我们的答案是积极的。 在本文中，我们对现实无线数据进行全面研究，以比较概率位置估计中的不同方法。我们的分析表明，对于在室内无线环境中收集的无线数据集，为了应用概率位置估计，信息理论提供了最好的特征选择方法，用于识别最重要的接入点和最小化决策所需的在线样本。我们还表明，为了确保能量效率计算，这是实时数据挖掘系统的新约束，可以使用基于多决策树的方法。我们开发了一种基于这种方法的算法，我们称之为CaDet，它代表（聚类和基于决策树的方法）。我们将位置估计的精度与采样时间和能量消耗联系起来。

我们的工作建立在以前基于使用廉价的无线局域网（LAN）作为基本基础设施的位置估计的工作。 为了检测用户位置，来自不同接入点（AP）的信号被收集并用作位置估计的基础。 利用估计的位置来进一步分析用户目标和目的的系统也正在出现。 在客户端仅感测来自不同AP的信号强度值的室内环境中，位置估计问题充满了不确定性。

据我们所知，我们在CaDet上的工作是第一个将能耗与无线领域中使用的数据挖掘方法联系起来的。我们的主要贡献是引入能量消耗作为建立预测模型的数据挖掘算法的设计的目标。在这一领域，我们提出了一个基于客户端的架构，在其上构建CaDet，客户端处理由位置估计中的各种AP发送的信号。在客户端上安装节能预测模型。我们在3.3节中展示了这种架构的优势。此外，我们提出了一种基于多决策树的方法，其中在离线阶段建立一个决策集合，每个集群一个。决策树允许使用最少的接入点，从而减少客户端上的计算和唤醒时间。我们演示了对客户端设备上的节能的影响。我们的设计为普及计算做出了贡献，因为基于客户端的位置估计被证明对于保持客户的隐私是有效的，但是在一般的计算中需要更多的功率。因此，普适计算的一个重要问题是如何节省能量，同时产生有效的预测。

本文组织如下：第2节讨论相关工作。 第3节介绍问题领域。第4节介绍我们的分析中使用的算法。第5节介绍实验。 本文在第6部分总结了对未来工作的讨论。

2.1 以前的工作地点估计

一般来说，位置估计研究可以分为两类：确定性技术和概率技术。确定性技术使用确定性推理方法来估计用户的位置。由Microsoft Research开发的RADAR系统提出最邻近启发法和三角测量方法来推断用户的位置。它维护无线电地图，其将从选定位置处的不同接入点接收的信号强度列表。然后将每个信号强度测量与无线电地图进行比较，并且对最佳匹配的坐标进行平均以给出位置估计。 RADAR的精度约为三米，概率为50％。LANDMARC系统利用参考点的想法来减轻由RFID信号强度波动所造成的影响。精度大致为1到3米。然而，参考标签的放置应仔细设计，因为它对系统的性能具有显着影响。移动，RFID读取器是如此昂贵，以致于其在大面积中的本地化是不可行的。在，基于来自用户的反馈的在线过程被采用以校正系统的位置估计。

由于客户端设备通常是小型的，依赖于电池电量的自我维护设备，如何节省能量的问题已经吸引了来自各种研究团队的很多关注。在普及计算领域，关于能量消耗有两个主要的研究问题：一个涉及硬件和其他软件。在硬件设计中，主要的问题是如何使移动设备更轻更紧凑，而不增加更多的功耗。已经有许多关于硬件电源管理的工作，其重点在于不同的组件，例如网络，磁盘和CPU。

在位置估计区域中，很少有工作解决了如何减少在线位置估计期间的计算开销的问题。In，Ladd et al。 使用概率推理方法进行本地化。 他们首先使用贝叶斯推理根据接收到的来自环境中9个接入点的信号强度测量结果计算位置上的条件概率。 然后，利用用户的移动轨迹的空间约束的后处理步骤用于细化位置估计并且拒绝具有位置空间中的显着变化的结果。

2.2

由于客户端设备通常是小型的，依赖于电池电量的自我维护设备，如何节省能量的问题已经吸引了来自各种研究团队的很多关注。 在普及计算领域，关于能量消耗有两个主要的研究问题：一个涉及硬件和其他软件。 在硬件设计中，主要的问题是如何使移动设备更轻更紧凑，而不增加更多的功耗。 已经有许多关于硬件电源管理的工作，其关注于不同的组件，例如网络和CPU。

对于软件方面的问题，移动软件的复杂性持续增长，因此增加了能源需求。有很多工作要从两个不同的方面解决节能问题：通信组件和计算组件。为了降低功耗，研究集中于通过尽可能地去激活无线电来优化通信成本，或者通过折衷用于通信的计算来进行研究。例如，在无线数据广播协议中，移动设备仅在所请求的数据帧的到达时间期间打开无线电。类似地，在传感器网络中，提出了局部网络架构以通过允许大多数传感器节点以保持在睡眠模式中并且通过减少远程传输的量。此外，呈现低能量自适应聚类分层结构以减少必须传输的信息量。然而，尽管计算组件比通信组件消耗更少的功率，但它们仍然是能量耗散的重要来源，特别是在通信成本被优化之后。因此，提出了基于预测的节能方案以减少用于对象跟踪的传感器网络的计算组件中的能量消耗。

在位置估计领域，小的工作已经解决了如何在在线位置估计的计算开销减少的问题。优素福等人。采用联合聚类技术为簇的位置来减少系统的计算成本。该方法定义了簇作为一组位置共享接入点相同。定位过程如下：在信号强度进行测量，最强的接入点来确定一个簇在寻找最有可能的位置，然后最大似然（ML）的方法被用来估计在簇中最可能的位置。然而，他们的方法有两个缺点：首先，聚类步骤只选择接入点相同的设置与信号强度最强的代表团；然而，不同的接入点和接入点的位置对同一组不同的信号值的判别能力没有被考虑。第二，ML方法需要几个条件概率分布的乘法。这仍然是苛刻的功率受限的客户端设备。我们的工作，在节约能源方面，有助于智能选择接入点聚类的目的，然后应用一个有效的估计方法，以减少计算成本。

3 无线环境

3.1 环境概述

我们的实验测试台设置在香港科技大学学术楼的计算机科学系的办公室。 该建筑配有在2.4 GHz频率带宽的IEEE 802.11b无线以太网。 地板的布局如图1所示。 该区域的尺寸为64米×50米。 在四个走廊（HW1〜HW4）和两个房间中进行实验，如图所示。 四个走廊的HW1为19.5，HW2为37.5，HW3为46，HW4为21。

在环境中总共有25个可检测的接入点，其中分布在该区域内的三个AP在图中用具体圆圈标记。 在其他22个AP中，一些位于该区域外的同一楼层上，而其他位于不同的楼层。使用我们开发的设备驱动程序和API，来自这些AP的信号由具有标准无线以太网卡的IBM膝上型计算机记录。当在办公区域内操作时，用户携带膝上型计算机用于数据收集和在线位置估计。

3.2 信号传播的特性

IEEE 802.11b标准在2.4GHz频带的无线电频率上工作。 它是广泛的，因为乐队在世界上大多数地方的免许可证。 它是有吸引力的，因为基于RF的技术是流行和廉价的，提供了很多无处不在的覆盖，并且需要很少的开销。

然而，由于信号传播的噪声特性，使用信号强度的测量的精确位置估计是长期困难的任务。受到结构甚至人体的反射，折射，衍射和吸收，信号传播在室内环境中遭受严重的多径衰落效应。结果，发送的信号可以通过不同的路径到达接收机，每个路径具有其自己的幅度和相位。 这些不同的组件组合和再现原始信号的失真版本。此外，即使环境条件的变化，例如温度或湿度，也在很大程度上影响信号。 因此，从固定位置处的接入点接收的信号强度随时间及其物理环境而变化。

图2给出从固定位置处的接入点接收的信号强度的归一化直方图的典型示例。 对几百个测量值进行采样以构建直方图。 很明显，从同一AP接收的信号强度随时间变化，即使在固定位置。 此外，覆盖位置的AP的数量也随时间变化。

3.3 我们的方法的理由

我们工作的一个关键新颖之处是通过应用数据挖掘算法为位置估计系统节省能量的能力。 我们采取了三种方法。 首先，我们采用了一种架构，其中我们只接收来自各种AP的信号，而不是发送它们。 正如我们将在后面看到的，这种感应模式节省了大量的电池电量。 第二，我们通过减少必须感测的AP的数量来优化位置估计算法，从而减少我们处理的数据量并节省计算时间。 第三种和最有效的方法是确保客户端设备的唤醒时间最小。

我们的架构是基于客户端 - 服务器的系统。在该系统中，客户端设备被保持在用户的手中，其由电池供电。手持式客户端系统通常消耗很多能量，这减少了位置估计系统的寿命，特别是当我们希望使整个操作尽可能普及时。论文和Ebert et al.检查了802.11范围内各种无线局域网的功耗。特别地，它们测量系统在不同模式下的功率消耗，包括系统休眠的休眠模式，系统在传输中不起作用的空闲模式以及两种传输模式。在传输模式中，系统可以传输包或接收包。在休眠模式下，平均功耗约为20MW。在空闲模式下，功率为110MW;在RX模式下，功率为900MW，而在TX模式下，功耗为2500 MW。影响功耗的其他因素包括分组的大小和传输速度。通常，当分组大且传输速度快时，系统使用较少的能量;然而，大包也增加了错误的机会，这又增加了能量消耗。

因此，在基于客户端的系统中，可以通过采用分组接收模式来节省传输能量消耗。然而，在基于服务器的系统中，必须将分组发送到服务器以便定位客户端。因此，为了使基于客户端的系统节省能量，一种方式是减少在客户端系统和服务器系统之间传输的信号量。一方面，在我们调查的位置估计系统中，RADAR系统[ 8]和市售的Ekahau系统需要传输信号到服务器，然后对客户端的位置进行预测。这些系统的优点在于，它们需要较少的离线训练，但是它们需要客户端设备上的大量电池电力。另一方面，我们在本文中提出的基于客户端的系统是仅接收由AP发送的包（即，使用RX模式），并且使用信号强度和离线获得的无线电地图来决定它在哪里这种基于客户端的架构需要更少的能量。另一个优点是由于主要计算是现场完成的，所以更容易保护客户承载的身份。

减少能量消耗的第二种方法是减少在客户端系统上进行的计算量。 在这种方法中，我们首先减少获取信号所需的AP数量，因为我们在应用聚类之后使用基于决策树的模型。 如我们稍后将展示的，减少AP的数量对应于数据挖掘中的维数减少，这减少了必须在系统上进行的乘法的数量。在我们的实验系统中，我们将演示通过该方法减少的计算量。

第三种方法是确保系统处于空闲或睡眠模式的时间量最大化。我们的方法通过最小化必须在系统可以做出可信决定之前实时接收的样本的数量来确保这种情况。 802.11b接口的最大比特率为11 Mbps，最大范围为100米。 Delaney使用PCMCIA 802.11b接口卡，并测量流入接口的平均电流以获得功耗。他使用开/关调度算法来降低802.11b设备的总能量消耗。当在802.11b电源管理模式下操作时，WLAN卡进入空闲状态。对于每100ms，它唤醒并接收业务指示图，其用于指示基站何时将向该特定移动主机发送数据。当存在严重的广播业务时，其发生在客户端设备进行信号传输时，WLAN接口将很少处于空闲状态，并且其将消耗与处于始终在线模式中一样多的功率。这是因为分析广播分组所需的时间大于睡眠模式。即使移动主机上没有运行应用程序，也会发生功耗增加。

在实验部分，我们将显示我们的系统在其用于位置估计的样本的一部分的准确性方面优于其他系统;这表明我们的系统在操作期间可以具有更长的睡眠时间。

表1显示了使用802.11b传输信号的HP Smartbadge IV嵌入式系统的功率测量从表中可以看出，无线通信占用了近一半的能量，而CPU活动占用了大约20％。 这意味着如果我们可以应用更智能的算法，增加设备的睡眠和空闲时间，并减少计算量，我们可以实现我们的节能目标。

4 cadet算法描述

我们的位置估计的CaDet算法分为两个阶段：

第一阶段离线完成，其主要目的是执行智能AP选择。 我们将此阶段划分为以下步骤：

1. 首先，应用特征选择算法来找到可以给出最佳性能的AP的子集S. 然后，该子集将被用作后续计算的基础。
2. 然后将后续聚类分析应用于在离线阶段收集的集合S和数据，以便将网格空间划分为群集。 然后每个群集将提供后续的位置模型。
3. 最后，基于在S中给出的AP为每个群集构建决策树模型。对于每个群集，仅选择来自S的AP的子集，这进一步减少每个群集内的位置估计所需的AP的数量。

第二阶段在线完成，其中以信号强度值的新轨迹作为输入，并且估计当前位置。 该阶段分两步完成：

1. 首先，使用来自集合S的所选AP的信号强度值来确定当前客户端的群集最可能位于其中。
2. 然后，来自所识别的集群的决策树用于在更细的级别确定客户端属于哪个网格。 该步骤将使用S中给出的AP的子集，这进一步减少了在计算中使用的AP的数量。 此外，所使用的AP仅涉及算术比较，其是关于计算能量的最便宜的计算之一。

4.1 在CaDet中离线AP选择

4.1.1 AP选择使用信息理论

通常，在环境中，来自许多AP的信号在这里或在所关注的区域内是可检测的。例如，在Rice大学的Duncan Hall地区可检测到9个AP，其用作实验的试验台。其中，5个位于该区域内，其他位于外部，包括在其他楼层。在许多其他环境中，例如我们的，可以检测到更多的AP。如3.1节所示，总共可检测到25个AP。来自每个AP的信号提供用于位置估计的一些信息，并且使用尽可能多的AP以提高位置估计系统中的准确性是自然的方式。然而，精度的提高是以对系统增加更多计算负担为代价的。因此，这种定位系统不仅具有差的可扩展性，而且当能量被约束在计算单元上时也是功率不足的。因此，重要的是仅使用目标系统能够承受的AP的数量，同时保持尽可能高的精度水平。

为了找到使用的AP数量和它们可以实现的准确性之间的权衡点，我们提出了一种基于信息增益的AP选择方法（简称InfoGain）。使用InfoGain进行AP选择的想法如下：假设在基于网格的定位系统中，n是网格的数量，m是可检测的AP的总数。每个AP（AP i ，l<=i<=m）被视为特征，并且每个网格（Gj,l<=j<=n）由这些m个特征描述。对于特定网格G，离线地收集来自AP的信号样本，并且来自AP i的平均信号强度被取为由于它们的物理位置以及信号传播的特性，也可能在G中未检测到一些AP。在这种情况下，对应的丢失的AP的特征采用默认值，其被设置为-95，在环境中接收的信号的最小强度。用于AP选择的InfoGain标准是根据其鉴别能力来评估每个特征（即，AP）的价值并选择最高的。当其值已知时，通过信息增益来测量特征AP i的辨别力。具体来说，它被计算为熵减少如下：



4.1.2 离线位置聚类在CaDet

在选择k个最有鉴别性的AP之后，下一步是聚集在环境中被建模为网格的位置。 聚类是模式分组的无监督分类。位置聚类的想法是接收的信号具有类似特性的位置形成簇。位置聚类是重要的，因为位置估计算法的复杂性可以通过首先识别未知样本所属的聚类，然后确定用户在聚类内的确切位置而大大减少。类似的权利要求在其中联合聚类技术。

我们将位置簇定义为从一组网格选k个接收类似信号的AP。每个网格可以由k信号强度的向量表示，其中第i个元素是在该网格处接收的平均信号强度。 直观地，在群集内的网格中接收的信号彼此更相似，而不是它们属于不同群集的网格的信号。两个网格的相似性可以根据它们的信号强度向量之间的欧几里得距离来测量与联合集群的概念相比，其中集群被定义为共享一组共同的AP的一组位置，我们的定义不仅考虑了AP在网格上的不同覆盖范围的方面，而且还反映了在信号强度的值。

在本文中，我们采用K均值聚类算法。 K-means是广泛使用的聚类算法，迭代地形成聚类。通过指定k，期望的群集的数目，算法通过任意选择k个网格作为k个群集中心开始。在每次迭代中，通过测量其信号强度与质心的距离之间的距离，将每个网格分配给最近的簇矩心。在分配了所有网格之后，通过取属于其的网格的信号强度的平均值来重新计算每个群集的质心。该迭代过程继续，直到实现收敛，其中k个质心不再移位。网格最终与其质心最接近该网格的k个聚类之一相关联。因此，在所有网格被分配之后，它们被分组成数量明显小于网格的数量的簇。正如我们将在实验（第5.3节）中看到的，K-means算法的使用通过聚类结果证明，几何上相互接近的网格聚集在一起。然而，群集k的数量仍然是开放的问题。如果我们指定k太大，则有太多的集群。结果，存在大量冗余计算，并且AP的数量没有减少。然而，如果我们指定太少的AP，则我们不能利用来自位置的局部化模式。在下文中，我们将经验地改变聚类k的数目，以便找到最佳平衡。

4.1.3 决策树的聚类内估计

在识别群集之后，CaDet确定用户在粗略级别（低分辨率）的位置。 在下一阶段中，需要区分相同群集中的网格，以使得位置估计在更精细的水平，导致高分辨率，同时减少所使用的AP的数量。

CaDet使用一种基于决策树的简单而有效的方法[31]。 决策树已经在广泛的模式分类应用中使用。决策树背后的想法是自然而直观的：通过一系列问题对测试样本进行分类。作为示例，在位置聚类的网格（G 1〜G 8）上构建的决策树在图5中示出。在图中，每个内部节点对应于关于特定AP的信号强度的值的测试问题。 从内部节点分支出几个子树，每个子树对应于不同的值范围。 从根节点开始，将向测试样本询问问题序列，直到它到达叶节点。 最低级别的叶节点表示样本属于哪个网格的决策。 更多的细节将通过4.2节中的例子给出。

4.2 在CaDet中的模型的在线应用

一旦建立模型，我们可以将它们应用于在线位置估计。 对于给定的接收信号样本，来自所选AP的信号强度值被用于确定当前客户端最可能位于的集群。随后，与所识别的集群相关联的决策树用于在更精细 级别，客户端属于哪个网格。任意破坏。我们现在用一个例子来说明这个过程。

作为示例，假设使用InfoGain算法选择的四个AP的向量。 测试样本T的信号强度列于表2中。令C j表示由离线聚类算法给出的总M个群集中的第j个群集的质心。

4.3 节能的分析

Once the models are built, we can then apply them to online location estimation. For a given received signal sample, the signal-strength values from the selected APs are used to determine the cluster that the current client is most likely located in. Subsequently,the decision tree associated with the identified cluster is used to determine,at a finer level,which grid the client belongs to.Ties are broken arbitrarily. We now illustrate this process using an example.