基于 Isomap 的 CWME 专家群体意见分析系统

杨伟杰, 戴汝为, 崔 霞
(中国科学院 自动化研究所 复杂系统与智能科学重点实验室, 北京 100080)

摘 要: 针对从定性到定量综合集成研讨厅体系, 对复杂问题分析过程中的专家群体的分散化思维, 提出并实现了一种基于自动调查问卷的群体意见相似性定性分析方法。此方法基于群体行为的相似性特征对调查问卷的相似性度进行量化统计, 并采用 Isomap 算法将问卷量化之后产生的多维数据集进行降维及可视化表示。

关键词: CWME; 群体意见相似性; Isomap 降维; 综合集成研讨厅
中图分类号: TP31 **文献标志码**: A **文章编号**: 1001-3695(2008)10-3188-03

Isomap-based experts' opinions evaluation system in CWME

YANG Wei-jie, DAI Ru-wei, CUI Xia
(Key Laboratory of Complex System & Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: There is the dynamic complexity in experts groups for man's social being determines his consciousness in the course of solving complex problem high related to OCGS. This paper gave a qualitative analysis method based on questionnaires, namely group coherence qualitative analysis method. Using the Isomap algorithm this method realized the data's dimension reduction of the opinions which the experts evaluated to argue scheme. At the same time it visualized the results in the low dimension space, so provided an ideal environment of the visualization for the hall for workshop of metasynthetic engineering, which would be clear to see the state of the clustering of the group of the experts' thinking.

Key words: CWME(cyberspace for workshop of metasynthetic engineering); coherence of group thoughts; Isomap dimension reduction; hall for workshop of metasynthetic engineering

综合集成研讨厅体系, 是思维科学的一项应用技术[1~3]。建立起能够激发出个人创造性的群体环境, 获取 1+1>2 的整体和综合优势。综合集成法的创建者早就指出, 与开放复杂巨系统有关的非常复杂的问题不是个人所能解决的, 必须依靠集体的智慧, 综合集成所有可以利用的知识资源, 发挥整体效应。同时这也就存在一个问题: 面对一个复杂问题, 由于群体中的个体成员对包含在复杂问题中各种因素的重要性的感知不同, 看法也各不相同, 这就是研讨厅中的专家群体分散化思维的表现[4]。

克服专家群体的分散化思维, 使参加研讨的专家群体的思维最终达成一致, 这就需要一个把握群体意见状态的工具, 而调查问卷就是其中一种有效的手段。它对发现和界定未知或模糊的复杂问题和现象具有启发作用。但是, 随着传统调查方法拒访率不断提高, 调查质量也在逐渐下降, 促使许多市场研究公司开始纷纷转向网络, 利用互联网为手段进行的市场调查已在我国悄然兴起, 其形式则是以网络调查问卷为主。现存的调查问卷系统已经很多, 问卷的管理、生成、分发功能基本完善, 但是在问卷结果的统计方面却存着很大的缺憾。因而本文基于综合集成研讨厅群体有效交互与组织, 融合群体行为相似性研究成果, 设计并实现了一种自动调查问卷系统, 它不仅可以完成一般问卷系统的基本功能, 并且针对问卷结果分析实现了多种定性、定量分析方法的融合, 对不同问题使用不同的分析方法, 并将调查问卷的结果可视化。本文系统最主要的贡献是: 针对调查问卷定性统计分析的难点, 即问卷的量化问题, 提出并实现了一种基于群体行为的相似性特征对调查问卷进行量化统计的方法, 并应用非线性机器学习算法对其进行可视化显示, 实现了对研讨厅中专家群体分散化思维导致的结果差异的定性分析。

1 基于 Isomap 算法的问卷结果定性分析方法

将专家的定性认识通过建模的方法来提高到定量的认识, 专家分析其结果, 不断修改模型或其参数直至达到一个满意的结果, 是从定性到定量综合集成思想的具体体现[5]。针对这一过程, 将专家对研讨方案评价意见的多维数据组进行降维, 并将降维后的结果在低维数据空间中进行可视化表示。通过这种对专家意见在低维空间内的可视化表示, 可直观观测到专家群体思维的聚类情况, 然后根据聚类情况分析专家之间的行为相似性, 逐渐引导专家意见趋于一致。本文中同样体现了这一点。因为调查往往是在一个群体内进行的, 而一个群体中的个体之间一定存在相似性, 只要能找到对这种相似性的合理描述, 就能引导群体意见达成一致。因而笔者提出了如下的方法: 首先对问卷结果进行统一的量化表示得到高维表示向量, 然后对此高维向量进行降维处理, 最后对降维之后的低维向量进行可视化表示, 得到个体之间意见相似性关系。

1.1 问卷结果的量化

在综合集成研讨厅中, 专家发表对方案的评价意见可以通过填写调查问卷来完成。其中选择题型调查问卷是目前市场

收稿日期: 2007-11-24; 修回日期: 2008-01-09

作者简介: 杨伟杰(1980-), 女, 山东潍坊人, 博士研究生, 主要研究方向为人工智能、模式识别、智能系统(weijie.yang@ia.ac.cn); 戴汝为(1932-), 男, 云南昆明人, 中国科学院院士, 研究员, 主要研究方向为自动控制、模式识别、人工智能、复杂性科学及思维科学; 崔霞(1975-), 女, 山东莒南人, 副研究员, 主要研究方向为人工智能、模式识别、智能系统.

调查中一种常用调查问卷,一份调查问卷有多道选择题,每道选择题有多个选项,可以单选,也可以多选。所以仅仅通过查看调查问卷的填写结果,难以区分专家之间观点上的差异。人们期待能用一种更加直接的方式来反映专家个体之间观点上的差异,如可视化数据显示方式。为了实现调查结果的可视化,首先要将问卷调查结果用数学形式量化表达出来。本系统采用的量化方式如下:首先统计一份调查问卷中所有的选项数,设定一个维数与该选项数相等的高维向量。该高维向量每一维的分量对应一个选项的选择情况,如果该选项被选中,则该分量置 1,否则置 0。这样每一份调查问卷都与一个高维向量相对应,同样一个高维向量也表示一份调查问卷的填写结果。

1.2　问卷结果的降维处理

如前所述,一份调查问卷的调查结果可以用一个高维向量表示。但只有三维以下的向量才能在坐标系中表示,所以为了实现可视化,需要对高维向量进行降维。如何将大量高维数据所包含的特征结构在低维空间中展现,即发现隐藏在高维数据中有意义的低维结构,成了问题的关键。传统的降维技术主要有 PCA 和 MDS 方法等,用 PCA 和 MDS 算法实现数据降维的主要特点是映射的低维数据点可以较好地保持高维数据之间的差异性,两者可较好地发现隐藏在高维数据线性空间中的数据结构,但对于高维数据集隐藏的非线性结构处理结果并不理想。因而本系统采用了 Isomap 非线性降维算法。Isomap 算法不仅融合了 PCA 和 MDS 的主要算法特征——计算有效性、全局的优化性和渐近收敛性,而且可以比较灵活地学习到数据点的非线性结构。它根据高维数据集包含的非线性结构进行点的非线性扩展,用欧式距离较好地揭示了高维数据集的自由度[6,7]。

Isomap 降维算法具体实现如下:

a) 构造关系图。将每个高维向量用一个节点表示,如果两个向量间的距离小于预设值 ε 则连接着两个节点。遍历所有节点,形成一张图 G。其中,定义向量 X 和向量 Y 的距离为

$$d(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

b) 计算最短距离矩阵。在图 G 中,计算所有节点 i 到其他所有节点 j 的最短距离,组成距离矩阵 $D_G = \{d_G(i, j)\}$。其中 $d_G(i, j)$ 表示节点 i 与 j 之间的最短连线距离。

c) 构造降维向量。将矩阵 D_G 按下述方法变换为矩阵 $\tau(D_G)$。

$$\begin{cases} \tau(D_G) = -HSH/2 \\ S_{ij} = D_{ij}^2 \\ H_{ij} = \delta_{ij} - 1/N,\ N\text{ 为矩阵的维数} \\ \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \end{cases} \tag{2}$$

令 λ_p 为矩阵 $\tau(D_G)$ 的第 p 大的特征值,V_p^i 为 λ_p 所对应特征向量的第 i 个分量。则如果要将高维向量 y_i 降为 d 维向量 y_i^*,y_i^* 的第 p 个分量由下式确定:

$$y_{ip}^* = \sqrt{\lambda_p V_p^i} \tag{3}$$

变量 d 可以等于 1、2、3,这时降维向量分别为一维、二维、三维,可以在不同的坐标系中显示。

d) 任取一个三维向量作为坐标原点,对其他向量按下面公式作相应的坐标变换:

$$\begin{cases} x' = x - x^* \\ y' = y - y^* \\ z' = z - z^* \end{cases} \tag{4}$$

其中:(x^*, y^*, z^*) 为原点向量坐标;(x, y, z) 为除原点向量以外的普通向量的原始坐标;(x', y', z') 为普通向量变换后的坐标。根据各个向量的坐标点和它们的连线关系在三维坐标系中绘图。以下举例说明。

假设有甲、乙、丙、丁四个人填写了一份调查问卷,在这份调查问卷中,有三道选择题。其中前两道选择题分别有两个选项,最后一道选择题有三个选项,选择题允许多选。他们的选择结果是这样的:甲选择了第一个选择题的第一个选项、第二个选择题的第二个选项、第三个选择题的第二和第三个选项;乙选择了第一个选择题的第二个选项、第二个选择题的第一个选项、第三个选择题的第一个选项;丙选择了第一个选择题的两个选项、第二个选择题的第二个选项、第三个选择题的第三个选项;丁选择了第一个选择题的第一个选项、第二个选择题的两个选项、第三个选择题的第二个选项。接下来用向量来表示调查问卷的填写结果,因为整个调查问卷共有七个选项,所以用一个七维向量来表示一份调查问卷的填写结果。根据上面 b) 所述的方法,将甲、乙、丙、丁四个人的问卷填写结果分别用 A、B、C、D 四个向量表示,它们分别为

$$A = (1, 0, 0, 1, 0, 1, 1)^T \qquad B = (0, 1, 1, 0, 1, 0, 0)^T$$

$$C = (1, 1, 0, 1, 0, 0, 1)^T \qquad D = (1, 0, 1, 1, 0, 1, 0)^T$$

根据上述的欧式距离计算公式,计算各向量间的欧式距离如下:

$$d(A, B) = d(B, A) = \sqrt{7} \approx 2.6 \qquad d(A, C) = d(C, A) = \sqrt{2} \approx 1.4$$

$$d(A, D) = d(D, A) = \sqrt{2} \approx 1.4 \qquad d(B, C) = d(C, B) = \sqrt{5} \approx 2.2$$

$$d(B, D) = d(D, B) = \sqrt{5} \approx 2.2 \qquad d(C, D) = d(D, C) = \sqrt{4} \approx 2$$

经过实验,在本例中无须通过设定 ε 值来滤除连线,所以预设 ε 值为 3,按照 a) 所述方法建立关系图如图 1 所示。

接下来在关系图中计算每两个节点之间的最短连线距离。从 A 到 B 有两条连线,分别是 ACB、ADB,计算这两条连线的距离:

$$d_{\overline{ACB}} = d_{\overline{AC}} + d_{\overline{CB}} = 1.4 + 2.2 = 3.6$$

$$d_{\overline{ADB}} = d_{\overline{AD}} = d_{\overline{DB}} = 1.4 + 2.2 = 3.6$$

所以从 A 到 B 的最短连线距离为 3.6。

同理,可以计算出从 A 到 C 的最短距离为 1.4,从 A 到 D 的最短距离为 1.4,从 B 到 C 的最短距离为 2.2,从 B 到 D 的最短距离为 2.2,从 C 到 D 的最短距离为 2。并且容易看出,从节点 $i \sim j$ 的最短距离与从节点 $j \sim i$ 的最短距离是相同的。

这样计算得到最短距离矩阵

$$D_G = \begin{bmatrix} 0 & 2.6 & 1.4 & 1.4 \\ 2.6 & 0 & 2.2 & 2.2 \\ 1.4 & 2.2 & 0 & 2 \\ 1.4 & 2.2 & 2 & 0 \end{bmatrix}$$

根据上述步骤 c) 的计算方法,计算变换矩阵

$$\tau(D_G) = \begin{bmatrix} 1.645 & -1.5 & 0.425 & -0.57 \\ -1.5 & 2.115 & -0.78 & 0.165 \\ 0.425 & -0.78 & 1.165 & -0.81 \\ -0.57 & 0.165 & -0.81 & 1.215 \end{bmatrix}$$

求解矩阵 $\tau(D_G)$ 的特征值

$$\lambda_1 = 3.905,\ \lambda_2 = 1.541,\ \lambda_3 = 0.694,\ \lambda_4 = 0$$

求解与各特征值相对应的特征向量

$$v_1 = (0.584, -0.673, 0.363, -0.274)^T$$

$$v_2 = (0.19, -0.41, -0.511, 0.731)^T$$

$$v_3 = (0.611, 0.36, -0.579, -0.374)^T$$

$$v_4 = (-0.5, -0.5, -0.5, -0.5)^T$$

根据步骤 d) 中的公式,将 $A1B1C1D$ 降为三维坐标向量。

$A^* = (1.154, 0.235, 0.509)^T$ $B^* = (-1.329, -0.509, 0.3)^T$

$C^* = (0.717, -0.635, -0.497)^T$ $D^* = (-0.542, 0.908, -0.312)^T$

以 A^* 为坐标原点，变换 A^*、B^*、C^*、D^* 坐标为

$A' = (0, 0, 0)^T$ $B' = (-2.483, -0.744, 0.209)I^T$

$C' = (-0.436, -0.870, -1.01)^T$ $D' = (-1.696, 0.672, -0.82)^T$

最后在三维坐标系中标出各个坐标点，绘制出图 2。图中各个点之间的距离就代表了不同个体之间意见的相似性关系。

2 系统功能实现

此系统可以完成问卷的管理、自动生成、自动分发等基本功能，并且针对问卷结果分析实现了多种定性、定量分析方法的融合，对不同问题使用不同的分析方法，并将调查问卷的结果可视化。同时本文系统通过采用 Isomap 算法对调查问卷结果的量化统计，实现了对研讨厅中专家群体分散化思维的定性分析。系统实现流程如图 3 所示。

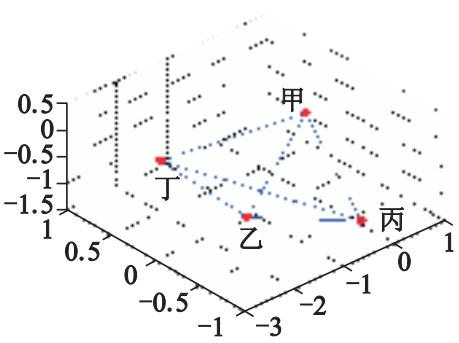

图2 Isomap方法结果

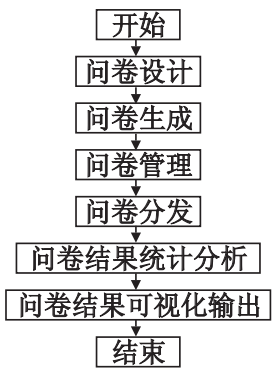

图3 调查问卷功能实现流程图

2.1 调查问卷的管理

在本系统中，问卷是以 XML 文件流的形式存放在数据库中，并以列表的形式在系统中提供了问卷管理功能，可以方便地增加、删除、编辑问卷。问卷管理工具界面如图 4 所示。

2.2 自动调查问卷设计

设计一份问卷包括一系列逻辑性的步骤，这些步骤由于调研人员不同而有所差别，但所有的步骤趋向于一个共同的顺序[8]即：

a) 提出分析架构，即首先是确定调研目标、资源和限制因素。这一步在本系统中是以设置问卷主题为主要任务，设置好主题，问卷中所有问题都围绕主题生成。

b) 决定调查方式和问卷形态，即确定数据收集方法。首先确定问题的回答形式。问卷设计中，问题的设计有不同的种类。其中开放性问题和封闭性问题是两个比较常用的种类[9]。开放性问题，就是不为受访者提供具体回答选项的问题，如简单回答问题等。封闭性问题，又称定选性问题，受访者从研究人员所给予的一个或更多的具体答案中选择回答，如单项选择题、多项选择题等。根据这两种问题的优缺点，本系统的自动问卷生成功能提供了以下几种问题类型模板，即单项选择、多项选择、简单回答、论述题和矩阵题。

c) 拟定问卷初稿、修订。本文设计的系统，在问卷创建者确定主题、创建问题及选项之后，系统自动生成问卷，并形成预览，直接可以看到分发状态的问卷形式，如果创建者不满意还可以继续进行修改。

d) 预试和评估。问卷创建完成，创建者可以在本地进行多次提交，然后查看多次调查结果的汇总，对问卷进行预试。如对产生的结果不满意可以根据需要继续修改问卷直至能够取得预期的效果。问卷设计工具界面如图 5 所示。

图4 问卷管理界面

图5 问卷设计界面

2.3 调查问卷的自动发放

系统采用了 Google 的 G-Talk 所采用的即时通信协议，即 XMPP 协议（也称 Jabber 协议），实现了问卷的自动发放。在研讨过程中，主持人对需要填写的问卷进行分发，专家电脑桌面自动弹出问卷，其间无须专家进行任何的操作。问卷发放界面如图 6 所示。

2.4 自动调查问卷结果统计分析

从统计分析的层次划分问卷的统计分析方法可分为两类，即定性分析和定量分析[9]。问卷的定量分析根据分析方法的难易程度可分为简单定量分析和复杂定量分析[10]。

简单的定量分析是对问卷结果作出一些简单的分析，本系统中对封闭问题的定量分析主要采用百分比、平均数、频数来进行分析。同时针对不同的被访群体之间进行比较则引入交叉分析来实现。对开放问题的定量分析主要采用分类处理之后用百分比之类的方法对各类型进行汇总。对于数量回答的定量分析，首先是对量化后的数据进行区间处理，然后使用位置测度、平均值、中位数等统计量来描述结果。

简单分析常用于单变量和双变量的分析，但是社会经济现象是复杂多变的，仅用两个变量难以满足需要。这时就需要用到复杂定量分析。在问卷设计中，常用的复杂定量分析有两种，即多元分析和正交设计分析。多元分析就是通过对观测数据的分析，由表及里来研究多个变量之间相互依赖的规律性，或根据实际问题的需要对研究对象作出某种评价、分类、判别，或从中发现各个指标都起作用的更一般的公共因素。在问卷分析中常用的多元分析主要有聚类分析、主成分分析、因子分析三类。正交分析则是首先对问卷进行正交分析，尽可能使问卷结构合理；然后进行问卷调查，并收集数据；最后整理数据，对调查结果进行分析（可采用正交实验中的方差分析），得出分析结果。

如前所述，由于被调查人群之间存在差异，结果同样存在差异，决定了结果分析中必定存在定性的因素，本系统中引入了对结果的定性分析。本系统中主要使用的定性分析方法包括前面提到的 Isomap 降维分析方法，以及现在克服专家群体的分散化思维常用的群体一致性算法[11,12]。

本文设计的系统中对问题的分析主要采用以上几种方法综合的方式。针对不同的问题类型设置了不同的分析方法：针对封闭性问题采用百分比的分析方法；针对开放性问题则是采用分类、分区间的定量方法与定性方法相结合的方式；对矩阵问题则采用层次分析法[8,11]进行分析。最后对整个问卷结果进行量化，通过降维及可视化技术对整体意见进行综合分析，得出群体意见发散程度，从而可以进一步协调专家个体之间的意见最终达到群体意见一致。

2.5 调查问卷结果可视化

结果的可视化有利于调查者直观地了解 （下转第 3194 页）

能量的数量、品质和时间,从库存、财务、MES 等系统获取相关信息,如各种能源介质的当前价格、燃烧品质、当前储量、可替代品种等,然后根据介质间的能量换算公式,获得最科学的能源介质搭配,最后将调配好的各种能源介质的数量发送给库存系统,进行配送或购买。模块的运行要求在保证满足能量需求的同时,将能源成本降到最低。如果该系统整体投入运行成功,预计能够为企业降低能耗 5% 以上,将会产生巨大的经济效益;同时让企业实实在在地体会到节能降耗的好处,具有长远的社会效益。

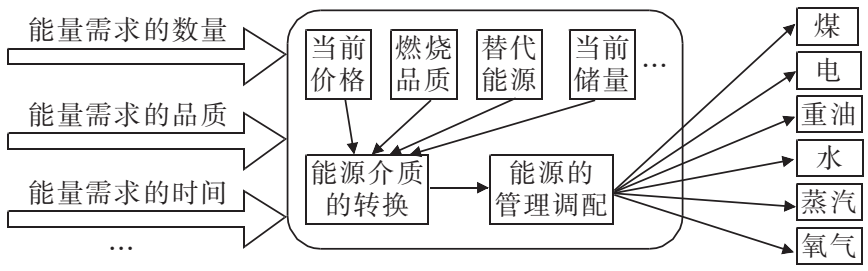

图 4 能源分析与分配模块

4 结束语

节能降耗是我国经济发展所迫切需要解决的问题和难点。虽然目前很多工矿企业已经意识到节能的重要性,但仅仅局限于对某些设备进行局部改造,缺乏对能源消耗进行系统化管理的理念。本文从企业管理角度出发,提出了节能降耗系统平台,从企业管理、流程监测、设备控制等三个层面,对企业中的能量转换平衡流程实现全面管控,实现了信息化与能源结合、能源与管理结合、能源管理与生产管理结合。

能源管理的信息化可以大大提高企业对能源的管理和利用能力,充分发掘企业节能潜力,降低企业能耗成本,在提高企业竞争力的同时,也为节约资源、保护环境做出了贡献。另外,企业实施信息化能源管理也有利于政府节能部门及时了解企业节能信息。总之,节能降耗系统平台能够发挥能源管理整体优化的特点和长期效应,将是企业节能发展的必然方向和实现长效节能的有效机制。

参考文献:

[1] nnual energy statistics, energy consumption by sector[EB/OL]. http://www.eia.doe.gov.

[2] International Energy Agency. Extended energy balances of OECD countries[R]. Paris: IEA, 2002.

[3] 戴彦德,朱跃中. 应慎重看待能源效率水平评价的国际比较[N]. 中国经济时报, 2005-06-17.

[4] 周伏秋. 国际能源评价指标体系及对我国的启示[J]. 中国战源, 2006, 28(11): 39-41.

[5] 丽贝卡·史密斯. 美国减少能源消耗量的十大举措[J]. 惠宇明,译. 电力需求侧管理, 2007, 9(2).

[6] LITTLE D. Overview of energy flow for industries in standard industrial classifications[R]. [S. l.]: Industrial Technologies Program (ITP), 2000.

[7] MAOYA M, PIETTE M A, KINNY S, *et al*. Case studies of energy information systems and related technology: operational practices, costs, and benefits[C]//Proc of International Conference for Enhanced Building Operations. 2003.

[8] 顾海涛,王书保,卢毅. 企业信息化能源管理系统的应用与企业节能[J]. 电力需求侧管理, 2006, 8(4): 40-42.

[9] 王永川,陈光明. 钢铁企业能源管理系统方案研究[J]. 冶金能源, 2003, 22(6): 5-6, 36.

[10] 冯为民,丛力群. 冶金企业能源管理系统[J]. 控制工程, 2005, 12(6): 597-600.

(上接第 3190 页) 调查结果。在此系统中,笔者主要利用了 Zed Graph 开源. NET 图表类库,根据问卷分析方法设计了几种问卷的结果可视化的显示方式。此处仅以 Isomap 降维定性分析方法可视化界面和百分比方法可视化界面作为例子,说明调查结果的可视化,如图 2、7 所示。

图6 问卷发放界面

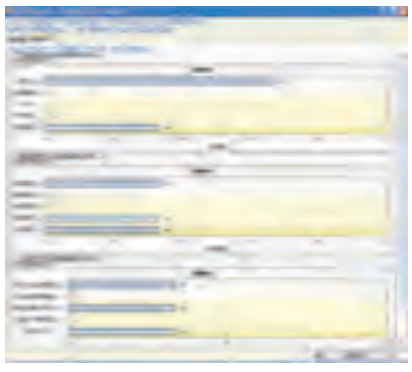

图7 百分比分析结果

3 结束语

在综合集成研讨厅中,不可避免地存在专家群体之间的分散化思维。调查问卷作为收集专家意见的一种常用方法,对其结果进行统计分析能比较客观地反映不同专家观点上的差异。本文设计的自动调查问卷系统采用定量与定性结合的方法完成对调查问卷结果的综合和意见一致性分析并最终实现了结果的可视化,从而有效分析专家个体之间行为相似性,并取得了不错的效果。本文提出的基于 Isomap 降维算法的问卷结果量化及可视化方法为综合集成研讨厅体系体现专家思维聚类情况提供了一种理想的可视化环境。

参考文献:

[1] 钱学森,于景元,戴汝为. 一个科学的新领域——开放的复杂巨系统及其方法论[J]. 自然杂志, 1990, 13(1): 3-10.

[2] 钱学森. 再谈开放的复杂巨系统[J]. 模式识别与人工智能, 1991, 4(1): 1-4.

[3] 戴汝为. 从定性到定量的综合集成技术[J]. 模式识别与人工智能, 1991, 4(1): 5-10.

[4] 王丹力,戴汝为. 专家群体思维收敛的研究[J]. 管理科学学报, 2002, 5(2): 1-5.

[5] 刘春梅. 综合集成研讨厅专家群体评估结果的可视化[J]. 模式识别与人工智能, 2005, 18(1): 6-11.

[6] TENNENBAUM J B, DeSILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(22): 2319-2323.

[7] ZHA Hong-yuan, ZHANG Zhen-yue. Isometric embedding and continuum ISOMAP[C]//Proc of the 20th International Conference on Machine Learning(ICML-2003). Washington DC: [s. n.], 2003: 21-24.

[8] 吕永红. 市场调查问卷的定量分析[J]. 集团经济研究, 2005(1): 52-53.

[9] 袁岳,周林古. 零点调查——民意测验的方法与经验[M]. 福州: 福建人民出版社, 2005.

[10] 刘菊红. 调查问卷中的统计分析方法[J]. 上海统计, 2002(2): 35-37.

[11] 王丹力,戴汝为. 群体一致性及其在研讨厅中的应用[J]. 系统工程与电子技术, 2001, 23(7): 33-37.

[12] 吕志坚. 综合集成研讨厅中的专家研讨与思维研究[D]. 北京: 中国科学院自动化所, 2004.