时间单位, 算法中设定的数据采集时间间隔的增加或者减小必须是此时间单位的整数倍;

数据采集初始时间间隔: 系统运行时, 初始配置的数据采集时间间隔, 其为数据采集基准时间间隔的整



数倍;

最大数据采集时间间隔:算法自动调整数据采集时间间隔时,不得大于此数据采集时间间隔,其为数据采集基准时间间隔的整数倍,一般情况下,数据变化量化区间  $m$  的大小应该大于此值;

最小数据采集时间间隔:算法自动调整数据采集时间间隔时,不得小于此数据采集时间间隔,其为数据采集基准时间间隔的整数倍。

该文的调整算法,根据 flatDegree 的值自动地选择相应的数据采集时间间隔,当 flatDegree 处于不同的区间段时,相应地选择不同的数据采集时间间隔,当 flatDegree 比较大时,则选择比较小的时间间隔,反之,当 flatDegree 比较小时,则选择比较大的时间间隔。

## 2 数据拟和精确度分析

利用基于 SNMP 的网络管理系统,设定数据采集的时间间隔为 4s,监控本地主机的网络接口的网络流量,截取其中一段时间的性能数据,通过平滑曲线逼近获得了如图 3 所示的性能曲线图,并以此作为真实的网络性能状况<sup>[10~12]</sup>。

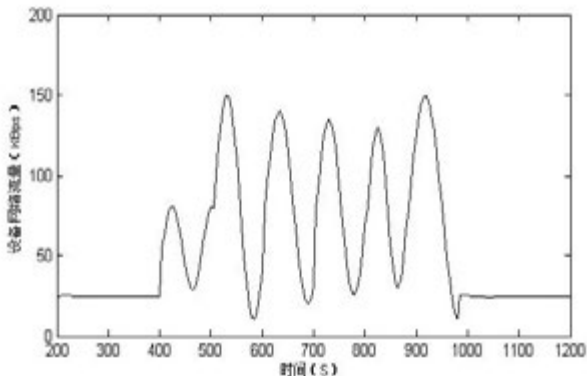


图 3 网络性能真实状况图

对于上面的性能数据中的时间段[200,1000],如果采用等时间间隔的数据采集方法,假设采集数据的时间间隔为 15s,对于上面一段时间的性能数据的拟和,可以得到如图 4 所示的结果。

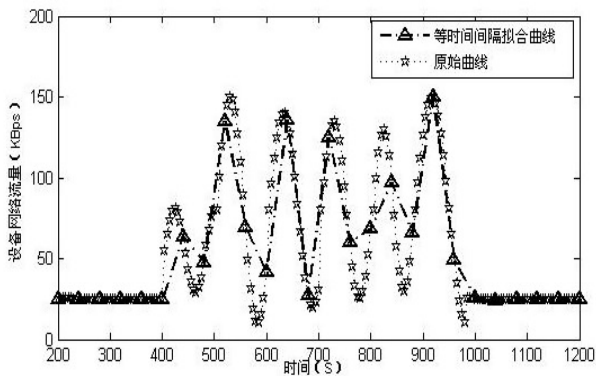


图 4 等时间间隔采集算法数据拟和曲线对比图

可以比较清楚地看出,在数据变化比较剧烈的状态下,这种拟和曲线并不能准确地描述出当前主机的网络接口流量状况,对真实情况进行拟和的失真比较严重。

利用文中所设计的自适应数据变化规律的数据采集算法,设定此算法中的一些参数如下:

数据采集基准时间间隔:5s;

数据采集初始时间间隔:10s;

最大数据采集时间间隔:30s;

最小数据采集时间间隔:10s;

数据变化量化区间  $m$ :60s;

数据变化平滑度区间  $n$ :200s。

自适应数据变化规律的数据采集时间间隔调整规则:

(1)  $0 < \text{flatDegree}(m, n) \leq 0.1$ , 设定数据采集时间间隔最大数据采集时间间隔;

(2)  $0.1 < \text{flatDegree}(m, n) \leq 0.4$ , 设定数据采集时间间隔为 25s;

(3)  $0.4 < \text{flatDegree}(m, n) \leq 0.7$ , 设定数据采集时间间隔为 20s;

(4)  $0.7 < \text{flatDegree}(m, n) \leq 1.0$ , 设定数据采集时间间隔为 15s;

(5)  $1.0 < \text{flatDegree}(m, n)$ , 设定数据采集时间间隔为最小数据采集时间间隔。

利用文中的算法,同样对原始性能数据中的时间段[200,1000]进行性能数据的拟和,可以得到如图 5 所示的拟和曲线。

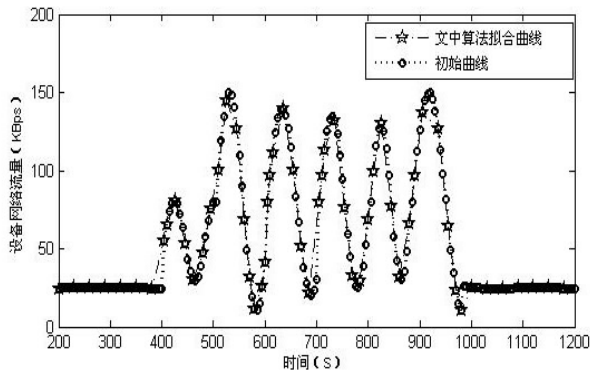


图 5 自适应数据变化规律数据采集算法的数据

对于同样的一段时间内的性能数据,通过对比可以清楚地看到,图 4 的采样点是均匀地分布在原来的数据曲线上,而图 5 的采样点的分布则是不均匀的,在数据曲线的平滑区域,采样点分布的比较稀疏,而在数据变化剧烈的区间,则相对要稠密得多,这正是本算法想要达到的,可以看出在等时间间隔采样的图中,在数据变化剧烈的地方,其拟和曲线的失真比较严重,而图 5 则失真很小。

对于同样的一段性能数据采集,等时间间隔采集方法在[200,1000]上共采集数据54次,而文中的算法在此段区间共采集数据40次,可以看出,在拟和曲线精确度相同的情况下,文中的算法,可以节约宝贵的网络带宽。另外,本算法还可以由用户自己来配置参数,可以使得对不同的网络性能参数进行灵活的配置,从而使此算法具有较大的适用范围。

3 结束语

当前,网络规模的不断扩大、功能复杂性的不断增加,给网络管理带来了前所未有的挑战。目前现有的大量网络管理系统软件中,一般采用的数据采集算法都是等时间间隔的数据采集算法:由系统或者用户事先设定好采样时间间隔,周期性地对性能数据进行采集和分析。针对传统的等时间间隔的数据采集算法,在占用网络带宽和拟和精确度上存在的缺陷,文中提出的自适应数据变化规律的数据采集算法在网络性能管理中对数据进行拟和时,可以在较小的管理通信开销下,得到更加精确的拟和曲线,有利于网络的健康发展。

参考文献:

[1] 王平,赵宏,陈海涛,等. 一个基于SNMP的简单网络管理系统的设计与实现[J]. 小型微型计算机系统,2001,22(9):1047-1050.

[2] 李木金,王光兴. 一种被用于网络性能管理的模型及实

[J]. 计算机学报,1999,22(11):1196-1203.

[3] 薛静,樊蓉,郑玉山. 基于回归分析的网络性能管理[J]. 微电子学与计算机,2005,22(2):161-164.

[4] 倪娜,晏蒲柳,夏德麟. 基于统计分析的网络性能管理与故障预测[J]. 计算机工程,2003,29(6):64-66.

[5] 杨军,张德运,张云翼. 基于分簇的无线传感器网络数据汇聚传送协议[J]. 软件学报,2010,21(5):1127-1137.

[6] Wu Y W, Li X Y, Liu Y H, et al. Energy-efficient wake-up scheduling for data collection and aggregation[J]. IEEE Transactions on Parallel and Distributed Systems, 2010, 21(2):275-287.

[7] Lin C, Tian Y, Yao M. Green network and green evaluation: mechanism, modeling and evaluation[J]. Chinese Journal of Computer, 2011, 34(4):593-612.

[8] Hao Z, Schizas L D, Giannakis G B. Power efficient dimensionality reduction for distributed channel aware Kalman tracking using WSNs[J]. IEEE Transactions on Signal Processing, 2009, 57(8):3193-3207.

[9] Jiang H B, Jin S D, Wang C G. Parameter-based data aggregation for statistical information extraction in wireless sensor networks[J]. IEEE Transactions on Vehicular Technology, 2010, 59(8):3992-4001.

[10] 许斌. JXTA-JAVA P2P 网络编程技术[M]. 北京:清华大学出版社,2003.

[11] Oaks S, Traversat B, Gong Li. JXTA 技术手册[M]. 朱剑平译. 北京:清华大学出版社,2004.

[12] 丁晓贵,刘桂江. 基于SOPC的远程数据采集系统设计[J]. 计算机技术与发展,2010,20(1):229-231.

(上接第156页)

乏深层次的加工、处理和运用。造成这种情况的重要原因之一即是分析和处理信息的手段尚显不足。今后工作中将考虑利用信息间的关联知识和因果知识来指导产品信息抽取的方法。

参考文献:

[1] 刘畅. 综合搜索引擎与垂直搜索引擎的比较研究[J]. 情报科学,2007,25(1):97-102.

[2] Cai Deng, Yu Shipeng, Wen Jirong, et al. Extracting content structure for Web pages based on visual representation[C]// Proceeding of the 5th Asia Pacific Web Conference. Berlin: Springer-Verlag, 2003:406-417.

[3] 林文清. B2B垂直搜索引擎在信息获取技术中的应用[J]. 情报杂志,2007(9):120-121.

[4] 余森,杨丹,赵俊芹. 垂直搜索引擎的关键技术研究[J]. 软件导刊,2007(12):31-33.

[5] 赵金仿,赵艳,缪建明. 网页信息抽取及其自动文本分类的实现[J]. 计算机技术与发展,2008,18(10):37-39.

[6] Cui Yang, Yang Bingru. A Method of Eliminating Noisy Infor-

mation in Web Pages for B2B Vertical Search Engine[C]// Proceedings of 2008 International Workshop on Information Technology and Security. [s. l.]:[s. n.], 2008:990-993.

[7] Caulkins J P, Ding W, Duncan G. A method for managing access to web pages: filtering by statistical classification (FSC) applied to text[J]. Decision Support Systems, 2006, 42:144-161.

[8] 李效东,顾毓清. 基于DOM的Web信息提取[J]. 计算机学报,2002,25(5):526-533.

[9] 李向阳,戴江山,张亚非. 一种Web信息抽取规则的优化方法[J]. 兰州理工大学学报,2006,32(1):90-93.

[10] 罗立宏,陈志. 基于语义分析的垂直搜索网络蜘蛛[J]. 计算机工程与设计,2008,29(18):4662-4665.

[11] 周明建,高济,李飞. 基于本体论的Web信息抽取[J]. 计算机辅助设计与图形学学报,2004,16(4):535-541.

[12] 王永庆. 人工智能原理与方法[M]. 西安:西安交通大学出版社,2001.

# 一种自适应数据变化规律的数据采集算法

作者: [庞希愚](#), [姜波](#), [全春玲](#), [王成](#)  
作者单位: [庞希愚, 全春玲, 王成\(山东交通学院 信息科学与电气工程学院, 山东 济南250357\)](#), [姜波\(中兴通讯股份有限公司 手机事业部, 江苏 南京210000\)](#)  
刊名: [计算机技术与发展](#)  
英文刊名: [Computer Technology and Development](#)  
年, 卷(期): 2013(2)

本文链接: [http://d.g.wanfangdata.com.cn/Periodical\\_wjtz201302042.aspx](http://d.g.wanfangdata.com.cn/Periodical_wjtz201302042.aspx)