信息系统中模糊数据聚类分析

□文/吴丽娜

由于计算机网络技术的迅猛发展，客户/服务器技术、大型分布式数据库系统软件的逐步成熟，管理信息系统正向着分布式数据库和分布式处理方向发展，因而计算机网络成为大型信息系统的主要支撑环境。目前网络设计的范例很多，但是纵观这些规划和设计，尤其是在信息系统的网络环境规划和设计中，很少有建立在充分的需求分析和数据库逻辑设计的基础之上，多数是以对象单位职能机构科室的区域划分为前提进行信息系统的网络环境规划和设计，不能真实的反映对象信息系统的内在信息处理和信息流动关系。显然，这样的信息系统网络环境物理设计带有一定的盲目性，使开发成的信息系统在不合适的网络物理环境下运行，以至整个信息系统的运行效率低下，甚至造成网络的频繁阻塞，使整个信息系统瘫痪不能运转。因而，信息系统网络环境逻辑设计作为一个大型信息系统开发的重要环节，其设计应建立在坚实的理论基础之上。

一、对象系统的分析

1、通过信息系统需求分析中生成的输出报表数据和需求报表数据以及信息定义表中获取关于信息的类型、产生此信息的频度等数据，然后生成一个数据属性表。

2、通过信息系统需求分析，获取有关应用功能访问、创建有关信息的数据，生成一个应用属性表。

二、应用和信息关联属性的量纲和数量级的标准化

通过数据属性表和应用属性表得到应用和信息的关联矩阵表。由于应用和信息的自身属性和它们之间的数据传递属性非常复杂，并且它们的量纲、数量级以及它们之间数据传递量的数量级变化幅度差异很大，如果用原始数据进行聚类分析有可能会突出某些数量级特别大的变量，而压低甚至排除某些数量级小的变量对分类的作用。

为了有利于分析、对比和分类清晰，常对原始的信息、应用及它们之间的关联进行一些必要的处理和变换，这样就可以消除量纲的不同和使第一变量都统一在某种共同的、相对均匀化的数值范围之内，为此通常在聚类分析之前对数据进行标准化处理，一般有两种方法：

（一）针对信息和应用的数据传递量对应用和数据的关联进行标准化处理。1、计算应用到数据的传递关系；2、对传递量进行标准化。

（二）针对应用对信息的创建、使用关系对应用和数据的关联进行标准化处理。管理信息系统中的所有应用都要和信息打交道，所以仅从应用和数据间的创建数据、使用数据及和与数据无关的角度出发考虑应用和数据的关联性，忽略掉应用对数据传递量的关系，通过引用数据属性表和应用属性表，可以将这种应用和数据的关联矩阵中的传递量直接量化为0、1或2。这种标准化处理较为简单，常被采纳。

三、利用应用/数据聚类分析法进行信息系统子系统划分

（一）对管理应用进行的聚类分析。通过对应用进行聚类分析，把经常访问相同数据的应用聚成一类（应用簇），这样可以使应用访问外部数据的次数大大减少，同时使原本分散的数据访问集中起来。因为各个应用之间都是通过访问相同的数据来表示它们之间的相关性，所以可以通过计算在应用之间的相似系数作为它们聚类分析的依据。

应用聚类步骤如下：

1、首先从对角距离矩阵中找出最大值 d_{ij}；

2、将应用 A_i 与应用 A_j 合并成新的应用 A_t，然后重新计算 $r_{tk}=Max\{r_{ik},r_{jk}\}$；

3、更新应用数据关联矩阵，更新对角距离矩阵；

4、重复1，直至对角距离矩阵的应用合并为一个大类。

经过上述各步生成应用聚类谱系图，其中横坐标为应用集，纵坐标为它们合并时的相似系数，通过经验定一个阈值，将该阈值以下的已经被合并的应用聚为一个应用簇，蓁的为松散偶合合并为一个应用簇，于是可将各个应用 $A_1,A_2\cdots A_n$ 聚为若干应用簇 AC_1,AC_2,AC_k（k<n）。

（二）对数据的进行聚类分析。数据之间的相关性是通过它们被相同的应用进行访问表现出来的，所以可用数据之间的距离系数来作为它们进行聚类分析的依据。

距离系数为 $I_{kl}=\Sigma Min\{r_{ik},r_{jl}\}/\Sigma r_{ik}$

其中 Σr_{ik} 表示数据 k 被所有应用访问的总数，$\Sigma Min\{r_{ik},r_{jl}\}$ 表示数据 k、数据 L 同时被应用访问的总数，I_{kl} 表示数据 k 与数据 l 的远近程度。

数据聚类步骤如下：

1、首先分配规模最大的数据作为树根。

2、找出一个数据 L 作为树中的一个节点 k 的子孙，且使 I_{kl} 最大，如果这个数据 L 同时使几个结点 k 的 I_{kl} 一样大，则使数据 L 作为以上几个结点 k 中分配规模最小的结点的子孙。

3、判断是否所有的数据都被联到了树上。

4、如果所有的数据都被联到了树上，就结束，否则就重复②。

通过上述各步生成数据块距离包含树，设置一个适当的相似值区间，把父亲结点和儿子结点之间的距离包含系数 I 在区间内的数据结点聚为一类，于是可将各个数据块 $D_1,D_2\cdots D_n$ 聚为若干数据簇 $DC_1,DC_2\cdots DC_t$（t<n）。

（三）利用相似距离聚类分析结果进行信息系统子系统划分。通过以上相似性、距离聚类分析的结果，重排应用数据关联表，可以生成一个应用簇和数据簇的关联。因为有些应用簇和数据簇之间的关系非常的松散，所以可以将这些松散的关系忽略掉。因为任何一个应用簇和数据簇都对应着一个本簇内应用和数据关系的关系矩阵（即应用簇和数据簇的关联矩

企业公司治理问题

□文 / 葛苏君

提要 公司治理的核心在于建立一系列制度安排，以提高企业决策效率，从而提升企业绩效。这种制度安排可分为基于代理理论、产权理论的公司内部治理和以市场、竞争为核心的外部治理。不同国家由于历史、文化、制度等因素的差异，企业融资结构各具特色，所采用的公司治理模式也不尽相同，但大体上可分为股权分散型、股权集中型和家族控制型等几类。

一、文献综述

自 Berle 和 Means 于 1932 年首次提出“所有权与控制权相分离”的论点以来，公司治理研究蓬勃兴起。20 世纪七八十年代，公司治理问题的研究集中于 Jensen 和 Mechling 等人运用委托—代理理论，以美国公司为对象，致力于解决股权分散化下“弱股东、强管理层”的问题。自 20 世纪九十年代起，公司治理的研究主要集中在所有权结构和控股股东（大股东）。因此，现代公司治理研究认为：公司治理的研究重点不应仅局限于外部股东和内部管理层之间的代理问题，而应更多地关注“强大股东，弱小股东”之间的利益冲突。如果把早期基于所有权分散的 Berle 和 Means 对公司治理的研究称为传统公司治理研究，那么近几年来，以 La Porta、Lopez-de-Silanes、Shleifer、Vishny（简称 LLSV）为代表，以所有权集中和控股股东为基础的研究则使得公司治理研究跨入了一个“革命化”的新阶段，成为公司治理研究的最新趋势。

（一）委托人——代理人问题。信息经济学针对新古典经济学完全理性和完全信息两大假定，提出人是有限理性的；同时由于搜寻成本的存在，信息是不完全的，并且信息的分布在个体间是不对称的。由此，该学派提出了“委托人—代理人问题”及由此产生的激励制度设计。在现代企业中，由于资产的复杂性和产权联结的复杂性，企业内部往往形成多环节、多层次的委托代理结构，即委托代理链。委托代理链主要涉及两大内容：一是委托代理主体本身的界定，二是各委托代理主体间的相关关系。各级委托人和代理人的利益和权利要通过契约来界定，这种界定是否有效则取决于产权的划分是否明晰以及契约的完备程度。契约的完备程度又进一步取决于信息是否完全、信息的分布是否对称。现实世界中信息不完全和信息分布的不对称，决定契约是不完全的，其中必然存在漏洞，这就需要制定一定的激励制度以降低不确定现象下代理人败德行为给委托人带来的损失。

（二）协作群生产假说与“状态依存所有权”理论。随着对产权结构研究的深入，研究者开始关注“如何计量投入的生产力”和“如何计算投入的报酬，使之等于其边际生产力”这两个企业内部的问题。针对以上两个问题，艾尔奇安和德姆塞茨提出了协作群生产假说。该理论认为，现代企业的生产关系是各种要素所有者之间的协作关系，各个所有者作为一个群体中的一员出现在生产过程中。这是难以观察和计量每个要素贡献大小的技术原因。在某些特定情况下，债权人、员工、管理者都有可能在事实上成为公司的所有者。从这

阵的子阵），通过对应簇和数据簇内部关系矩阵的元素求平均值可以得出应用簇和数据簇的关系度量矩阵。

通过对上述的关联数据分析，设置一个关联的阈值，将此阈值以下的应用簇和数据簇关联值忽略（标记为空白），将应用簇和数据簇关联值大于阈值的设置为 Primary（标记为 P），阈值的设定必须满足任何一个应用簇仅能有一个数据簇和它具有 Primary 关系。

最后将每一个数据簇与所有和该数据簇具有 Primary 关系的应用簇合并为一个子系统，这样就可以将整个管理信息系统划分成各个业务子系统。

（四）子系统间数据流量的计算。根据管理信息系统的需求分析调查表统计各个子系统之间的数据流量。任何两个应用和数据之间的数据流量计算如下：

D_{ij}=应用 i 访问数据 j 的数据量×访问频度

各个子系统之间的数据流量 Z_{ij} 为：子系统 i 的所有应用对子系统 j 的所有数据的数据流量之和以及子系统 j 的所有应用对子系统 i 的所有数据的数据流量之和。于是，生成各个子系统之间的流量矩阵。

四、信息系统网络环境逻辑设计

通过信息系统的子系统划分图我们可以将各子系统的数据和应用放入相应的数据库服务器、应用服务器或数据/应用服务器，然后通过子系统之间的数据流量矩阵生成子系统流量无向图，其中以各子系统为子系统流量无向图的结点，各子系统之间的数据流量作为子系统流量无向图的边。Z_{12}，…，Z_{kn} 表示子系统之间的数据流量，也即子系统流量无向图中的各个边的权值。

这样生成的流量无向图反映了包含所有子系统接点的主要连接关系，用它可以作为今后网络环境物理设计的主要参考依据。如果满足以上依据，并且把流量无向图的所有边的权值（也就是各个子系统之间的数据流量）都乘以-1，再计算该流量无向图的最小生成树，则该最小无向图就能够最大的反映出各个子系统的流量关系和连接关系。

通过这个最小生成树来确定主干和拓扑结构，然后再进一步划分二级网络逻辑结构，然后将这些主干结点作为信息系统的网络逻辑主干，自然网络主干上的数据流量也就最大。这样从需求分析过渡到信息系统网络环境物理设计就有了一个坚实的理论基础。至于应用服务器和数据服务器的硬件选型可以对它们的原型——应用簇和数据簇进行量化，然后划分一些等级，并把这些等级作为它们硬件选型的依据。■