第5章　数据质量评估与优选

随着嵌入式设备、无线传感网络、物联网、智能移动终端等的快速发展，普适计算逐步融入人们的日常生活环境中，所获取的数据也越来越多样。传统的数据收集方式主要依赖专业感知设备，如摄像头、空气质量检测仪等。虽然采集数据准确，但具有部署难，维护成本高，采集数据单一和范围受限等问题，因而在应用范围和应用效果上受到很大限制。为了解决相应问题，移动群智感知将采集对象转向大量普通用户来提高覆盖范围，通过用户随身携带的智能设备进行大规模、随时随地的采集方式。

这一方式虽然能够大范围覆盖城市区域，收集多种多样的城市信息，但由于不同用户在活动范围上有一定重叠，群智感知采集到的数据中可能存在大量冗余。而大量未经训练的用户作为基本感知单元会带来感知数据多模态、不准确、不一致等质量问题。在此背景下，如何在数据冗余、质量良莠不齐情况下对参与者上传到中心服务器的感知数据进行质量评估和优选，并将得到的智能信息运用到各种创新型服务中是群智感知领域的研究热点。

5.1　移动群智感知数据质量

数据的收集和传输是群智感知网络实现其作用的必经途径，也是群智感知网络任务完成的根本保证。传统的无线/有线传感器网络所用的终端感知设备大多为专业传感器，如温湿度传感器、环境污染物传感器等，一般可以在感知任务部署前进行设备的校正或误差修正，因此，相应的感知数据质量评估相对容易。然而，在移动群智感知网络中，一方面众多参与者在操作技能、任务认知等方面存在较大差异，另一方面不同参与者所拥有智能移动设备的种类多样、属性各异，这两个要素均会对所收集的感知数据带来显著的影响，从而导致收集到的感知数据存在质量参差不齐的现象。因此，如何科学衡量群智感知系统收集数据的质量并优选高质量数据是一个亟待解决的问题。

现有的研究大多将群智感知数据质量与能耗、隐私、感知源等方面进行联合研究[1-2]，希望能够获得高质量的数据。然而，影响数据质量的因素极其复杂，很难依靠单独因素的改进而显著提高。具体来说，群智感知数据质量受多方面因素的影响，主要包括：①用户使用的感知设备类型与属性。例如，高端手机所配置的传感器种类一般比低端手机所配置的传感器种类多，同时传感器精度也高。②用户采集数据的环境和方式。例如，手持移动设备采集环境噪声的数据质量比将移动设备放置在衣服口袋或手提包里采集环境噪声的数据质量高。③用户的认知/技能水平。例如，基于移动群智感知的图像搜索应用依赖用户对图像的识别能力，而不同用户对同一图像的认知可能存在偏差。④用户的主观参与意愿。例如，有些用户会严格按照任务要求采集数据，而有些用户会比较随意，甚至存在恶意用户上传虚假伪造数据以骗取奖励的情况。

总而言之，数据质量的好坏是衡量感知系统优劣的直接指标，是提升群智感知性能和应用服务的基础与保障，也是影响移动群智感知系统能否得以普及的关键因素。通过对所收集的数据质量进行有针对性的评估，并根据评估结果采取相应的措施与方法增强数据质量，可以有效地支撑上层相关应用需求。为了提高群智感知系统的效率，感知数据质量的合理评估是必须解决的问题。

5.2　移动群智感知数据选择

由于不同感知参与者在时空范围上的重叠现象以及部分未经训练用户在任务执行阶段的不熟练，所收集的感知数据不准确且冗余。为了避免大量低质量数据上传对系统和服务器工作效率的影响，过滤冗余数据，筛选优质数据是常用方式之一。

由于移动设备的续航能力有限，大多数移动群智感知应用都没有让智能设备承载过多的计算任务，因而将大量数据不经筛选地上传至服务器是现有阶段群智感知系统进行数据优选的基本方式。但是，随着智能设备软硬件的发展，移动群智感知的数据优选功能可移至智能设备执行，从而提高数据收集质量并降低通信负载和数据中心负载。

一般来说，移动群智感知数据从用户端上传到服务器，主要经历两个阶段：数据选择和数据移交。数据选择可分为前置选择和后置选择。数据移交可分为实时移交和容延移交。前置选择指数据在上传之前对其效用进行评估，然后选择可用数据上传至服务器；而后置选择指数据上传至数据服务器或交付给任务发布者之后再进行数据选择操作。前置选择适用于网络带宽有限或网络流量费用较高的情况，通过前置选择，在数据上传前过滤掉一些低质、低效用和冗余的数据，有助于节省带宽和流量。后置选择适用于对数据时效不敏感的应用，原始数据可通过免费或低收费的网络全部上传至服务器，优点是数据传输成本低，而且由于服务器可以针对完整的数据集进行去噪和去重，因此其数据选择的精度和召回率理论上高于前置选择方式。在数据移交方面，实时移交指参与者采集感知数据之后立即上传，数据失效快，但需要良好的网络通信环境。容延移交中数据的时效性较低，参与者可以在接入免费或低收费的网络后上传数据。容延移交可以由即时移交代替，反之则不行。感知数据优选的目的在于如何在数据大量冗余、质量良莠不齐的情况下实现优质数据的甄别与选择，因此，本小节主要介绍数据选择的具体流程和方法，数据移交的相关内容将在下一小节详细的介绍。

对于多媒体文件（如视频、音频等）的数据选择而言，其数据传输所需资源（如带宽、流量等）较多，因此需要通过感知节点与后台数据中心之间的交互，实现前置数据选择（Pre-selection）。在前置选择方式下，客户端首先提取数据的关键信息，然后上传给数据中心，数据中心根据感知任务对数据集的多维约束（如时空特征、质量、采样间隔、方向等）评价数据质量，最终只有高质量的数据被完整上传至数据中心。根据任务需求从原始数据集中优选出理想的数据子集是降低通信、计算等资源消耗的有效途径。而低冗余、高覆盖是理想子集的通用化标准，其中低冗余指数据子集中没有相同或相似的数据，高覆盖是指数据子集能够最大限度地全方位反映感知对象的真实情况，如图 5-1 所示。

实际中，不同的移动群智感知任务在评价冗余度和覆盖度时的标准是不同的，也就是在数据选择时需要遵循的选择约束是不同的，因此数据优选的方法也不同。为了提高系统效

率，降低数据传输成本，如何在数据大量冗余、质量良莠不齐的情况下实现优质数据的甄别与选择是数据优选需要解决的问题。

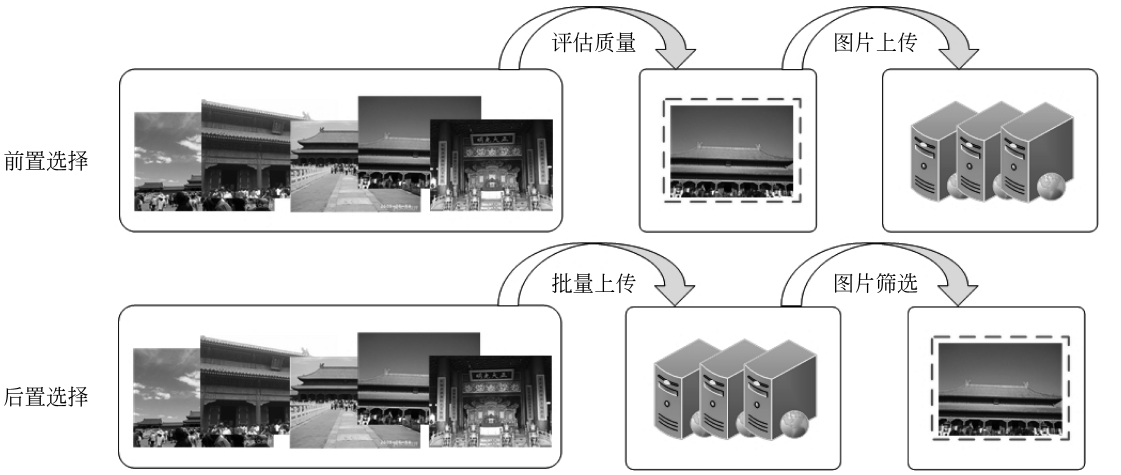

图 5-1　不同的数据选择方法

5.3　研究进展

作为一个新的研究领域，移动群智感知数据的质量评估与优选有其自身的研究挑战和问题，包括用户质量不高、数据大量冗余和数据欺骗等问题。下面分别有针对性地对其现阶段所取得的研究进展进行介绍。

5.3.1　移动群智感知数据质量评估

根据感知任务或对象的类型的不同，有 3 种典型的感知数据的质量评估和保障方法能够合理评估数据质量，包括面向二值型任务的方法、面向多类别型任务的方法和面向连续信号型任务的方法。在面向二值型任务的方法中，任务的结果只有两种："是"或者"否"。事件检测是一种典型的二值型任务，即判断某种事件是否发生。最简单的方法是基于投票统计的方法[3]，即当判定事件发生的用户数量超过特定阈值的时候，才最终确定事件发生。在面向多类别型任务的方法中，多类别型任务的结果多于两种，例如，用户对某个事物的评价可以打分为 1～5 的某个分数。投票法虽然也可以用来度量结果的不确定性，但还不够准确。最大期望法[4]是一种常用的更准确的方法，它以迭代的方式首先根据用户的感知数据估计用户的可靠性，然后根据用户的可靠性估计最终的任务结果，并不断重复上述过程。面向连续信号型任务的方法不同于前两种，主要针对连续信号任务监测，例如对区域环境现象的连续监测属于连续信号型任务。Koutsopoulos 针对这类任务提出了一种感知数据质量度量方法，该方法将某个用户提交的历史数据与所有用户数据的平均值之间的累积误差，作为该用户的感知数据质量指标[5]。

以上 3 种方法的基本思想都是通过多人重复采集抵消个人数据不准确的影响，从而提高整体数据的可靠性。然而，相比于群智感知网络中其他方面的研究，关于感知数据质量的

研究还比较少。文献[6-9]只考虑群智感知网络中最简单的情况，即参与者“是/否”感知。但在实际的群智感知网络任务中，其所要求的感知数据信息远比“是/否”要复杂得多，例如各地信号强度值和污染物指数等数据，这些都不是仅用“是/否”就能完整描述的。文献[10]基于真实的污染源监控场景实际需求，依靠参与者上传的不完整数据，采用最大期望算法准确识别污染源。文献[11]针对群智感知网络中参与者上传数据的不完整问题，提出一种基于动态规划与信誉反馈的参与者选择方法，解决在参与者人数不确定的情况下如何选择参与者，从而提高感知数据质量的问题。

以上这些方法主要解决从用户采集和数据汇聚角度提高数据质量，并未充分应对因为用户参与主观意愿所导致的数据质量低下甚至虚假数据的问题。针对该问题，有两类方法解决感知数据的可信性问题，分别是可信平台模块和信誉系统。可信平台模块在用户的移动感知设备中设置专门的硬件模块，保证用户感知和上报到数据中心的数据是由真实的、授权的感知设备所采集。文献[12]基于“安全数码相机”的思想，利用 MD5 算法和基于随机数的加密算法设计了一个图像篡改检测的可信平台模块方法来保障用户上传图像数据的真实性。信誉系统用于评估和记录用户的历史感知数据的可信性，并将其用在未来的系统交互过程中，对于信誉度低的用户，优选其所采集的感知数据的可能性也会相应较低，同时会兼以相应的激励或惩罚措施[13]。此外，Mousa 等人总结了串谋攻击、女巫攻击、GPS 欺骗等 11 种恶意用户攻击方式[14]，为后续数据质量的提高提供了理论支持。

5.3.2　移动群智感知数据优选

对感知数据质量完成评估后，系统需要筛选冗余数据，提高分析效率。现有的许多工作都基于内容进行质量评估，分析群体与感知对象交互时产生的多维物理情境信息（如光强、加速度、拍摄角度等），提出群体数据质量评估模型。当前对移动群智感知数据质量量化评估方面的研究较少，大部分研究侧重于数据分析阶段使用数据挖掘[15]、机器学习[16]等方法识别和过滤异常数据实现数据优选。文献[17]从数据可信度出发，结合聚类和逻辑推理手段，提出一种识别错误或正确数据的方案，并实现虚假数据的过滤，从而实现数据质量的提升。Guo 等[18]提出了一种基于上下文感知的数据质量估计方法，通过历史数据训练上下文质量分类器，捕获上下文信息和数据质量之间的关系，据此设计激励手段引导移动用户的参与度和贡献度。

除数据分析外，也有的工作尝试通过消除冗余达到数据优选的目的。在这一方面，已有工作通过语义相似度进行冗余发现。语义相似度计算主要用于对两幅图像之间内容的相似程度进行打分，根据分数的高低判断图像内容的相近程度。具体地，语义相似度可通过时空情境信息、拍摄角度、远近等进行刻画。文献[19]通过计算其与图像库中所有图像的检索矢量的范式距离和欧里几得距离，得到与检索图像相似度值最高的一组图像数据并将其作为优质数据。Guo 等[18]提出基于分层金字塔树的冗余发现方法，每一层根据不同的约束阈值可以形成不同的分支。某层分支涵盖的数据代表该层以上特征联合聚类的结果。该方法能根据数据流和任务的语义约束，在线构造分层金字塔树，实现满足多维覆盖的群体感知冗余数据分组。此外，Uddin 等[19]研究了灾后现场在容延网络环境下的传输问题，在数据上传前根据时空和内容相似度约束进行照片选择，提高了群智感知移交效率。Wang 等[20]提出一系列摄影采集规则，实现对群体贡献视频数据的融合和集成。Tuite 等[21]通过群体感知

收集建筑物照片，用于城市 3D 建模，其主要思路是采用一种游戏的方式，训练“玩家”成为采集照片的“专家”，使“专家”们能够从不同的角度高密度地采集城市中建筑物的照片，实现对感知数据对象的多角度覆盖。Kawajiri 等[22]采用动态激励机制提高感知任务不同侧面的覆盖。

5.4 代表性工作：视觉群智感知与数据优选模型

文献[23]针对移动群智感知场景下感知节点的异构和网络的弱连接性带来感知数据不完整、不均匀和不可靠问题，解决数据采集和收集阶段的挑战，重点针对视觉群智感知任务（图像类数据）收集过程中的独特问题并提出可行的解决方案，高质量地收集感知数据。该文针对视觉群智感知数据的异构特征、动态数据流和任务多样化等问题，提出自适应任务约束的数据流聚类方法，实现高效率、高精度逼近数据选择最优解的目标。同时，考虑在客户端消除冗余数据，以有效降低存储空间和流量的消耗。为此，提出一种基于塔形树的照片信息存储结构 PTree，以及基于 PTree 的照片流数据聚类方法和数据选择方法。该方法适用于不同的感知任务，并且能够快速甄别冗余图像数据，从而降低由于参与者的移动性而造成网络通信中断所带来的数据丢失风险，提高数据收集的可靠性。

5.4.1 视觉群智感知数据优选模型

面向移动群智感知的照片不同于传统意义上只关注图像信息的照片，照片数据除了图像文件外，还包括大量的传感器数据，这些传感器数据常被用于移动群智感知应用系统。人们在互联网中（特别是移动社交网络）分享了大量照片，机会式感知利用这些照片及附带的信息（如作者、拍照地点、拍照时间、标签等）完成感知任务。此时，分享照片的人成为无意识地完成感知任务的参与者。而参与式感知通过招募参与者采集数据完成感知任务，参与者必须按任务要求使用定制的 App 拍摄特定对象的照片。此时，拍摄照片的人成为专门完成感知任务的参与者。不同于传统意义上的高质量照片，本文提及的高质量照片数据需要满足两个标准：①满足任务需求的数据。②低冗余的数据。

移动群智感知的数据集对感知目标的覆盖度越高且数据集的规模越小，数据集的质量越好。导致数据集质量降低的两个主要原因是低质采集和重复采集。移动群智感知中导致低质采集的原因分为非主观原因和主观原因。非主观原因是指：①任务的描述存在二义性，即参与者和任务发布者对人物的理解有偏差。②针对多样化的感知目标，参与者的感知行为难以规范化，其主观原因是指参与者敷衍完成采集任务。

移动群智感知中产生冗余采集有两个原因：①移动群智感知为保证数据集对感知目标的覆盖度，参与者实际雇用量往往大于理论雇用量。②陌生的参与者之间是一种“隐式”协作关系，因而不同的参与者就可能采集到相同或相似的照片。

以图像感知数据为例，其冗余关系较为特殊，如图 5-2 所示，照片 A 和照片 B 部分相似，照片 B 和照片 C 部分相似，但照片 A 和照片 C 不相似，也就是相似关系不具有传递性。原始照片和照片之间的相似关系可以用图表示。依相似度进行聚类时，如果以照片 B 为聚类簇中心，那么照片 A 和照片 C 就可能归入同一个簇，这显然是错误的。所以，为了保证对感知目标的最大覆盖，这里使用图的最大独立集（max independent set）作为数据选择的最

优解。例如，图 5-2 中有 4 个建筑物，如果选择照片 A 和 C，则使用两张照片即可覆盖所有建筑物。照片是否能够被选中依赖于该照片是否同时满足任务需求，即参与者是否按照要求拍摄了正确的照片，并且该照片不能由其他照片代替。

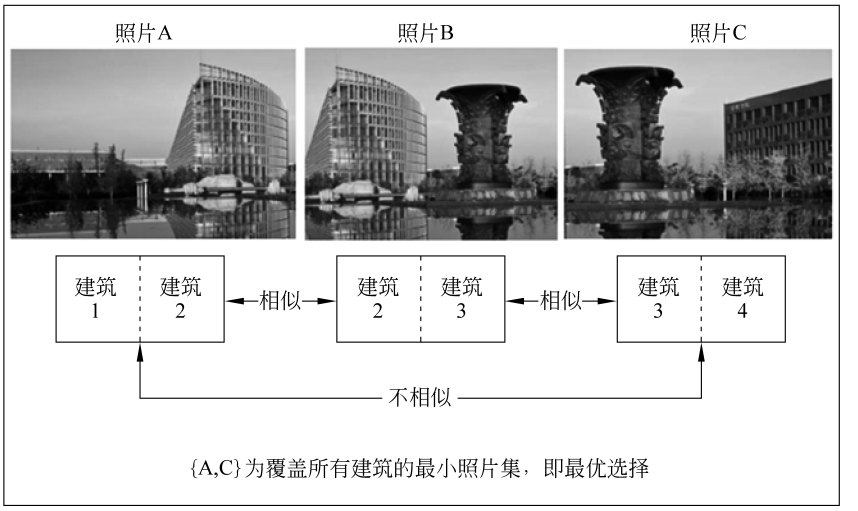

图 5-2　移动群智感知数据冗余示例与数据过滤理论基础

为了能够及时收集数据，并降低收集到冗余数据的概率，工作者在提交照片后，后台服务器需要快速判断照片的价值，并决定是否留用该工作者上传的照片。利用照片的情境信息可以粗略判定两幅图像是否相似，而照片的情境信息（如时间、地点、拍摄角度）由低维异构的数据组成，因此需要研究面向低维异构大数据的数据聚类方法，从而快速评估新数据的价值。

1. 交互式数据移交

群智感知受分布式采集模式的影响，冗余数据常常是在工作者不知情的情况下产生的，所以，通过工作者与平台之间的交互，可以达到对数据集进行前置选择的目的，从而降低感知成本。

采用前置选择方式的交互式选择移交的流程如图 5-3 所示：①客户端 App 首先向感知平台的服务器发送感知数据的基本项、情境项和缩略图，这些数据大约 10KB。②服务器根据这些数据判断该感知数据是否与数据中心的其他数据相似，如果相似，则拒绝上传感知数据的完整图片，如果不相似，则请求客户端 App 发送完整图片。③服务器计算应该为该数据支付的报酬，包括拍照报酬和流量补贴两部分，并告知工作者。

为了利用前置选择机制，大量的移动群智感知任务需要实时对拟上传的数据进行价值评估，即该数据是否值得收集。移动群智感知的数据价值体现在该数据是否能够对整个数据集的价值做出贡献。移动群智感知数据由异构的数据项组成，分两步评估一个数据的价值：①根据移动群智感知的任务约束，评估数据的质量，即是否满足任务要求。②在已收集到的数据中检索可以被替代的数据（可相互替代的数据对数据集的价值贡献是相等的，即相似数据）。为了适应不同的任务，数据相似关系计算采用参数化的函数。如前面所述，移动

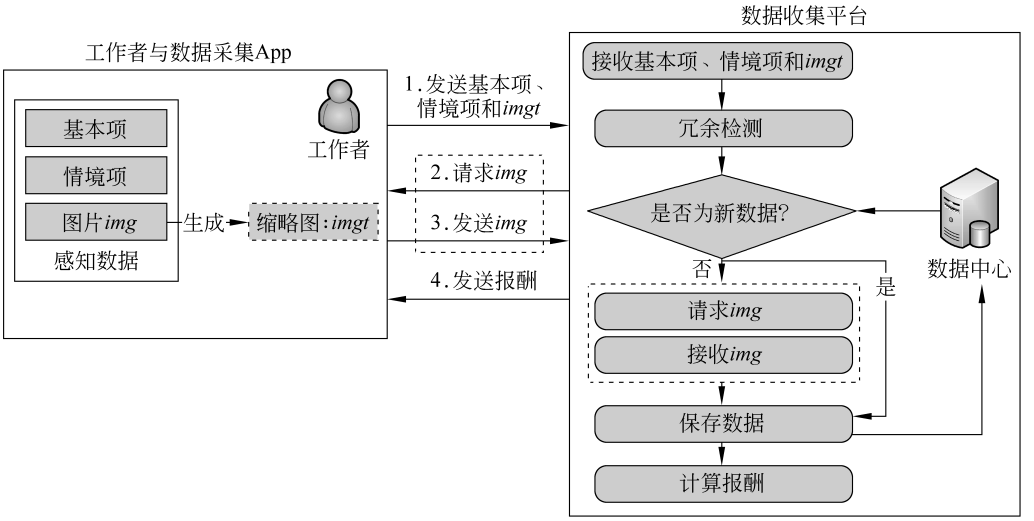

图 5-3　移动群智感知的前置数据选择工作流程图

群智感知数据的相似关系不具有传递性，所以图的最大独立集为最优解。为了合理聚类，生成最大独立集，下面研究针对动态数据流的逼近最优解的数据聚类和数据选择方法。具体来说，根据数据之间的相似关系，对动态的数据流进行聚类，使相同聚类簇中的数据对整个数据集的价值贡献相等，然后从不同的簇中选择数据。

2. 基于感知覆盖度的数据集质量评价方法

假设原始数据集的数据相似关系矩阵为$\boldsymbol{\lambda}$，$\boldsymbol{\lambda}_{i,j}=1$ 表示照片 P_i 和 P_j 之间的相似关系为真，此时，P_i 和 P_j 对优选后的感知数据集合的价值贡献是相等的，即 $Q((O-\{P_i\})\cup\{P_j\})=Q((O-\{P_j\})\cup\{P_i\})$。

为了及时向参与者支付报酬，并判断数据是否有保留的价值，在每个数据到达数据中心时，任务平台需要及时对数据进行价值评估，并支付报酬。若 $Q(S)$ 表示 S 中数据的总价值，那么如果 $Q(S\cup\{p\})>Q(S)$，则数据 p 值得保留。如果数据子集 S 中的数据两两都不相似，那么这个子集中的数据对任务都是有价值的，增加任何一个与子集中的元素相似的数据都是无意义的，但是，如图 5-2 所示，我们更希望得到对感知对象覆盖最全面的最小数据子集，即高质量的数据集，这个问题可以转化为求最大独立集。

假设某一任务的感知数据集为 $P=\{P_1,P_2,\cdots\}$，数据上传至数据中心的时间为ta_1，ta_2，…。由于在任务结束之前，P 一直增加且时间无法预知，因此 P 为数据流。由于群智感知是分布式数据采集，受移动通信速度的影响，采集时间 ts 和上传成功的时间 ta 并不同步，所以，及时判断数据价值非常重要。移动群智感知任务的时间跨度差别很大，短则几十分钟（如事件感知[22]），长则数月（如环境污染[24]）甚至数年（如生物科学研究[25]），所以我们无法每次都能够对完整的数据集进行分析。

3. 基于任务约束的语义相似度计算方法

为了使计算机能够判断两张照片是否相似，本章将在感知任务中定义的照片相似度判定条件称为任务约束条件。根据任务约束条件，可以从两方面衡量数据的相似度：图像相

似度和语义相似度[26]。

1）图像相似度

图像相似度指采用传统的图像相似度计算方法计算图像之间的相似程度，从而判定两张图片是否相似。提取图像特征的方法很多，如 SURF（Speed-Up Robust Features，加速健壮特征）、SIFT（Scale-Invariant Feature Transformation，尺度不变特征变换）、颜色直方图、边界检测，然后计算特征之间的距离，如欧氏距离和 KL 距离（Kullback-Leibler Divergence，相对熵）。

2）语义相似度

本章采用语义相似度识别冗余数据，数据之间的语义相似度度量以任务约束为基础。由于群智感知数据的特征是异构的，因此采用式(5-1)的布尔函数计算。

$$S(P_i,P_j)=\bigwedge_{k=1}^{n}(dist_k(p_{i,k},p_{j,k})\leqslant cth_k) \tag{5-1}$$

式中：P_i——感知数据，有 n 个特征；

P_j——感知数据，有 n 个特征；

$p_{i,k}$——感知数据 P_i 的第 k 个特征；

$p_{j,k}$——感知数据 P_j 的第 k 个特征；

$dist_k$——第 k 个特征的距离计算函数；

cth_k——在任务中定义的针对第 k 个特征设定的相似度阈值。

式中，当函数值$dist_k(p_{i,k},p_{j,k})\leqslant cth_k$时，表示两个数据的第 i 个特征是相似的。当所有的特征计算得到的结果都为“真”时，两个数据相似关系成立。函数$dist_k$的意义是由第 k 个特征决定的。若第 k 个特征表示定位，则$dist_k$计算地理距离；若第 k 个特征是图像，则$dist_k$即图像相似度。

5.4.2　基于 PTree 的高质量数据选择方法

两个感知数据的相似度是由式(5-1)计算的，根据布尔函数的特性，当有一项为假时，无论后面几项是真或是假，表达式的结果都是假。如图 5-4 所示，4 个感知数据依次到达，这里使用树的形式表达计算过程和结果，虚线框内为数据特征之间的关系，其中因为$agl_1\neq agl_2$，所以img_1和img_2无须比较而被省略。由于无法确定后来的数据是否与前面的数据有相似关系，因此，所有的数据都需要自顶向下保留所有的属性，由于第 $n+1$ 层的结点数不小于第 n 层的结点数，所以这棵树的形状像佛塔，被称为“塔形树”（Pyramid Tree，PTree）。

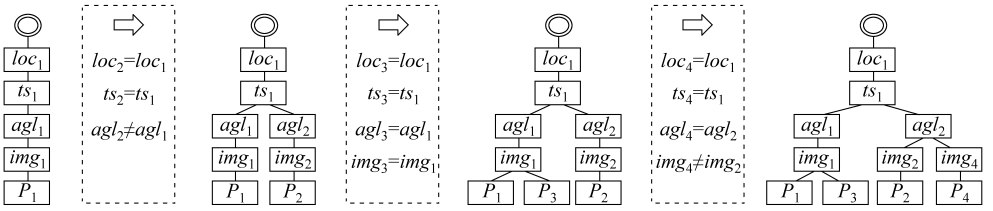

图 5-4　感知数据的相似度计算和聚类过程示例

如果群智感知数据包含 d 维，那么塔形树包含 $d+2$ 层，其中根结点在第 0 层，叶子结点在第 $d+1$ 层，其他层结点称为非叶子结点，每个叶子结点分别对应一个感知数据，每个非叶子结点对应了数据基于某特征的聚类。如图 5-5 所示，塔形树的 $1\sim d$ 层分别对应感知数据的一维属性，从根结点至叶子结点的分支上的结点则代表分支上的数据对感知对象

某一侧面的覆盖。非叶子结点代表了所有子结点的某一个共同特征，如结点$N_{1,1,1}$代表感知数据$\{P_1,P_3,P_5\}$的某个特征是相似的，而结点$N_{1,1}$则代表数据$\{P_1,P_2,\cdots,P_6\}$的某个特征是相似的。所以，第i层的结点的叶子结点数据组成第i层的微簇。

1. 基于 PTree 的数据选择算法

PTree 的每个分支都代表一个数据或一组相似的数据，分支的非叶子结点代表了该组数据的各个特征，多个分支之间的部分重合关系表示叶子结点里的数据的部分特征是相似的，因此，由不同非叶子结点组成的分支上的数据是不同的。如图 5-5 所示，感知数据 P_1 和 P_3是重复的，但 P_1 和 P_2 是不重复的。

在服务器端，每个任务可以对应一个或多个 PTree，PTree 初始为空，当服务器接收到感知数据后，PTree 开始生长，生长的过程中如果一个感知数据使 PTree 长出新的非叶子结点，则表示该数据为新数据，即高价值数据，因此，根据图 5-5 所示的交互式移交模型，所有在兄弟结点中排行老大的叶子结点（即最先到达）最终被选中。

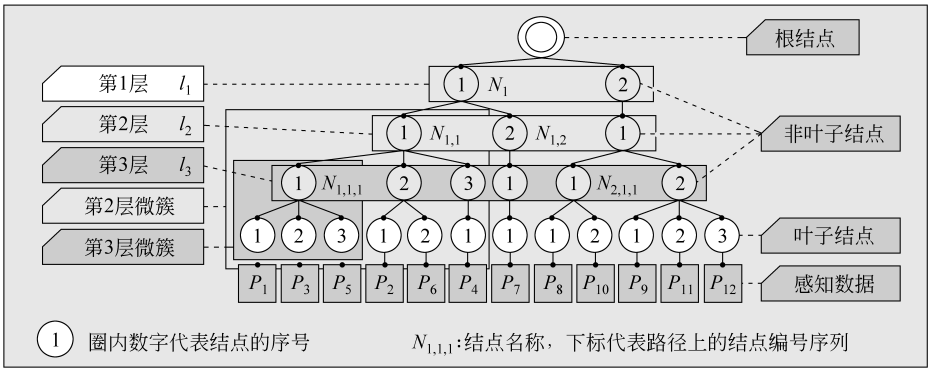

图 5-5　PTree 的基本结构与定义示例（此处 $d=3$，感知数据集为$\{P_1,P_2,\cdots,P_{12}\}$）

PTree 的生长过程示例如图 5-6 所示，每个数据对应一个叶子结点，PTree 的生长相当于判定在哪个非叶子结点处分叉。

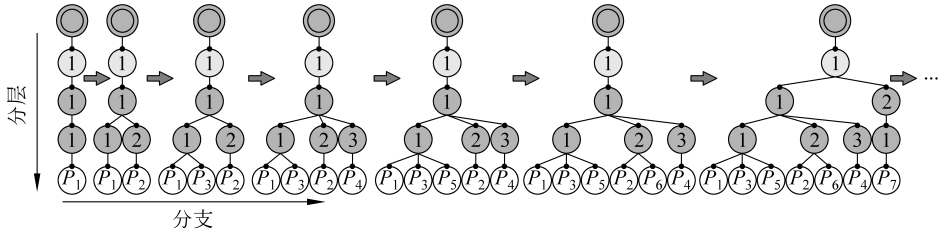

图 5-6　PTree 的生长过程示例

匹配结点是一个非叶子结点，如果一个非叶子结点是一个感知数据的匹配结点，那么该数据必须能够包含在该非叶子结点构成的微簇中。任务约束被用来查找匹配结点，而找不到匹配结点时，PTree 产生分支，因此，任务约束可以看作 PTree 的分支参数。

以图 5-6 为例，PTree 生长过程如算法 1 所示。PTree 的生长过程就是不停地为数据流的最后一个数据在已构造的 PTree 中寻找匹配结点。树的分层、结点的属性和匹配结点的查找方法都对 PTree 的构造造成影响，也就是对重复数据的识别结果有影响。

算法 1　PTree 生长算法

输入：数据流 P
输出：塔形树 PTree
for each $p \in P$ do
　N＝PT 的根结点；
　//从根结点开始，自顶向下广度优先搜索数据 p 的匹配结点
　While N 不是叶子结点 do
　　从 N 的直接后继子结点中检索数据 p 的匹配结点 M
　　if M 被找到 then
　　　$N \leftarrow M$；
　　else if M 没有被找到 then
　　　自结点 N 处开始构造分支
　　　break；
　　end if
　end while
　构造 N 的兄弟叶子结点，在结点内保存数据 P
end for

2. 树的分层与树结点的属性

由于 PTree 的分层会影响计算效率，为了适应数据流的特性以及塔形树与分层配置有关的特性，PTree 需要采用动态的分层配置。假设移动群智感知平台能够支持的感知数据的特征数为 $\boldsymbol{F}$，某任务收集的感知数据的特征集表示为 $F=\{f_1, f_2, \cdots\}$，PTree 的层集为 $\boldsymbol{L}=\{l_1, l_2, \cdots\}$，$|F| \leqslant |\boldsymbol{F}|$，$|F| \leqslant |\boldsymbol{L}|$，那么 PTree 的分层规则可以定义为 $F \rightarrow L$ 的映射关系，表示为集合 $LM=\{(f_i, l_j) | f_i \in \boldsymbol{F}, l_j \in \boldsymbol{L}\}$。

PTree 的生长是感知数据与 PTree 的对比和查找匹配结点的过程，树结点的参数需要支持这个查找过程，因此，基于感知任务的定义和感知数据的定义，PTree 的非叶子结点的数据结构见表 5-1。不同的非叶子结点可以将数据分为不同的微簇，该微簇的中心值是依据微簇内的数据对应的属性值计算的。如表 5-1 所示，在定位层，微簇中心以 *nloc* 表示，与此类似，参数(*ntl*，*ntu*)、*nagl* 和 *nimg* 都表示不同层的微簇中心属性。由于并不是所有的感知数据的属性都可以进行算术运算，如图像，所以本文采用不同的方法计算微簇的中心值。

表 5-1　PTree 的非叶子结点的数据结构

层	对应的感知数据项	对应的感知任务项	树结点的属性
任务	*tid*	*tid*	(*ntid*)
定位	*loc*：(*px*，*py*)	*c_loc*	(*nloc*，*c_loc*)，*nloc*：(*px*，*py*)
时间戳	*ts*	*c_tm*	(*ntl*，*ntu*，*c_tm*)
拍照方向	*agl*	*c_agl*	(*nagl*，*c_agl*)
图像	*img*	*mt_is*，*th_is*	(*nimg*，*mt_is*，*th_is*)