移动计算环境中易变数据的在线广播调度

许华杰 李国徽
(华中科技大学计算机学院 武汉 430074)

摘 要 随着无线传感器网络、GPS等技术的广泛应用,产生了易变数据这种区别于传统静态数据的新型数据类型,对数据处理方法提出了新的要求。在移动计算环境中,数据广播是一种有效的数据访问方式。针对易变数据的特点提出数据平均不确定率的概念并在此基础上提出一种易变数据在线广播调度策略CEDB-M。仿真实验表明该策略在无传输差错发生、有传输差错发生和多信道广播条件下在获得较优的访问延迟的同时有效降低通过广播读取易变数据的不确定性,有利于基于这些数据的查询结果质量的提高。

关键词 移动计算环境,易变数据,广播,调度

On-line Scheduling for Constantly-evolving Data Broadcasting in Mobile Computing Systems

XU Hua-jie LI Guo-hui
(College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract With the increasing popularity of wireless sensor network and GPS, constantly-evolving data as a new type of data bring new challenge for the data processing methods. Data broadcasting is an effective method for data dissemination in mobile computing systems. Definition of the mean uncertainty ratio of data was presented. Furthermore, a broadcasting scheme CEDB-M was proposed for constantly-evolving data dissemination. Simulation results testify that CEDB-M can reduce the uncertainty of the broadcasted constantly-evolving data effectively at the cost of minor increase in data access time, in the case of no transmission error, presence of transmission errors, and multiple broadcast channels, thus benefit the qualities of the query results based on the data.

Keywords Mobile computing systems, Constantly-evolving data, Broadcasting, Scheduling

1 引言

随着传感技术、定位技术和无线通信技术的发展,近年来直接从外部世界中获取信息的系统的研究引起了学术界和产业界的广泛兴趣。这样的系统(例如无线传感器网络和全球定位系统GPS)从外部世界采集信息,并支持基于这些信息的新的应用,但同时也对现有的数据处理方式提出了新的挑战。在这些系统中,极度受限的系统资源,如网络带宽和电能供给,只能够实现数据以离散的方式进行采集,而外部环境中的数据(如温度、压力和位置等)是连续、不断变化的(此类数据称为易变数据,constantly-evolving data)。这就产生了一对矛盾:系统中保存的数据(值)和当前物理环境中的实际数据(值)可能不一致。尤其是在无线网络环境下,数据包常常会被延迟或丢失,这更加重了数据的不确定性。因此数据不确定性对数据处理系统提出了一个严峻的挑战:随着时间的推移,外部环境中的测量值可能已经发生改变,与数据库中保存的对应易变数据的值不再一致,易变数据的不确定性将会对数据处理系统查询结果的准确性产生巨大影响。图1(a)所示的是一个确定两个传感器中哪个检测到的温度更低(x或y)的查询。如果基于保存在数据库中的当前数据,查询的结果是"x",但实际上两个传感器所测量到的温度读数可能已经变成"$x1$"和"$y1$",在这样的情况下正确的查询结果应该是"y"。可见在易变环境中,数据库中的数据往往无法真正捕捉到外部世界的真实状态,外部世界值的变化无法及时反映到数据库中,涉及这些不正确数据的查询就可能会产生不正确的结果。

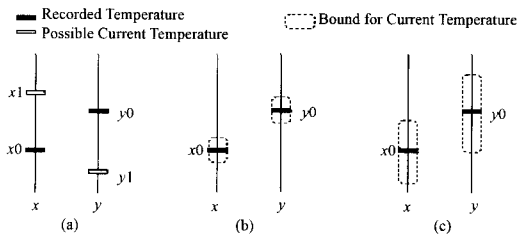

图1 易变数据的不确定性示意图

根据以上讨论,数据库中保存的易变数据给出对当前情况正确的查询结果似乎是非常困难的,但实际上外界数据的值一般不会发生突变,这一重要性质为通过保存在数据库中的数据来对外界易变环境下的测量值进行估计提供了可能。

到稿日期:2008-01-30 本文研究得到国家高技术研究发展计划(863计划)项目(No. 2007AA01Z309)和国家自然科学基金项目(No. 60203017)资助。

许华杰 博士生,CCF会员,研究方向为无线传感器网络、移动数据管理、不确定性数据处理,E-mail:hjxu@smail.hust.edu.cn;**李国徽** 博士,教授,博士生导师,研究方向为移动实时数据库、流数据处理、无线传感器网络。

如在某些移动对象数据库系统[1]中，传感器将值变化范围 q 连同测量值 a 本身一同传送回来，因此即使不知道传感器当前的实际读数也可以将测量值粗略地表示为$[a-q,a+q]$的形式。值的变化范围 q 是对数据的精确性和通信代价的一种折衷：q 的值越大，所需要的数据更新频率就越低，但是数据不确定性范围就越大，数据的不确定性程度就越高。这一"不确定性范围"信息对于提供正确的查询结果可能很有用。如图 1(b)所示，假设在发起查询的时候可以保证 x 和 y 当前的实际测量值与数据库中保存的值 $x0$ 和 $y0$ 之间的偏差被限制在一定范围内，利用不确定性范围信息我们就可以确定传感器 x 的读数较小。通过对图 1(b)和图 1(c)进行对比可以看出，数据的不确定区间大小对查询结果的质量影响很大，针对不确定区间较小的数据的查询结果质量较高、得到正确结果的概率较大；相反，当数据的不确定区间较大时，即使采用概率性查询处理方法，所得到的结果可信度也不高，意义不大。国际学术界对不确定性易变数据处理方法的研究方兴未艾，当前的成果主要包括易变数据的概率查询技术[2,3]、概率索引技术[4]、概率连接技术[5]以及易变数据挖掘技术[6-9]。

在移动计算环境中，相对于下行带宽，上行带宽有限甚至没有，因此数据广播的传送方式具有得天独厚的优势。数据广播主要分为 push-based，pull-based 和 hybrid3 种模式：push-based 广播[10]中服务器根据对数据需求的估计周期性地产生广播调度，其优点是可扩展性强。push-based 广播又分为离线和在线这两种方式，离线广播调度一旦生成，在广播的过程中就保持不变，而在线广播调度[12]是在广播过程中产生的，以适应数据需求的变化。pull-based 广播又称为按需广播[11]，根据客户机通过上行信道提交的数据请求产生广播调度，优点是可以充分考虑各个客户机的需求，缺点是可扩展性差、易产生传输瓶颈。hybrid 广播[13]综合 push-based 和 pull-based 两种广播模式。据了解，到目前为止出现的关于数据广播的研究成果考虑了传输错误、多信道、缓存和需求改变等因素，但是对具有一定不确定性的易变数据广播的研究成果尚未见报道。本文对具有一定不确定性的易变数据的广播调度问题进行探讨。

2 定义和问题描述

随着无线传感器网络技术的快速发展和广泛应用，当前国际学术界一种流行的观点是开发相应的技术以实现将广泛部署在自然界中的无线传感器网络整体看作是一个巨大的数据库，而对自然界监测得到的数据就是这个巨大数据库的数据源。数据库中的数据是现场采集的"获得性数据"而不是传统意义上存在硬盘中的数据[14]。如果这一目标最终得以实现，利用无线通信技术人们就能实现在任何时间、任何地点对自己所感兴趣的外部环境进行检索的梦想，这将大大拉近人与自然之间的距离。本文基于这样的思想，无线传感器网络或 GPS 从自然界采集的数据传送到服务器，生成数据库 DB，移动支持基站 MSS 根据移动客户机 MC 的需要调度数据，经无线信道进行广播，用户通过客户机 MC 可以对广播信道进行监听并读取自己感兴趣的数据。系统框架如图 2 所示。之所以采用数据广播的形式，是考虑到移动计算环境的特点和系统的可扩展性，因为客户机的数量可能非常巨大且不同客户机之间所需的数据重合的概率比较大。自然界变化的连续性与数据采样的离散性之间的矛盾决定了采集的数据必然是具有一定不确定性的易变数据，且其不确定区间随着时间的推移不断增大，因此数据必须根据其变化率及时传送到客户机。基于以上考虑，push-based 在线广播比较适合该应用环境。

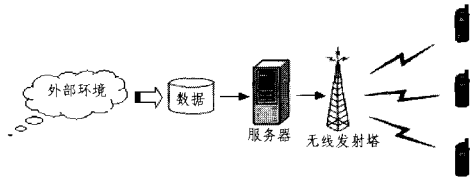

图 2 系统框架示意图

为了便于说明，假设数据以数据项为基本单位，每个数据项的长度相等(本文提出的方法稍加改动也适用于数据项长度不等的情况)。以在信道中广播一个数据项所需的时间作为单位时间。需要广播的数据项在广播中的位置安排称为广播调度。设需要广播的数据项数目为 M，一个广播周期为 N 个单位时间。数据项 i 出现在整个广播周期中的次数称为数据项 i 的广播频率 f_i。数据项 i 在广播中的一次出现称为 i 的一个实例。广播中数据项 i 的两个相邻实例之间的间隔称为数据项 i 的广播间隔，以 s_{ij} 表示数据项 i 的第 j 个实例与第 $j+1$ 个实例之间的间隔，$1 \leqslant j \leqslant f_i$。需求概率 p_i 表示根据以往的访问历史统计出来的数据项 i 被一个客户机访问的概率。数据项 i 的平均访问时间 t_i 和数据总平均访问时间 t 分别为：

$$t_i=\sum_{j=1}^{f_i}\frac{s_{ij}}{2}\frac{s_{ij}}{N}=\frac{1}{2}\sum_{j=1}^{f_i}\frac{s_{ij}^2}{N},\ t=\sum_{i=1}^{M}p_it_i=\sum_{i=1}^{M}p_i\left(\frac{1}{2}\sum_{j=1}^{f_i}\frac{s_{ij}^2}{N}\right)$$

易变数据的不确定性程度用数据的不确定率来衡量。不确定率越大，说明相应数据项的不确定程度越高。设 v_i 表示数据项 i 值的变化率，即单位时间数据项 i 值变化量与数据项 i 值之比，I 为服务器最近一次数据更新到目前所经历的时间，数据项 i 的平均不确定率 u_i 和数据总平均不确定率 u 分别为

$$u_i=\frac{\sum_{j=1}^{f_i}Iv_i}{f_i},\ u=\sum_{i=1}^{M}p_iu_i=\sum_{i=1}^{M}\frac{p_i\sum_{j=1}^{f_i}Iv_i}{f_i}$$

数据的不确定率对应的是数据不确定区间的相对大小，不确定率越小的数据越能如实反映数据所代表的物理量当前的真实状态，基于这些数据的查询所得到的结果就越可信。易变数据广播调度的目标是在满足数据的总平均访问时间 t 较优的前提下，使数据的总平均不确定率 u 最小。

3 易变数据在线广播调度策略

以下依次提出在移动计算环境中分别在不存在传输差错、存在传输差错和多信道广播情况下易变数据的在线广播调度策略。

3.1 易变数据在线广播调度策略(无传输差错)

文献[12]证明了获得最优平均访问时间数据广播所要满足的条件：

(1)同一个数据项的不同实例在广播中的间隔相等，即 $s_{ij}=s_i, j=1,\cdots,f_i$；

(2)$f_i \propto \sqrt{p_i}$ 或 $s_i \propto \frac{1}{\sqrt{p_i}}$ 或 $s_i^2\sqrt{p_i}=$常量(平方根准

则)。

根据平方根准则,并考虑到易变数据的不确定性,本文提出一种易变数据广播的调度策略:

易变数据在线广播调度策略 CEDB-M(Constantly-Evolving Data Broadcasting in Mobile Computing Systems)

1)计算每个数据项 i 的 $G(i)$ 值,其中 $G(i)=(T-R(i))^2 p_i v_i, i=1,\cdots,M$;

2)$G(j)=G_{max}=\mathrm{Max}\{G(i)\}, i=1,\cdots,M$;即选取 G 值最大的数据项 j 作为下一个广播的数据项;

3)在当前时间 T 广播数据项 j;

4)$R(j)=T; T=T+1$。

其中 T 表示当前时间;$R(i)$ 表示数据项 i 最近一次广播的时间,若数据项 i 还未广播过则 $R(i)$ 的值设置为初始值 -1;p_i 为数据项 i 的访问概率(需求概率);v_i 为数据项 i 值的变化率。由于在线调度方式下数据项相邻实例之间的间隔 s_{ij} 不尽相同,算法通过 $T-R(i)$ 避免对 s_{ij} 的计算。需要说明的是,数据库服务器根据从自然界采集的数据每隔一定时间 r 对数据库中的数据进行更新,更新的时间间隔 r 由数据采集系统(如传感器网络)通过对数据的流行性要求和通信代价两者进行权衡决定。在广播过程中,易变数据的不确定性(用不确定区间表示)随着时间的推移在不断增加,直到下一次数据更新易变数据的不确定性才又恢复到最小。由于服务器端数据的更新是同步进行的,且各数据项的更新间隔 r 相等,因此 r 的取值不会影响到 CEDB-M 调度中各数据项在广播中的相对位置。服务器完成对数据的更新后即进入下一个数据广播周期。移动客户机 MC 通过监听广播信道读取所需的、带不确定区间的数据,基于读取的数据可以采用文献[2,3,6-9]提出的不确定性数据的概率处理方法对带有不确定区间的易变数据进行值查询、范围查询、最近邻居查询、连接甚至是数据挖掘。CEDB-M 调度策略的计算复杂度为 $O(M)$,可以通过文献[12]中提出的将需要广播的 M 个数据项划分为 K 块再按块进行广播调度的方法降低计算复杂度为 $O(K)$,$K\leqslant M$。

3.2 传输差错对广播调度的影响

以上提出的在线广播调度策略 CEDB-M 假设服务器广播的每一个数据项都能准确无误地被客户机接收,并没有考虑到数据传输差错的发生。与传统有线网络环境不同,移动计算环境由于受使用环境、传输媒介等因素的影响,发生传输差错的概率相对较大。且在移动计算环境下的数据广播,若出现传输差错,无法直接要求服务器重传,而必须等到下个广播周期,因此在进行数据广播调度时有必要考虑传输差错造成的影响。设数据项 i 发生传输差错的概率为 e_i,则广播调度策略应做适当修改(由于篇幅有限,分析过程从略):

易变数据在线广播调度策略(考虑传输差错)

1)计算每个数据项 i 的 $G(i)$ 值,其中 $G(i)=(T-R(i))^2 p_i v_i \frac{1+e_i}{1-e_i}, i=1,\cdots,M$;

2)$G(j)=G_{max}=\mathrm{Max}\{G(i)\}, i=1,\cdots,M$;即选取 G 值最大的数据项 j 作为下一个广播的数据项;

3)在当前时间 T 广播数据项 j;

4)$R(j)=T; T=T+1$。

3.3 多信道数据广播

以上讨论的都是数据在单一信道广播的情况,多个客户机同时对同一广播信道进行监听。但实际上,现在很多无线设备都支持多信道传输,允许服务器同时以多个信道进行广播,而客户机根据自身的性能和需求对其中的若干信道进行监听。多信道的使用从某种意义上说相当于增加了广播带宽,显然有助于广播性能的提高。为了适应易变数据的多信道广播,CEDB-M 广播调度策略必须进行适当修改。具体来说,在广播的开始阶段对第一个信道的广播可以直接采用 CEDB-M 调度策略,而对于第二个信道开始的其他信道的广播,在采用 CEDB-M 调度策略时所使用的 $R(i)$ 的意义发生改变,它表示的是数据项 i 最近一次在已进行广播调度的信道中任何一个信道广播的时间。设可用广播信道数为 c,当前调度的广播信道为第 h 个信道,$1\leqslant h\leqslant c$,则考虑多信道的易变数据在线广播调度策略如下:

易变数据在线广播调度策略(对 h 信道广播的调度)

1)计算每个数据项 i 的 $G(i)$ 值,其中 $G(i)=(T-R(i))^2 p_i v_i, R(i)=\mathrm{Max}_{1\leqslant k\leqslant h}\{R_k(i)\}, i=1,\cdots,M$;

2)$G(j)=G_{max}=\mathrm{Max}\{G(i)\}, i=1,\cdots,M$;即选取 G 值最大的数据项 j 作为下一个广播的数据项;

3)当前时间 T 在信道 h 中广播数据项 j;

4)$R_h(j)=T; T=T+1$。

其中 $R_k(i)$ 表示最近一次在信道 k 上广播数据项 i 的时间,$R_h(j)$ 表示最近一次在信道 h 上广播数据项 j 的时间。

4 性能分析

在实际应用中,客户机对各个数据项的访问往往是不均匀的,呈现出明显的偏斜,如典型的"80%~20%"现象,即80%的访问作用在20%的数据项上。Zipf 分布模型可以很好地描述这种非均匀访问需求的概率分布。按照 Zipf 分布模型,数据项 i 的访问概率为

$$pi=c(\frac{1}{i})^{\theta},\ \text{其中}\ c=\frac{1}{\sum_{i=1}^{M}(\frac{1}{i})^{\theta}},\ 1\leqslant i\leqslant M$$

其中,θ 的大小反映数据访问的偏斜程度,称为访问偏斜因子。Zipf 分布模型中,当 $\theta=0$ 时,Zipf 分布等同于均匀分布;当 $\theta=1$ 时,基本上符合"80%~20%"分布。仿真实验中对数据项的访问概率符合 Zipf 分布。文献[12]提出的方法是比较经典的在线广播调度方法,且该方法在平均访问时间方面可以达到接近理论最优值的性能,调度效果比较理想。本文通过仿真实验对所提出方法在性能上与文献[12]提出的方法进行比较。为了方便比较,实验所用的参数及其取值与文献[12]中的基本相同,但只考虑数据项长度均匀分布(默认每个数据项的长度都为单位长度)的情况,原因是实际应用中数据项长度一般都是相等的,且本文提出的方法稍加改动即可适用于数据项长度不等的情况。另外增加两个参数:s 和 r。参数 s 表示数据项值的变化率,其上限范围是 $sb\in[0.1\%, 1\%]$,如不特别说明其默认值为 0.5%。sb 的取值虽然不大,但由于本文以广播一个数据项所需时间作为单位时间,而在移动计算环境下广播一个数据项所需时间非常短,因此在该数值范围的数据项值变化率实际上不小,客观反映了数据的易变性。仿真实验中各数据项值的变化率在 $[0,sb]$ 范围内随机分布。参数 r 表示服务器端数据更新的时间间隔,r 的取值不会对广播调度过程产生影响,只会影响数据不确定率的绝对大小,不失一般性,实验中取 $r=1000$ 单位时间。

4.1 CEDB-M 广播调度策略(不存在传输差错)

首先在不存在网络传输差错的情况下考察方法的性能。图 3 显示的是在不同的访问偏斜因子 θ 下的总平均访问时间,图中用 VHM(Vaidya-Hameed Method)代表文献[12]提出的方法,而本文提出的方法用 CEDB-M 表示。从图 3 可以看出,随着 θ 的增大,采用两种方法所获得的总平均访问时间都在减少。在总平均访问时间上 CEDB-M 方法的性能略逊于 VHM 方法,这是由于在广播调度时 CEDB-M 方法除了要考虑数据的访问概率之外还要考虑数据值的变化率。从图 4 可以看出,CEDB-M 方法在改善广播数据的不确定率上效果比较明显,因此可以说 CEDB-M 方法是以牺牲少量访问时间的代价换取数据不确定性的显著降低,从而有利于基于这些数据的查询质量的提高。

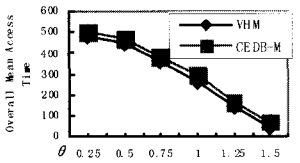

图 3 总平均访问时间与偏斜率的关系(sb=0.5%)

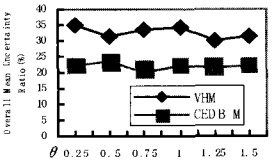

图 4 总平均不确定率与偏斜率的关系(sb=0.5%)

图 5 显示的是在数据访问偏斜率 θ=1 时总平均不确定率与数据项值变化率的关系。从图中可以看到,随着数据项值变化率的提高,两种广播调度方法所达到的数据总平均不确定率都逐步增加,但是 CEDB-M 方法增加的幅度比 VHM 方法的幅度平缓。特别是当数据项值变化率比较大(sb = 1%)时,CEDB-M 方法的优势就体现出来了,其数据总平均不确定率只有 VHM 方法的将近一半,效果比较明显。对比图 4 和图 5 还可以看出,数据总平均不确定率受数据访问偏斜因子 θ 的影响不大,而主要是受到数据项值变化率 sb 的影响。因此本文余下的实验在考察数据总平均不确定率时只考察其与数据项值变化率的关系,而忽略与数据访问偏斜因子的关系。

4.2 传输差错对性能的影响

为了便于分析,考虑较为常见的一种网络传输差错发生模型:差错的发生呈现以 λ 为参数的泊松分布,即数据项 i 发生传输差错的概率 $e_i=1-e^{-\lambda}$。此时,由于 $\frac{1+e_i}{1-e_i}=2e^{\lambda}-1$ 为常数,因此考虑传输差错的易变数据在线广播调度策略的调度过程和结果都与不考虑传输差错时相同,只是传输差错会导致总平均访问时间的绝对值有所改变,仿真结果见图 6。由于是否考虑传输差错并不影响调度的结果,因此总平均不确定率的值及其变化趋势均与图 5 相同,在总平均不确定率方面 CEDB-M 方法的性能优于 VHM 方法。

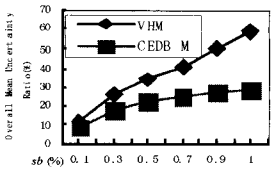

图 5 总平均不确定率与数据项值变化率的关系(θ=1)

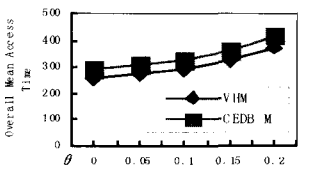

图 6 总平均访问时间与 λ 的关系(θ=1,sb=0.5%)

4.3 多信道广播调度

如果有多个网络通信信道可用,在多个信道上进行数据广播调度可以有效降低数据访问延迟,不同广播信道数所对应的总平均访问时间的仿真结果如图 7 所示。从这一点上看,广播信道的增加可以看作是广播带宽的增加。图 8 所示为在广播信道数 c=3、访问偏斜因子 θ=1 情况下分别采用 CEDB-M 方法和 VHM 方法广播数据的总平均不确定率与数据项值变化率的关系。通过对比图 8 和图 5 可以看出,无论采用 CEDB-M 方法还是 VHM 方法,多信道的使用对广播数据平均不确定率值的大小及其随 sb 变化的趋势影响不大。出现这一现象的原因是,由前面对多信道广播调度策略的讨论可知,无论是 CEDB-M 还是 VHM,其广播调度的过程与单信道广播调度类似,因此调度的结果(即各数据项在广播中出现的相对顺序)与是否使用多个信道无关。而广播数据的总平均不确定率是由各数据项在广播中出现的相对顺序决定的,因此在广播数据的平均不确定率方面多信道广播相对于单信道广播并没有什么优势。从广播调度的角度,多信道广播调度可以看作是将单信道广播调度的结果分散到多个信道上。从图 8 还可以看到,与单信道广播的情况类似,在多信道广播的情况下在平均不确定率方面 CEDB-M 的性能也明显优于 VHM。

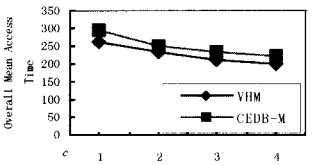

图 7 总平均访问时间与信道数的关系(θ=1,sb=0.5%)

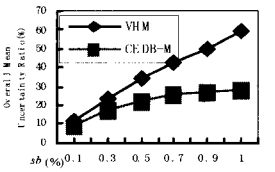

图 8 总平均不确定率与数据项值变化率的关系(θ=1,c=3)

结束语 无线传感器网络、GPS 等技术的广泛应用使从外部世界直接获取数据的数据库应用成为可能,而无线通信技术使人们可以不受时空限制对这些数据进行访问。由于受网络带宽等因素的限制,从外界获取数据只有采用离散采样的方式进行,自然界变化的连续性与数据采样的离散性之间的矛盾决定了从外部世界获得的数据本质上是不确定性随时间增长的易变数据,易变数据的处理对传统的数据处理方法提出了新的挑战。本文探讨的是易变数据的有效传播问题,针对易变数据的特点提出数据平均不确定率的概念,进而提出一种在移动计算环境下易变数据的在线广播调度策略,并通过仿真实验对该策略在无传输差错发生、有传输差错发生、多信道广播等情况下的性能进行测试。结果证实,本文提出的策略可以通过牺牲少量访问时间的代价换取数据不确定性的显著降低,从而有利于基于这些数据的查询质量的提高。未来的研究方向包括结合缓存、概率查询对查询数据的概率性要求的易变数据广播调度策略。

参 考 文 献

[1] Wolfson O, Sistla P, Chamberlain S, et al. Updating and querying databases that track mobile units. Distributed and Parallel Databases, 1999, 7(3)

[2] Cheng Chun-kong. Managing Uncertainty in Constantly-evolving Environments. PhD dissertation. 2005

[3] Cheng R, Kalashnikov D, Prabhakar S. Evaluating probabilistic queries over imprecise data // Proc. of the ACM SIGMOD Intl.

Conf. on Management of Data. June 2003

[4] Cheng R, Xia Y, Prabhakar S, et al. Efficient indexing methods for probabilistic threshold queries over uncertain data// Proc. of the 30th Intl. Conf. on Very Large Data Bases. 2004

[5] Cheng R, Xia Y, Prabhakar S, et al. Efficient join processing over uncertain-valued attributes// CIKM'06. November 2006

[6] Chau M, Cheng R, Kao B. Uncertain Data Mining: A New Research Direction // Invited Paper, the Workshop on the Sciences of The Artificial (WSA) 2005. National Dong Hwa University, Taiwan, Dec. 2005

[7] Chau M, Cheng Reynold, Kao B, et al. Uncertain Data Mining: An Example in Clustering Location Data // the Methodologies for Knowledge Discovery and Data Mining, Pacific-Asia Conference(PAKDD 2006). Singapore, April 2006

[8] Wang Kay Ngai, Ben Kao, Chun Kit Chui, et al. Efficient Clustering of Uncertain Data // the Proc. of the Sixth International Conference on Data Mining (ICDM'06). 2006

[9] Sau Dan Lee, Ben Kao, Reynold Cheng. Reducing UK-means to K-means// the Proc. of the 1st Workshop on Data Mining of Uncertain Data (DUNE2007). 2007

[10] Acharya S, Alonso R, Franklin M, et al. Broadcast disks: Data management for asymmetric communications environments // Proc. of the ACM SIGMOD Intl. Conf. on Management of Data. May 1995

[11] Acharya S, Muthukrishnan S. Scheduling on-demand broadcasts: New metrics and algorithms// Proceedings of the 4th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'98). October 1998

[12] Vaidya N H, Hameed S. Data broadcast in asymmetric wireless environments// Workshop on Satellite Based Information Services(WOSBIS). Nov. 1996

[13] Acharya S, Franklin M, Zdonik S. Balancing push and pull for data broadcast// Proc. of the ACM SIGMOD Intl. Conf. on Management of Data. May 1997

[14] Govindan R, Hellerstein J, Hong W, et al. The sensor network as a database. Technical Report, 02-771. Computer Science Department, University of Southern California, 2002

(上接第 23 页)

[13] Dewes C, Wichmann A, Feldmann A. An analysis of Internet chat systems// IMC '03. Miami Beach, FL, USA, October 2003

[14] ez-Campos F H, Smith F D, Jeffay K, et al. Statistical Clustering of Internet Communications Patterns. Computing Science and Statistics, July 2003, 35

[15] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques// SIGMETRICS'05. Banff, Alberta, Canada, June 2005

[16] Crotti M, Dusi M, Gringoli F, et al. Traffic Classification through Simple Statistical Fingerprinting. ACM SIGCOMM Computer Communication Review, January 2007

[17] Williams N, Zander S, Armitrage G. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. ACM SIGCOMM Computer Communication Review, October 2006

[18] McGregor A, Hall M, Lorier P, et al. Flow ClusteringUsing Machine Learning Techniques// PAM '04. Antibes Juan-les-Pins, France, April 2004

[19] Zander S, Nguyen T, Armitage G. Automated Traffic Classification and Application Identification using Machine Learning // LCN'05. Sydney, Australia, November 2005

[20] Bernaille L, Teixeira R, Salamatian K. Early Application Identification// CoNEXT'06. Lisboa, Portugal, December 2006

[21] Erman J, Arlitt M, Mahanti A. Traffic Classification Clustering Algorithms// SIGMETRICS'06 (MineNet). Pisa, Italy, September 2006

[22] Erman J, Mahanti A, Arlitt M, et al. Offline / Online Traffic Classification Using Semi-supervised Learning. Technical report. University of Calgary, 2007

[23] Bernaille L, Teixeira R, Akodkenou I, et al. Traffic classification on the fly. ACM SIGCOMM Computer Communication Review, 2006, 36(2): 23-26

[24] Duda R O, Hart P E, Stork D G. Pattern Classification. Second edition. Wiley, 2001

[25] Bernaille L, Soule A, Jeannin M-I, et al. Blind application flow recognition through behavioral classification. Technical report. Laboratoire d'Informatique de Paris 6, Universit'e Pierre et Marie Curie, 2005. http://www-rp. lip6. fr/ bernaill/techreport. pdf

[26] Bernaille L, Teixeira R. Early Recognition of Encrypted Applications// PAM'07. Louvain-la-neuve, Belgium, April 2007

[27] Moore A W, Zuev D. Discriminators for use in flow-based classification. Technical report. Intel Research, Cambridge, 2005

[28] Yu Lei, Liu Huan. Feature selection for high-dimensional data: A fast correlation-based filter solution// ICML'03. Washington DC, USA, August 2003

[29] Hongbo J, Moore A W. Lightweight Application Classification for Network Management // INM' 07. Kyoto, Japan, August 2007

[30] Guyon I, Elisseef A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 2003, 3: 1157-1182

[31] Chapelle O, Scholkopf B, Zien A. Semi-supervised Learning. Cambridge, MA: MIT Press, 2006

[32] Paxson V. Bro: A System for Detecting Network Intruders in Real-time. Computer Networks, 1999, 31(23/24): 2435-2463

[33] Erman J, Mahanti A, Arlitt M, et al. Identifying and Discriminating Between Web and Peer-to-Peer traffic in the Network Core// WWW'07. Banff, Canada, May 2007

[34] Erman J, Mahanti A, Arlitt M. Byte Me: A Case for Byte Accuracy In Traffic Classification // SIGMETRICS'07 (MineNet). San Diego, CA, June 2007

[35] Crotti M, Dusi M, Gringoli F, et al. Detecting HTT PTunnels with Statistical Mechanisms // CC'07. Glasgow, Scotland, June 2007

[36] Salgarelli L, Gringoli F, Karagiannis T. Comparing Traffic Classifiers. ACM SIGCOMM Computer Communication Review, 2007, 37(3): 65-68