(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104462187 B
(45)授权公告日 2017.09.08

(21)申请号 201410568300.X

(22)申请日 2014.10.22

(65)同一申请的已公布的文献号
申请公布号 CN 104462187 A

(43)申请公布日 2015.03.25

(73)专利权人 上海交通大学
地址 200240 上海市闵行区东川路800号

(72)发明人 闻于天 张奇 田晓华 杨峰 王新兵

(74)专利代理机构 上海汉声知识产权代理有限公司 31236
代理人 郭国中

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
CN 100391255 C,2008.05.28,
CN 102546972 A,2012.07.04,
CN 103312698 A,2013.09.18,
US 7278028 B1,2007.10.02,
CN 101345601 A,2009.01.14,

审查员 倪礼

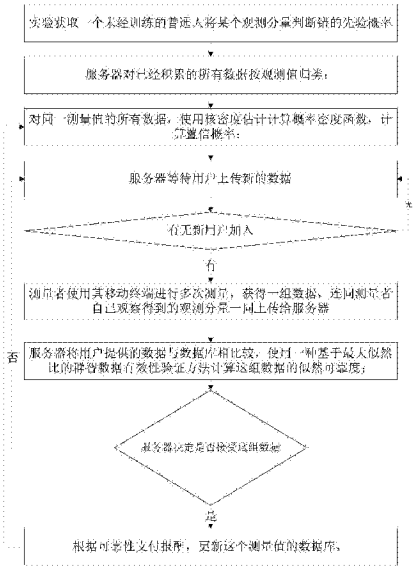
权利要求书3页 说明书6页 附图1页

(54)发明名称

基于最大似然比的群智数据有效性验证方法

(57)摘要

本发明提供了一种基于最大似然比的群智数据有效性验证方法,包括步骤:实验获取一个未经训练的普通人将某个观测分量判断错的先验概率;服务器对已经积累的所有数据按观测值归类;对同一测量值的所有数据,使用核密度估计计算概率密度函数,计算置信概率;服务器等待用户上传新的数据;测量者使用其移动终端进行多次测量,获得一组数据,连同测量者自己观察得到的观测分量一同上传给服务器;服务器将用户提供的数据与数据库相比较,使用一种基于最大似然比的群智数据有效性验证方法计算这组数据的似然可靠度;服务器决定是否接受这组数据,根据可靠性支付报酬,更新这个测量值的数据库,重新计算概率密度函数和置信概率。



CN 104462187 B

1.一种基于最大似然比的群智数据有效性验证方法,其特征在于,包括如下步骤:

步骤1:实验获取先验概率p_{1j},其中,p_{1j}表示对于某个观测分量j,一个未经训练的测量者将该观测分量j判断为观测分量1的概率;

步骤2:服务器对已经积累的所有数据按观测值归类;对同一观测分量j的所有数据,使用核密度估计计算概率密度函数,计算置信概率α_j;

步骤3:服务器等待用户上传新的数据;

步骤4:测量者i使用其移动终端进行多次测量,获得一组数据,这组数据连同测量者自己观察得到的观测分量一同上传给服务器;

步骤5:服务器将用户提供的数据与数据库相比较,计算这组数据的似然可靠度;

步骤6:服务器决定是否接受这组数据,根据可靠性支付报酬;如果服务器接受这组数据,返回步骤2,更新这个观测分量j的数据库,重新使用步骤2中的方法计算概率密度函数和置信概率α_j。

2.根据权利要求1所述的基于最大似然比的群智数据有效性验证方法,其特征在于,所述步骤1包括如下步骤:

步骤1.1:对于基于Wi-Fi信号强度的室内定位的训练过程中,测量者需要确定自己所处室内的位置,产生观测误差;测量者的观测误差被抽象为其处于房间中一点时对于房间最近的两个墙壁的距离的估计误差;

步骤1.2:通过预先的一次实验确定先验概率p_{1j}并将先验概率p_{1j}应用于所有室内定位的活动中;具体为,令多个测量者在一个没有距离参照物的房间里某些观测分量j判断观测分量1,收集该多个测量者的判断结果分布情况即作为p_{1j};

步骤1.3:对于不能通过预先的一次实验确定的p_{1j},取克罗内克函数:

$$p_{lj} = \delta_{lj} = \begin{cases} 0 \text{ if } l \neq j \\ 1 \text{ if } l = j \end{cases}$$

其中,δ_{1j}表示克罗内克函数。

3.根据权利要求1所述的基于最大似然比的群智数据有效性验证方法,其特征在于,所述步骤2包括如下步骤:

步骤2.1:服务器的数据库中的每个观测分量对应积累数据集D_j,j=1,2,3,…,N,N表示观测分量的总数,D_j中的各个元素D_j^k,k=1,2,3,…T,服从$f^j(x)$分布,T表示每个观测分量的数据总数,$f^j(x)$表示观测分量j所服从的概率密度函数;T=|D_j|>>M,M表示测量者一次上传的数据总数,则

$$f^j(x) = \frac{1}{T}\sum_{k=1}^{T} K_h(x - D_j^k)$$

其中,K_h表示核密度函数,x表示数据变量;

步骤2.2:设$n_s(x) = \sum_{k=1}^{T} 2h\, K_h(x - D_j^k)$,即$n_s(x)$表示[x-h,x+h]内数据库中已存在数据个数,h表示核密度函数K_h的带宽;

$n_s(x)$可能有T+1个取值,服从分布:

$$P(n_s(x) = n_s) = C_T^{n_s} P\left(\left|x - D_j^k\right| < h\right)^{n_s} \left(1 - P\left(\left|x - D_j^k\right| < h\right)\right)^{T - n_s}$$

其中，P(•)表示$n_s(x)$的概率质量函数，$n_s(x)$表示表示[x-h,x+h]内数据库中已存在数据个数，n_s取0,1,…,T,T+1中的一值，$C_T^{n_s}$表示从T个不同元素中取出n_s个的组合数，h表示表示核密度函数K_h的带宽；

步骤2.3：通过数据库大小确定r_{il}的期望，将这个期望作为置信概率α，其中，r_{il}表示观测者i所上传的数据属于观测分量l的概率密度；显然，不同观测值对应的积累数据量是不同的，因此对于不同观测值有不同的置信概率α_j。

4.根据权利要求3所述的基于最大似然比的群智数据有效性验证方法，其特征在于，所述步骤4包括如下步骤：

步骤4.1：测量者获得一组M个数据记作下式

$$\overrightarrow{x_{ij}}=\{x^1{}_i,x^2{}_i,x^3{}_i,\ldots,x^M{}_i;\ j\},$$

其中，$\overrightarrow{x_{ij}}$表示测量者i对同一观测分量进行多次测量获得的一组数据，j表示这组M个数据的一个需要观测的分量的真实值，$j\in\{1,2,3,\cdots,N\}$，N表示观测分量的总数；$x^t{}_i$服从分量j对应分布$f^j(x)$，$x^t{}_i$表示测量者i上传的第t个数据；

步骤4.2：观测误差体现为测量者将j判断为j′上报给服务器，即$\overrightarrow{x_{ij'}}$。

5.根据权利要求4所述的基于最大似然比的群智数据有效性验证方法，其特征在于，所述步骤5包括如下步骤：

步骤5.1：服务器取得数据$\overrightarrow{x_{ij'}}$后计算所有$\{r_{il}\}$：

$$r_{il}=\prod_{t=1}^{M}f^l(x^t{}_{ij'}),l=1,2,3,\ldots,N$$

其中，M表示测量者一次上传的数据总数，f(•)表示观测分量所服从的概率密度函数，l表示观测分量编号，$x^t{}_{ij'}$表示观测者i上传的第t个数据，并将其判断为观测分量j′，N表示观测分量的总数，r_{il}的物理意义为$\overrightarrow{x_{ij'}}$属于观测分量l的概率密度；显然，当l＝j时最大；

步骤5.2：定义参数$L(\overrightarrow{x_{ij'}})$，

$$L(\overrightarrow{x_{ij'}})=\max_l\left(\log(\frac{r_{il}p_{lj'}}{\alpha_j})\right)$$

其中α_j称为置信概率，$p_{lj'}$表示对于观测分量j′，测量者将该观测分量j′判断为观测分量l的概率；当α_j＝1时$L(\overrightarrow{x_{ij'}})$的意义为测量数据的最大可能概率密度的对数；显然对于相同长度的一组数据，$L(\overrightarrow{x_{ij'}})$较大者更可信；

步骤5.3：通过$L(\overrightarrow{x_{ij'}})$，能够对所有群智数据的有效性进行排序，取其中的前若干个。

6.根据权利要求5所述的基于最大似然比的群智数据有效性验证方法，其特征在于，

在步骤2.1中，取核密度函数为均匀核函数：$K_h=\frac{1}{2h},|x|<h$，h足够小使得数据在带宽范围内近似均匀分布，落到这个区域内的概率$P_s=P(|x-D_j^k|<h)=f(x)2h$；

在步骤2.3中，所有$L(\overrightarrow{x_{ij'}})>0$的数据都具有采用的价值，下面是一种计算$r_{il}$的期望$E\{r_{il}\}$的方法：

$$\alpha_j = E\{r_{il}\}$$

$$= E\{[f^l(x^t)]^M\}$$

$$= \sum_{x=-\infty}^{\infty}\left\{\sum_{n_s=1}^{T}\left(\frac{n_s}{T}\right)^M\left(\frac{P_s^{n_s}}{n_s!}\right)e^{-P_s}\right\}P_s$$

其中，$f^l(x^t)$表示观测分量l取值为x^t的概率密度，l表示第l个观测分量，t表示观测者上传的第t个数据，M表示测量者一次上传的数据总数，!表示阶乘，e表示自然底数，$P_s=P(|x-D_j^k|<h)=f(x_i)2h$，$f(x_i)$用核密度估计得出；上式中不存在T以外的变量，故确定了置信概率α_j与每个观测分量的数据总数T的关系。

基于最大似然比的群智数据有效性验证方法

技术领域

[0001] 本发明涉及通信技术领域，具体地，涉及一种基于最大似然比的群智数据有效性验证方法。

背景技术

[0002] 群智(crowdsourcing)在智能手机的应用中有十分广阔的前景。随着互联网技术的飞速发展，网络中个体的数量飞速增长，个体相互之间的联系也越来越紧密。在这样的大环境下，群智服务应运而生。如何有效的构建群智服务平台，促进社会中的资源共享，是下一代互联网研究需要解决的重要问题。

[0003] 如今，信息提供商往往采用群智激励机制(Crowdsourcing Incentive Mechanism)，将采集信息的工作交由分散的用户来做，并为他们提供的信息或服务给予一定的回报。例如有人想知道某段道路的拥堵情况，由正在该路段上的用户提供的信息不仅比提供商派人去勘察得到的信息更快也更准确。如今手机传感技术(Mobile Phone Sensing)正在蓬勃的发展之中，多种多样的传感设备正在被安装到智能手机上，例如加速传感器，GPS，距离传感器，相机等。利用这些分散的用户的智能手机传感技术获取到所需的信息并上传给提供商是现阶段逐渐流行的手段。

[0004] 尽管群智有众多优点，但是其弊端也是不可避免的。由于数据的测量者没有经过专业训练，测量的数据的观测误差总体来说会比较大，而且，由于测量者未经训练，不同数据的有效性的差异也会比通过传统方法获得的数据更大。极端情况下，如果测量者对测试对象非常陌生，甚至误操作，导致数据严重偏离了正常水平，采用这个数据将会对样本的有效性造成一定损害。

[0005] 这是群智场景中特有的一种误差，以下称为观测误差；其余的称为测量误差。这两种误差通常都可以用更大的样本量来弥补，但是我们的目的在于通过概率论的方法对群智数据进行定量评价与比较。进一步地，目的在于能从中筛选出相对有效性更高的一部分，也就是观测误差较小的一部分。

[0006] 经过对现有技术文献的检索发现，M.Ramadan等2008年在International Symposium on Personal，Indoor and Mobile Radio Communications发表的“Implementation and evaluation of cooperative video streaming for mobile devices”中提出了基于合作下载的视频分享机制，但该机制要求所有参与用户都相互认识并主动组成无线局域网，因而应用场景受到了极大限制。L.Keller等2012年在International Conference on Mobile Systems，Applications，and Services发表的“MicroCast:cooperative video streaming on smartphones”中提出了一种利用手机之间无线通信实现的视频协作下载加速机制。但该机制要求所有参与用户都希望下载同一个视频，该条件在大部分情况下都得不到满足，因而有很大的局限性。

发明内容

[0007] 针对现有技术中的缺陷，本发明的目的是提供一种基于最大似然比的群智数据有效性验证方法，通过利用服务器数据库中已经积累的大量数据内容更好地筛选有效的数据，减少录入错误数据造成的判断偏差。

[0008] 根据本发明提供的一种基于最大似然比的群智数据有效性验证方法，包括如下步骤：

[0009] 步骤1：实验获取先验概率p_{1j}，其中，p_{1j}表示对于某个观测分量j，一个未经训练的测量者将该观测分量j判断为1的概率；

[0010] 步骤2：服务器对已经积累的所有数据按观测值归类；对同一测量值j的所有数据，使用核密度估计计算概率密度函数，计算置信概率α_j；

[0011] 步骤3：服务器等待用户上传新的数据；

[0012] 步骤4：测量者i使用其移动终端进行多次测量，获得一组数据，这组数据连同测量者自己观察得到的观测分量一同上传给服务器；

[0013] 步骤5：服务器将用户提供的数据与数据库相比较，计算这组数据的似然可靠度；

[0014] 步骤6：服务器决定是否接受这组数据，根据可靠性支付报酬；如果服务器接受这组数据，返回步骤2，更新这个测量值j的数据库，重新使用步骤2中的方法计算概率密度函数和置信概率α_j。

[0015] 优选地，所述步骤1包括如下步骤：

[0016] 步骤1.1：对于基于Wi-Fi信号强度的室内定位的训练过程中，测量者需要确定自己所处室内的位置，产生观测误差；测量者的观测误差被抽象为其处于房间中一点时对于房间最近的两个墙壁的距离的估计误差；

[0017] 步骤1.2：通过预先的一次实验确定先验概率p_{1j}并将先验概率p_{1j}应用于所有室内定位的活动中，具体为，令多个测量者在一个没有距离参照物的房间里某些固定点j判断自己的位置1，收集该多个测量者的判断结果分布情况即作为p_{1j}；

[0018] 步骤1.3：对于不能通过预先的一次实验确定的p_{1j}，可取克罗内克函数：

[0019] $$p_{lj} = \delta_{lj} = \begin{cases} 0 \text{ if } l \neq j \\ 1 \text{ if } l = j \end{cases}$$

[0020] 其中，δ_{1j}表示克罗内克函数。

[0021] 优选地，所述步骤2包括如下步骤：

[0022] 步骤2.1：服务器的数据库中的每个观测分量对应积累数据集D_j，j＝1，2，3，...，N，N表示观测分量的总数，D_j中的各个元素D_j^k，k＝1，2，3，...T，服从$f^j(x)$分布，T表示每个观测分量的数据总数，$f^j(x)$表示观测分量j所服从的概率密度函数；T＝|D_j|＞＞M，M表示测量者一次上传的数据总数，则

[0023] $$f^j(x) = \frac{1}{T}\sum_{k=1}^{T} K_h(x - D_j^k)$$

[0024] 其中，K_h表示核密度函数，x表示数据变量；

[0025] 步骤2.2：设$n_s(x) = \sum_{k=1}^{T} 2h\, K_h(x - D_j^k)$，即$n_s(x)$表示[x-h，x+h]内数据库中已存在数据个数，h表示核密度函数K_h的带宽；

[0026] $n_s(x)$可能有T+1个取值，服从分布：

[0027] $$P(n_s(x)=n_s)=C_T^{n_s}P(|x-D_j{}^k|<h)^{n_s}(1-P(|x-D_j{}^k|<h))^{T-n_s}$$

[0028] 其中，P(•)表示$n_s(x)$的概率质量函数，$n_s(x)$表示表示[x-h，x+h]内数据库中已存在数据个数，n_s表示可能的取值，可取0，1，...，T，T+1中的任一值，$C_T^{n_s}$表示从T个不同元素中取出n_s个的组合数，h表示表示核密度函数K_h的带宽；

[0029] 步骤2.3：通过数据库大小确定r_{il}的期望，将这个期望作为置信概率α，其中，r_{il}表示观测者i所上传的数据属于观测分量l的概率密度；显然，不同观测值对应的积累数据量是不同的，因此对于不同观测值有不同的置信概率α_j。

[0030] 优选地，所述步骤4包括如下步骤：

[0031] 步骤4.1：测量者获得一组M个数据记作下式

[0032] $$\overrightarrow{x_{ij}}=\{x^1{}_i,x^2{}_i,x^3{}_i,\ldots,x^M{}_i;\ j\},$$

[0033] 其中，$\overrightarrow{x_{ij}}$表示测量者i对同一观测分量进行多次测量获得的一组数据，j表示这组M个数据的一个需要观测的分量的真实值，$j\in\{1,2,3,\ldots,N\}$，N表示观测分量的总数；$x^t{}_i$服从分量j对应分布$f^j(x)$，$x^t{}_i$表示测量者i上传的第t个数据；

[0034] 步骤4.2：观测误差体现为测量者将j判断为j′上报给服务器，即$\overrightarrow{x_{ij'}}$。

[0035] 优选地，所述步骤5包括如下步骤：

[0036] 步骤5.1：服务器取得数据$\overrightarrow{x_{ij'}}$后计算所有$\{r_{il}\}$：

[0037] $$r_{il}=\prod_{t=1}^{M}f^l(x^t{}_{ij'}),l=1,2,3,\ldots,N$$

[0038] 其中，M表示测量者一次上传的数据总数，f(•)表示观测分量所服从的概率密度函数，l表示可能的观测分量编号，$x^t{}_{ij'}$表示观测者i上传的第t个数据，并将其判断为观测分量j′，N表示观测分量的总数，r_{il}的物理意义为$\overrightarrow{x_{ij'}}$属于观测分量l的概率密度；显然，当l＝j时最大；

[0039] 步骤5.2：定义参数$L(\overrightarrow{x_{ij'}})$，

[0040] $$L(\overrightarrow{x_{ij'}})=\max_l(\log(\frac{r_{il}p_{lj'}}{\alpha_j}))$$

[0041] 其中α_j称为置信概率，$p_{lj'}$表示对于观测分量j′，测量者将该观测分量j′判断为观测分量l的概率；当α_j＝1时$L(\overrightarrow{x_{ij'}})$的意义为测量数据的最大可能概率密度的对数；显然对于相同长度的一组数据，$L(\overrightarrow{x_{ij'}})$较大者更可信；

[0042] 步骤5.3：通过$L(\overrightarrow{x_{ij'}})$，能够对所有群智数据的有效性进行排序，根据需要取其中的前若干个。

[0043] 优选地，

[0044] 在步骤2.1中，取核密度函数为均匀核函数：$K_h=\frac{1}{2h},|x|<h$，h足够小使得数据在

带宽范围内近似均匀分布，落到这个区域内的概率$P_s=P(|x-D_j^k|<h)=f(x)2h$；

[0045] 在步骤2.3中，所有$L(\overline{x_{ij'}})>0$的数据都具有采用的价值，下面是一种计算r_{il}的期望$E\{r_{il}\}$的方法：

$$\alpha_j = E\{r_{il}\}$$

$$= E\{[f^l(x^t)]^M\}$$

[0046]

$$= \sum_{x=-\infty}^{\infty}\left\{\sum_{n_s=1}^{T}\left(\frac{n_s}{T}\right)^M\left(\frac{P_s{}^{n_s}}{n_s!}\right)e^{-P_s}\right\}P_s$$

[0047] 其中，$f^l(x^t)$表示观测分量l取值为x^t的概率密度，l表示第l个观测分量，t表示观测者上传的第t个数据，M表示测量者一次上传的数据总数，!表示阶乘，e表示自然底数，$P_s=P(|x-D_j^k|<h)=f(x_i)2h$，$f(x_i)$用核密度估计得出；上式中不存在T以外的变量，故确定了置信概率α_j与数据库大小T的关系。

[0048] 与现有技术相比，本发明具有如下的有益效果：

[0049] 1、本发明可以通过预先实验矫正群智数据观测者的判断误差；

[0050] 2、本发明可以基于现有的可靠数据集，评价新进群智数据的有效性，从而合理对新进群智数据做出有效取舍。

附图说明

[0051] 通过阅读参照以下附图对非限制性实施例所作的详细描述，本发明的其它特征、目的和优点将会变得更明显：

[0052] 图1为本发明的步骤流程图。

具体实施方式

[0053] 下面结合具体实施例对本发明进行详细说明。以下实施例将有助于本领域的技术人员进一步理解本发明，但不以任何形式限制本发明。应当指出的是，对本领域的普通技术人员来说，在不脱离本发明构思的前提下，还可以做出若干变形和改进。这些都属于本发明的保护范围。

[0054] 本发明提供了一种基于最大似然比的群智数据有效性验证方法，包括步骤：实验获取一个未经训练的普通人将某个观测分量判断错的先验概率；服务器对已经积累的所有数据按观测值归类；对同一测量值的所有数据，使用核密度估计计算概率密度函数，计算置信概率；服务器等待用户上传新的数据；测量者使用其移动终端进行多次测量，获得一组数据，连同测量者自己观察得到的观测分量一同上传给服务器；服务器将用户提供的数据与数据库相比较，使用一种基于最大似然比的群智数据有效性验证方法计算这组数据的似然可靠度；服务器决定是否接受这组数据，根据可靠性支付报酬，更新这个测量值的数据库，重新计算概率密度函数和置信概率。

[0055] 具体地，本发明提供一种基于最大似然比的群智数据有效性验证方法，通过利用服务器数据库中已经积累的大量数据内容更好地筛选有效的数据，减少录入错误数据造成的判断偏差。

[0056] 参见附图1，本发明是通过以下技术方案实现的，本发明包括如下步骤：

[0057] 第一步：实验获取先验概率p_{1j}，表示对于某个观测分量j，一个未经训练的普通人将之判断为l的概率。

[0058] 第二步：服务器对已经积累的所有数据按观测值归类。对同一测量值j的所有数据，使用核密度估计计算概率密度函数，计算置信概率α_j。

[0059] 第三步：服务器等待用户上传新的数据。

[0060] 第四步：测量者i使用其移动终端进行多次测量，获得一组数据，连同测量者自己观察得到的观测分量一同上传给服务器。

[0061] 第五步：服务器将用户提供的数据与数据库相比较，使用一种基于最大似然比的群智数据有效性验证方法计算这组数据的似然可靠度。

[0062] 第六步：服务器决定是否接受这组数据，根据可靠性支付报酬；如果服务器接受这组数据，返回步骤2，更新这个测量值j的数据库，重新使用步骤2中的方法计算概率密度函数和置信概率α_j。

[0063] 下面更详细地将本发明的实施过程进行阐述。

[0064] 步骤一，假设服务器需要通过群智数据对某测量值进行测量，该测量值包含若干个观测分量。受观测误差的影响，测量者以概率p_{lj}将某个观测分量j误判为另一个观测分量l。实验首先获取先验概率p_{lj}。

[0065] 例如，对于基于Wi-Fi信号强度的室内定位的训练过程中，测量者需要确定自己所处室内的位置，产生观测误差。测量者的观测误差可以被抽象为其处于房间中一点时对于房间最近的两个墙壁的距离的估计误差。通过预先的一次实验就可以确定这个分布p_{lj}并将其应用于所有室内定位的活动中。招募大量志愿者在一个没有显著距离参照物的房间里某些固定点j判断自己的位置l，收集他们的判断结果分布情况即可视作p_{lj}。

[0066] 若不能通过预先的一次实验确定的p_{lj}，可以取Kronecker Delta函数。

[0067] $$p_{lj} = \delta_{lj} = \begin{cases} 0 \text{ if } l \neq j \\ 1 \text{ if } l = j \end{cases}$$

[0068] 步骤二，服务器的数据库中的每个观测分量对应积累数据集D_j，j＝1，2，3，...，N，其中各个元素D_j^k，k＝1，2，3，...T，服从$f^j(x)$分布，T＝|D_j|为数据集的大小。假设可以对其通过核密度估计足够精确地恢复出$f^j(x)$。则

[0069] $$f^j(x) = \frac{1}{T}\sum_{k=1}^{T} K_h(x - D_j^k)$$

[0070] 核密度函数可以取其他的任意形式，本领域技术人员可以在权利要求的范围内做出各种变形或修改，这并不影响本发明的实质内容。例如，取核密度函数为均匀核函数：$K_h(x) = \frac{1}{2h}, |x| < h$，h足够小使得数据在带宽范围内近似均匀分布，落到这个区域内的概率P_s＝P(|x-D_j^k|＜h)＝$f^j(x)$ 2h。

[0071] 设$n_s(x) = \sum_{k=1}^{T} 2h\, K_h(x - D_j^k)$，即[x-h，x+h]内数据库中已存在数据个数。$n_s(x)$可能有T个取值，其分布满足

[0072] $$P(n_i(x)=n_s)=C_T^{n_s}P(|x-D_j^k|<h)^{n_s}(1-P(|x-D_j^k|<h))^{T-n_s}$$

[0073] 由于不同的观测分量积累不同的数据量，因此不同的观测分量有不同的置信概率α_j。置信概率α_j用于衡量用户上传数据的采用价值$L(\overrightarrow{x_{ij'}})$，其中$\overrightarrow{x_{ij'}}$表示用户i上传数据，且该用户将其判断为观测分量j′。若用r_{il}表示$\overrightarrow{x_{ij'}}$属于观测分量l的概率密度，则r_{il}的期望就可以作为置信概率α。下面是一种计算E{r_{il}}的方法。

$$\alpha_j=E\{r_{il}\}=E\{[f^l(x^t)]^M\}$$

[0074] $$=\sum_{x=-\infty}^{\infty}\left\{\sum_{n_s=1}^{T}\left(\frac{n_s}{T}\right)^M\left(\frac{p_s{}^{n_s}}{n_s!}\right)e^{-P_s}\right\}P_s$$

[0075] 其中$P_s=P(|x-D_j^k|<h)=f(x_i)2h$，$f(x_i)$用核密度估计得出。式中不存在T以外的变量，故确定了置信概率α_j与数据库大小T的关系。

[0076] 步骤三，服务器等待用户上传新的数据。

[0077] 步骤四，测量者i对某个测量分量获得一组M个数据记作下式

[0078] $$\overrightarrow{x_{ij}}=\{x^1{}_i,x^2{}_i,x^3{}_i,\ldots,x^M{}_i;\ j\},$$

[0079] j表示这组数据测量分量的真实值，$j\in\{1,2,3,\ldots,N\}$。$x^t{}_i$服从分量j对应分布$f^j(x)$。观测误差体现为测量者将观测分量j判断为j′，并上报给服务器，即$\overrightarrow{x_{ij'}}$。

[0080] 步骤五，服务器取得数据$\overrightarrow{x_{ij'}}$后计算所有{r_{il}}：

[0081] $$r_{il}=\prod_{t=1}^{M}f^l(x^t{}_{ij'}),l=1,2,3,\ldots,N$$

[0082] 显然，当l＝j时最大。定义参数$L(\overrightarrow{x_{ij'}})$，

[0083] $$L(\overrightarrow{x_{ij'}})=\max_l\left(log(\frac{r_{il}p_{ij'}}{\alpha_j})\right)$$

[0084] 通过$L(\overrightarrow{x_{ij'}})$，系统可以对所有群智数据的有效性进行排序，根据需要取其中的前若干个。

[0085] 本实施例的环境参数为：

[0086] 移动终端设备：六部Android智能手机，都是Nexus 4，每部智能手机都配置有1.5GHz Snapdragon APQ8064 CPU和2 G RAM六部智能手机的操作系统都是Android Jelly Bean(4.2)。这六部智能手机并列作为测试手机进行室内定位。

[0087] 服务器：宏基4930G笔记本电脑，酷睿双核处理器，2G的内存，2G的主频。

[0088] 以上对本发明的具体实施例进行了描述。需要理解的是，本发明并不局限于上述特定实施方式，本领域技术人员可以在权利要求的范围内做出各种变形或修改，这并不影响本发明的实质内容。

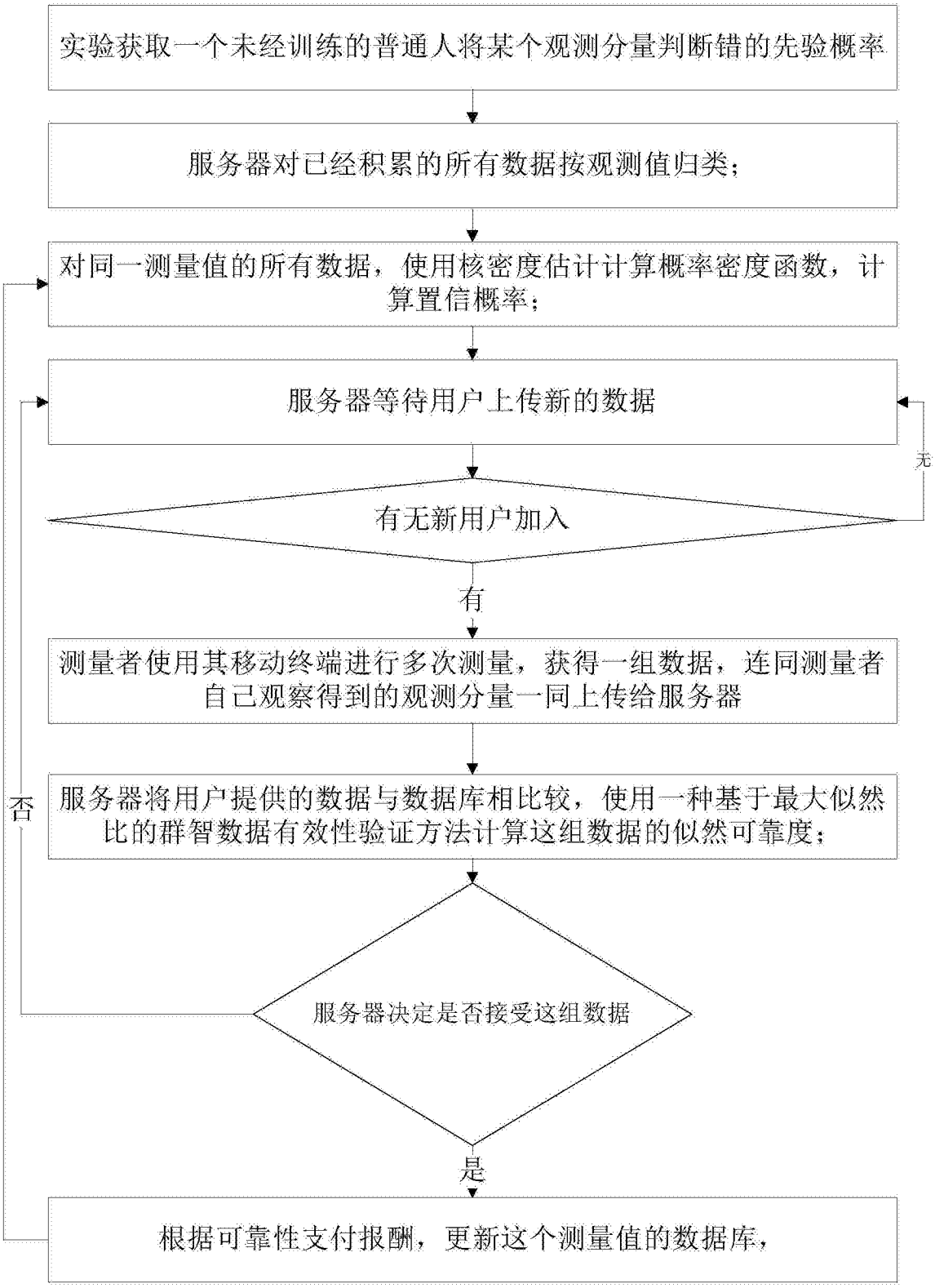

实验获取一个未经训练的普通人将某个观测分量判断错的先验概率
服务器对已经积累的所有数据按观测值归类；
对同一测量值的所有数据，使用核密度估计计算概率密度函数，计算置信概率；
服务器等待用户上传新的数据
无
有无新用户加入
有
测量者使用其移动终端进行多次测量，获得一组数据，连同测量者自己观察得到的观测分量一同上传给服务器
服务器将用户提供的数据与数据库相比较，使用一种基于最大似然比的群智数据有效性验证方法计算这组数据的似然可靠度；
否
服务器决定是否接受这组数据
是
根据可靠性支付报酬，更新这个测量值的数据库，

图1