# 基于信息粒度的知识网的模糊分类与检索方法

杨人子 [1,2] 严洪森 [1,3]

**摘 要** 针对知识化制造系统自重构中知识网检索方法过于主观以及重复检索和运算等问题, 提出基于信息粒度的知识网的模糊分类和检索方法. 知识网复杂度解决了自重构运算导致的知识网存在多样性的问题. 相似度考虑知识网在 "质"、"量" 和复杂性等方面的差异, 具有反映知识网运算规律的特征. 知识网模糊聚类方法不需要确定分类数, 并且能够同时获得关于目标知识网的排序. 以各聚类中心为中心确定的检索空间实现了问题由细粒度空间转化为粗粒度空间.

**关键词** 知识网, 信息粒度, 相似度, 聚类

**DOI** 10.3724/SP.J.1004.2011.00585

# The Method of Fuzzy Classification and Searching for Knowledge Meshes Based on Information Granularity

YANG Ren-Zi[1,2] YAN Hong-Sen[1,3]

**Abstract** To solve over-subjective retrieval method, repetition retrieval and operation of knowledge meshes in the self-reconfiguration of knowledgeable manufacturing system, the method of fuzzy classification and searching based on information granularity is proposed. The complexity of knowledge mesh solves the problem about the diversity of knowledge mesh caused by self-reconfiguration computing. The similarity degree which considers the differences in three aspects of quality, quantity and complexity has the characteristics of reflecting operation laws. The fuzzy clustering method of knowledge mesh does not need the classes number, meanwhile the permutation about target knowledge mesh is obtained. The search space is determined by centering at the clustering, which converts the problem from fine-grained space to coarse-grained space.

**Key words** Knowledge mesh, information granular, similarity, clustering

知识化制造是 2000 年提出的一种新的制造理念[1], 其致力于解决现有制造模式中存在的模式单一、缺乏灵活性、不能满足制造企业需求的多样性以及重复研发等问题. 自重构是知识化制造系统的重要特征, 已经建立了知识网多重集理论和自重构算法[2]、知识网自动生成方法[3] 以及重构后的模式决策方法[4]. 已有的研究表明在知识网多重集理论和自重构算法的基本理论基础之上, 利用基于模糊综合评判的知识网运算方法可以获得在每个功能需求下的满意度高的几个知识网, 进一步通过改进的混合遗传算法对知识网多重集运算的表达式进行优化从而获得新知识网. 但在此过程中, 参与自重构的知识网是通过模糊满意关系获得, 人为因素影响太大; 系统是对用户的每个需求检索, 用户需求越多, 检索次数和其后的自重构运算次数也就越多; 不同需求下相同的知识网, 会进行重复的检索过程和自重构运算; 即使将检索的相同知识网进行合并再自重构运算, 也仍旧会面临过多的自重构运算, 使得系统过于复杂而不适于实际应用. 已有的文献对这些问题并没有讨论, 因此如何能够更为合理地评价知识网, 减少知识网的检索次数, 简化自重构运算以及新知识网的复杂性, 并尽可能地获得各需求下满意度均为最高的新知识网, 是本文要解决的问题.

粒度计算是描述问题空间和解决问题的有效手段, 在软计算、数据挖掘、知识发现等[5−10] 领域中得到了广泛的应用并取得较为理想的效果. 但将粒度原理用于解决自重构问题尚未见报道. 另外, 知识网的自重构运算会导致不同的知识网多重集对应相同的知识网, 这就需要对知识网的复杂性进行分析. 虽然 Yan 等[4] 在模式决策方法中对知识网的复杂度有所讨论, 但其不能反映知识网在自重构运算过程中的变化特点. 熵函数是评价制造系统复杂性的有效方法[11−13], 系统的规模性、元素种类及元素间的关系均可以影响系统的熵值. 而知识网作为一种新的先进制造知识表示, 已有的熵模型[11−13] 不能应用到知识网上, 需要建立一种新的熵模型用于知识网的复杂性分析.

收稿日期 2010-06-11 录用日期 2011-01-12
Manuscript received June 11, 2010; accepted January 12, 2011
国家自然科学基金重点项目 (60934008) 资助
Supported by the Key Program of National Natural Science Foundation of China (60934008)

1. 东南大学自动化学院 南京 210096 2. 东南大学数学系 南京 210096 3. 东南大学复杂工程系统测量与控制教育部重点实验室 南京 210096
1. School of Automation, Southeast University, Nanjing 210096 2. Department of Mathematics, Southeast University, Nanjing 210096 3. Key Laboratory of Measurement and Control of CSE, Ministry of Education, Southeast University, Nanjing 210096

一个知识网可以看成由一族等价关系得到的不同类中的基本知识点构成, 其中等价关系为用户需求[14]. 若需求不同, 同一个知识网所对应的类也不同. 这其实是粒度思想的一种体现. 不同的需求对应不同的粒度表示, 同一个知识网可以有不同的粒度表达形式. 基于这一思想启发, 本文提出一种基于信息粒度原理的知识网的模糊分类与检索方法. 相似性度量是对知识网进行分类的前提. 丁雪峰等[15] 用知识点相同功能的数量定义相似度, 这仅仅是从定量的角度考虑, 忽略了知识网的固有特性. 文中的相似度函数是从定性定量结合的角度, 依据用户提出的需求和知识网的固有特征来构造相似度, 特别对知识网的复杂性进行了专门的讨论. 构造的复杂度能够反映知识网在运算过程中的变化特征, 并且相似度也具有该特点. 考虑与目标知识网匹配的最大化和知识网库中知识网间相似的最小化, 对知识网进行模糊聚类. 模糊聚类方法[16−17] 往往需要确定分类数, 而文中方法的类是迭代构建的, 因此无需事先给定类的个数. 作为聚类中心的知识网依次是某些需求下的最佳状态, 并由此确定类的特征和用户需求的合并, 以及其他知识网在类特征下的关于目标知识网的排序. 权重的优化使得聚类中心的类特征更为明显. 以各聚类中心为中心, 建立的检索空间将知识网的检索问题由细粒度空间转化为粗粒度空间.

## 1 知识网的相似度函数的定义

### 1.1 知识网[2]

任何一种先进制造系统都是一种制造模式的具体实现, 都由若干功能相对独立且相互有联系的模块构成. 将每个功能模块看作是一个 Agent, 则先进制造系统可张成由 Agent 及它们之间的联系组成的网, 称为 Agent 网. Agent 网是实际制造系统, 不能进行逻辑的运算, 将每个 Agent 映射为知识点, Agent 之间的联系映射为知识点之间的联系, 就构成了知识网, 其反映的是 Agent 网的静态信息. 将先进制造模式转化为知识网, 存储在知识化制造系统知识网库中, 进行逻辑运算, 获得新知识网, 再映射为 Agent 网, 从而构建新的先进制造系统.

知识网是 Agent 网的一个集合表达、图形表达或关系数据库表达, 其集合表达为由知识点、继承流、信息流和功能等构成的大集合, 可以简单地记为知识网 $W = \{x_1, x_2, \cdots, x_n\}$. 知识网的元素与相应的 Agent 网元素之间是一一对应的, 两者之间映射关系的图形见图 1.

图 1 中只表示 Agent 网和知识网的形式定义中的知识点和各种信息联系, 而知识点功能、信息流和继承流在面向关系数据库的表示中实现. 图 1 中, 细圆圈表示知识点, 细圆圈中的字符表示知识点名, 粗圆圈表示 Agent, 粗圆圈中的字符表示 Agent 名, 粗曲线表示复合联系, 细曲线表示除了复合联系以外其他的信息联系, 虚曲线表示知识点和 Agent 之间的一对一的映射.

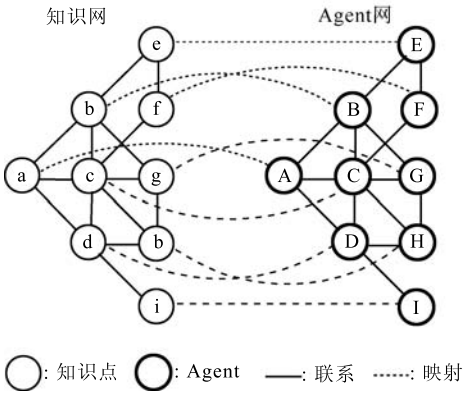


图 1 知识网和 Agent 网

Fig. 1 Knowledge mesh and Agent mesh

根据用户需求对知识网进行自重构, 需求的不同层次决定了 Agent 粒度的不同和知识网在不同层次上的自重构. 若用户对敏捷制造系统在执行层进行自重构, 则每个 Agent 为执行单元或者功能实体, 将敏捷制造系统抽象为 Agent 网, 再映射为知识网, 即可进行知识网的自重构, 如图 2 所示.

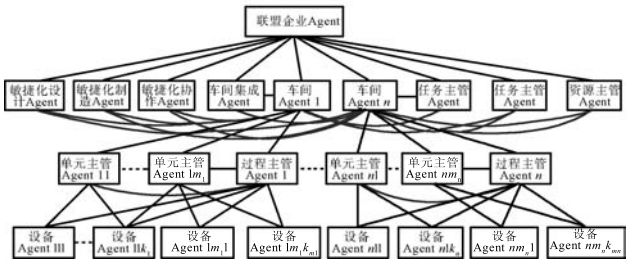


图 2 敏捷制造系统对应的 Agent 网

Fig. 2 Agent mesh corresponding to agile manufacturing system

### 1.2 复杂度

知识网多重集理论[2] 解决了知识网运算中采用经典集合运算引起的元素信息丢失和集合还原失真问题, 但也使得知识网库中的知识网, 有的为已有知识网, 有的为自重构得到的知识网, 完全相同的知识网对应的多重集未必相同. 将这种知识网间的差异通过对复杂性的定量分析来区别, 并且将其建立在知识网多重集的基础之上. 知识网 $W = \{x_1, x_2, \cdots, x_n\}$ 的多重集 $W_M^{[2]}$ 为: $W_M = \{\alpha_1 x_1, \alpha_2 x_2, \cdots, \alpha_n x_n\}$, 其中 $\alpha_1, \alpha_2 \cdots, \alpha_n$ 为有界实数, 称为

元素的多重数.

Shannon 首次提出用熵来度量信息量, 但其形式局限于概率形式; 并且在公式 $H=-\Sigma p_i \log p_i$ 中, 当任一 $p_i=1$ 时, $H=0$, 这不符合复杂性问题的特点. 因此将信息熵进行改进和推广, 给出如下定义.

**定义 1.** 称知识网 $W=\{x_1,x_2,\cdots,x_n\}$ 的复杂度为

$$G(W)=\sum_{j=1}^{n}\lambda_j I(\alpha_j)\cdot\log\left(\frac{\sum\limits_{i=1}^{n}\lambda_i I(\alpha_i)}{\lambda_j I(\alpha_j)}+1\right) \quad (1)$$

其中, $I(\alpha_i)=\begin{cases}\alpha_i, & \alpha_i>0\\ 0, & \alpha_i\le 0\end{cases}$, $\alpha_i$ 为元素 $x_i$ 的多重数; $\lambda_i$ 表示元素 $x_i$ 在整个多重集中的重要程度 ($\lambda_i\ge 0$).

复杂度 $G(W)$ 具有下列性质: $G(W)$ 是元素个数和多重数的单增函数; 当知识网的元素多重数非负, 知识网的交和差运算不会使得复杂度增加, 知识网的并运算不会使得复杂度减少. 该性质证明见本文附录. 当制造企业对知识化制造系统提出新需求时, 如生产新产品、增加新功能或完善功能模块等, 制造系统的复杂程度会发生改变, 反映在知识网上是元素、元素数目和多重数的改变以及知识网的各种运算. 特别是知识网的并运算虽然可以发挥不同制造模式的优势, 但合并后的新知识网的复杂度会大于合并前的任何一个知识网的复杂度. 进一步, 还可以证明合并知识网元素, 如制造系统的单元化构造可以降低知识网的复杂度; 合并后的知识网的复杂度大于合并前所有知识网复杂度之和. 因此知识网的复杂度能够反映知识网的各种变化, 反映制造系统复杂程度的改变, 是知识网的固有性质.

### 1.3　相似度

**定义 2.** 设知识网 $V=\{v_1,v_2,\cdots,v_m\}$, 知识网 $W=\{x_1,x_2,\cdots,x_n\}$, $W$ 关于 $V$ 的相似度为

$$sim(V,W,\boldsymbol{\Omega})=\frac{\overset{m}{\underset{i=1}{T}}\left(s(\mu_{W(v_i)}f(v_i,W),\omega_i^2)\right)}{\varepsilon\cdot g(V,W)} \quad (2)$$

其中, $T(\cdot)$, $s(\cdot)$ 分别表示模糊集的 $t-$模和 $s-$模, 将其具体化, 本文采用 "积" 和 "概率和" 作为相应形式, 即 $T(x,y)=xy$, $s(x,y)=x+y-xy$. $f(v_i,W)$ 为 $W$ 关于 $v_i$ 的匹配度. $\mu_{W(v_i)}$ 表示知识网 $W$ 中与 $v_i$ 相匹配的元素的完善度. $g(V,W)$ 为 $W$ 相对 $V$ 的复杂性系数. $\mu_{W(v_i)}f(v_i,W)$ 称为知识网 $W$ 关于 $v_i$ 的满意度, 记为 $\alpha^i_{W(V)}$. $\boldsymbol{\alpha}_{W(V)}=(\alpha^1_{W(V)},\alpha^2_{W(V)},\cdots,\alpha^m_{W(V)})$ 为知识网 $W$ 关于知识网 $V$ 的满意度向量. $\varepsilon$ 称为相似度的调节系数, 其作用是放大分子的计算数值便于相似度的比较. 权重 $\omega_i\in[0,1]$ 是衡量满意度分量对最终相似度 $sim(\cdot)$ 的影响, $\boldsymbol{\Omega}=(\omega_1,\omega_2,\cdots,\omega_m)$.

匹配度 $f(v_i,W)$ 是反映知识网的匹配程度, 表现知识网之间在 "量" 上的不同. 定义为 $f(v_i,W)$: $V\times W\to[0,1]$. 当 $W$ 中存在元素与 $v_i$ 完全匹配, 则 $f(v_i,W)=1$; 当 $W$ 中任何元素与 $v_i$ 完全不匹配, 则 $f(v_i,W)=0$; 当 $W$ 中的元素与 $v_i$ 是部分匹配, $0<f(v_i,W)<1$. 匹配度的具体函数形式由元素 $v_i$ 的类型来决定.

完善度 $\mu_{W(v_i)}$ 是反映知识网对应的制造模式的完善程度. $\mu_{x_i}:W\to[0,1]$ 称为知识网元素 $x_i$ 的完善度. $\boldsymbol{\mu}=(\mu_{x_1},\mu_{x_2},\cdots,\mu_{x_n})$ 称为知识网 $W$ 的完善度. $\mu_{x_i}=0$ 表示知识网不具备元素 $x_i$, 数值越大说明元素 $x_i$ 的特征越完善. 该数值是知识网存储到知识网库中同时存入的数据 (可以通过专家对知识网中的各元素完善程度进行等级划分, 采用模糊分析方法获得). 在知识网库中, 任何一个知识网都具有一个完善度. 完善度是知识网的固有信息, 相对于匹配度, 反映的是知识网之间 "质" 的比较.

复杂性系数 $g(V,W)$ 是对知识网复杂程度的比较. 因为相同知识网对应的多重集未必相同, 满足用户需求的知识网会多样化. 用户到底应该选择满意度低一些的已有知识网, 还是经过多次自重构而满意度高但复杂度也高的知识网, 复杂性系数起到一个权衡作用.

$$g(V,W)=\begin{cases}\dfrac{G(W)}{G(V)}, & G(W)>G(V)\\ 1, & G(W)\le G(V)\end{cases} \quad (3)$$

$g(V,W)\ge 1$. 当 $W$ 比 $V$ 复杂, 则 $g(V,W)>1$; 当 $V$ 比 $W$ 复杂, 或者 $V$ 和 $W$ 的复杂度相同, 则取 $g(V,W)=1$.

相似度 $sim(\cdot)$ 是综合考虑了知识网在 "质"、"量" 和复杂性三个方面的差异, 具有下列性质: 当知识网 $V$ 和 $W$ 完全相同的情况下, 相似度为 1; 当相似度为 0 时, 知识网 $W$ 关于某个 $v_i$ 为完全不匹配; 相似度函数是满意度和权重的单增函数, 是复杂性系数的单减函数; 在复杂性系数为 1 的条件下, 子集、交集的相似度不大于原集合的相似度, 并集的相似度不小于任何一个原集合的相似度. 该性质证明见本文附录. 该性质说明若将用户需求的知识网视为目标知识网, 将知识网库中知识网与其比较, 则满意度高的知识网是用户所期待的; 若用户对需求有所偏好, 权重分量的调节可以体现这一点; 又若权重是由满意度决定, 即满意度分量值越大, 对应的权重分量值也越大, 则最终相似度值会使得知识网间的相似程度体现得更加明显; 复杂性系数可

以区分那些虽然完全满足需求但过于复杂的知识网, 如具有冗余功能和信息联系的知识网、多重数不为 1 的知识网等情况. 该性质也说明当知识网复杂度不大于目标知识网复杂度时, 并集运算会使得它们的相似度提高. 但要注意的是, 并不是合并的知识网就一定是用户的最佳选择, 若合并的知识网过于复杂使得复杂性系数大于 1, 相似度可能反而会降低. 因此知识化制造系统在通过知识网的各种运算获得期望知识网的同时, 应当尽量减少复杂度的增加.

## 2 基于信息粒度原理的知识网的模糊分类与检索方法

将知识网库中的知识网根据用户需求, 按照相似性进行聚类, 将满意度值相似的知识网聚成同一类, 在类中对知识网排序选择用于自重构. 换句话说, 就是把自重构问题从用户每个需求都需要检索的 "细" 粒度空间转化为在类中检索的 "粗" 粒度空间进行处理.

### 2.1 聚类中心和分类数的确定

将完全满足用户需求的知识网设为目标知识网, 若知识网库中存在多个完全满足用户需求的知识网, 则按照式 (2) 对其进行排序选择. 若知识网库中没有完全满足条件的知识网, 对知识网进行聚类分析, 选出类中代表知识网参与下一步自重构.

首先, 一些相似度值极低的知识网不属于讨论的范围, 但又为了避免漏掉一些在个别需求上具有很高满意度的知识网, 因此确定某个数值 $\tau$, 对由满意度分量不全小于 $\tau$ 的知识网构成的集合进行聚类. 设知识网库中满足 $\tau$ 条件的知识网有 $N$ 个 $W_1, W_2, \cdots, W_N$, 目标知识网 $V = \{v_1, v_2, \cdots, v_m\}$.

第 1 个聚类中心 $V_1$ 选作知识网集合中的一个知识网 $W_j$ ($V_1 = W_j, j = 1, 2, \cdots N$), 其使得式 (2) 的值最大化:

$$sim(V, V_1, \boldsymbol{\Omega}_1) = \max_{j=1,\cdots,N} sim_1(V, W_j, \boldsymbol{\Omega}_1) \tag{4}$$

设 $V_1$ 对应的满意度为 $\boldsymbol{\alpha}_{V_1(V)} = (\alpha^1_{V_1(V)}, \alpha^2_{V_1(V)}, \cdots, \alpha^m_{V_1(V)})$, 取定阈值 $\theta$, 由满足 $\alpha^{k_1}_{V_1(V)} \geq \theta$ 的集合 $\{\alpha^{k_1}_{V_1(V)}\}$ 所对应的 $\{v_{s_1}\}$ 作为以 $V_1$ 为聚类中心的类所具有的特征, 即 $V_1$ 的满意度分量 $\alpha^{k_1}_{V_1(V)}$ 具有较高的数值. 记 $\{\alpha^{k_1}_{V_1(V)}\}$ 的上标的集合为 $\{k_1\}$, 个数为 $l_1$.

第 2 个聚类中心 $V_2$ 不能距第 1 个聚类中心太近而成为第 1 个聚类中心的复制品, 为避免这种效果, 考虑下面表达式求最大值:

$$\begin{aligned} &sim(V, V_2, \boldsymbol{\Omega}_2) = \\ &\quad \max_{j=1,\cdots,N}(1 - \varepsilon sim_V(V_1, W_j, \boldsymbol{0}))^{l_1} \times \\ &\quad sim_2(V, W_j, \boldsymbol{\Omega}_2) = (1 - \varepsilon sim_V(V_1, V_2, \boldsymbol{0}))^{l_1} \times \\ &\quad sim_2(V, V_2, \boldsymbol{\Omega}_2) \end{aligned} \tag{5}$$

其中, $sim_V(V_1, V_2, \boldsymbol{0}) = \dfrac{\overset{m}{\underset{i=1, i\notin\{k_1\}}{T}}(s(\alpha^i_{V_2(V_1)}, 0))}{\varepsilon g(V_1, V_2)}$ 表示在舍去 $\{v_{s_1}\}$ 的 $V$ 前提下, $V_2$ 关于 $V_1$ 的相似度. $\boldsymbol{\alpha}_{V_2(V_1)} = \{\alpha^1_{V_2(V_1)}, \alpha^2_{V_2(V_1)}, \cdots, \alpha^m_{V_2(V_1)}\}$ 为 $V_2$ 关于 $V_1$ 的满意度向量. 各分量的值通过 $V_2$, $V_1$ 关于 $V$ 的满意度取得, 设 $V_2$ 对应的满意度为 $\boldsymbol{\alpha}_{V_2(V)} = (\alpha^1_{V_2(V)}, \alpha^2_{V_2(V)}, \cdots, \alpha^m_{V_2(V)})$, 则

$$\alpha^i_{V_2(V_1)} = \begin{cases} \dfrac{\alpha^i_{V_2(V)}}{\alpha^i_{V_1(V)}}, & \alpha^i_{V_2(V)} \leq \alpha^i_{V_1(V)},\ \ i \notin \{k_1\} \\ \dfrac{\alpha^i_{V_1(V)}}{\alpha^i_{V_2(V)}}, & \alpha^i_{V_2(V)} > \alpha^i_{V_1(V)},\ \ i \notin \{k_1\} \end{cases}$$

$(1 - \varepsilon sim_V(V_1, V_2, \boldsymbol{0}))^{l_1}$ 表达了 $V_2$ 尽可能地远离 $V_1$. $sim_2(V, V_2, \boldsymbol{\Omega}_2)$ 表示 $V_2$ 和 $V$ 均去掉与 $\{v_{s_1}\}$ 对应的项后, $V_2$ 关于 $V$ 的相似度. $\alpha_{V_2(V)}$ 去掉与 $\left\{\alpha^{k_1}_{V_1(V)}\right\}$ 对应的项后, 由满足 $\alpha^{k_2}_{V_2(V)} \geq \theta$ 的 $\{\alpha^{k_2}_{V_2(V)}\}$ 所对应的 $\{v_{s_2}\}$ 作为以 $V_2$ 为聚类中心的类所具有的特征. 记 $\{\alpha^{k_2}_{V_2(V)}\}$ 的上标的集合为 $\{k_2\}$, 个数为 $l_2$. $g(V_1, V_2)$ 为 $V_2$ 相对 $V_1$ 的复杂性系数,

$$g(V_1, V_2) = \begin{cases} \dfrac{g(V, V_2)}{g(V, V_1)}, & g(V, V_1) \leq g(V, V_2) \\ 1, & g(V, V_1) > g(V, V_2) \end{cases}$$

继续确定第 3 个聚类中心, 依次类推, 第 $L$ 个聚类中心按以下形式:

$$\begin{aligned} Q(L) = &\max_{j=1,\cdots,N}(1 - \varepsilon sim_V(V_{L-1}, W_j, \boldsymbol{0}))^{l_{L-1}} \times \\ &(1 - \varepsilon sim_V(V_{L-2}, W_j, \boldsymbol{0}))^{l_{L-2}} \cdots \times \\ &(1 - \varepsilon sim_V(V_1, W_j, \boldsymbol{0}))^{l_1} \times \\ &sim_L(V, W_j, \boldsymbol{\Omega}_L) = (1 - \varepsilon \times \\ &sim_V(V_{L-1}, V_L, \boldsymbol{0}))^{l_{L-1}} (1 - \varepsilon \times \\ &sim_V(V_{L-2}, V_L, \boldsymbol{0}))^{l_{L-2}} \cdots (1 - \varepsilon \times \\ &sim_V(V_1, V_L, \boldsymbol{0}))^{l_1} sim_L(V, V_L, \boldsymbol{\Omega}_L) \end{aligned} \tag{6}$$

当寻找当前聚类中心时, 式 (6) 考虑了以前所有的聚类中心. 当 $\left\{\alpha^{k_1}_{V_1(V)}\right\}, \left\{\alpha^{k_2}_{V_2(V)}\right\}, \cdots, \left\{\alpha^{k_L}_{V_L(V)}\right\}, \cdots$ 所对应的 $\{v_s\}$ 最终构成 $V$, 则聚类过程结束, 分类数 $A$ ($A \geq L$) 也随即确定. 可以看到第 1 个聚类中心是全体知识网的关于知识网 $V$ 的最好代表, 以后

诸次找到的聚类中心为 $V$ 中去掉某些项以后的最好代表. 当 $\{v_s\}$ 不能构成 $V$, 说明关于 $V-\{v_s\}$ 的需求在阈值 $\theta$ 的条件下没有知识网可以选择, 此时可以调低 $\theta$ 值, 重新进行聚类分析, 直到 $\{v_s\}$ 能够构成 $V$, 聚类过程结束. 整个聚类过程是与阈值 $\theta$ 相关的, $\theta$ 的不同取值将获得不同聚类中心和分类结果.

### 2.2　权重的优化

在上述过程中, 对权重 $\boldsymbol{\Omega}$ 没有任何规定. 如何通过权重使得每个知识网的优势得到更好的发挥, 下面给出关于 $\boldsymbol{\Omega}$ 优化问题的结论.

**定理 1.** 存在 $\boldsymbol{\Omega}_L$ 使得 $Q(L)$ 取得最大值, 且 $\boldsymbol{\Omega}_L$ 各分量 $\omega_{L_i}\in[0,1]$ 满足归一性质, 即 $\sum\omega_{L_i}=1$.

**证明.** 将式 (6) 中 $(1-\varepsilon sim_V(V_{L-1},V_L,\boldsymbol{0}))^{l_{L-1}}\cdot(1-\varepsilon sim_V(V_{L-2},V_L,0))^{l_{L-2}}\cdots(1-\varepsilon sim_V(V_1,V_L,0))^{l_1}$ 记为 $G$, 其值不依赖 $\boldsymbol{\Omega}_L$. 记 $\{k_1\}\cup\{k_2\}\cup\cdots\cup\{k_{L-1}\}=\{k\}$, 则 $Q(L)=Gsim_L(V,V_L,\boldsymbol{\Omega}_L)=G\dfrac{\overset{m}{\underset{i=1,i\notin\{k\}}{T}}(s(\alpha^i_{V_L(V)},\omega^2_{L_i}))}{\varepsilon g(V,V_L)}$. 应用拉格朗日乘子法, 构造辅助函数 $F$, 求解 $\dfrac{\mathrm{d}F}{\mathrm{d}\omega_{L_t}}=0$, $\dfrac{\mathrm{d}F}{\mathrm{d}\lambda}=0$.

$$F=Gsim_L(V,V_L,\boldsymbol{\Omega}_L)+\lambda\Big(\sum_{i=1,i\notin\{k\}}^{m}\omega_{L_i}-1\Big)$$

设 $\varepsilon g(V,V_L)=\beta_{V_L}$, $\overset{m}{\underset{i=1,i\neq t,i\notin\{k\}}{T}}(s(\alpha^i_{V_L(V)},\omega^2_{L_i}))=\gamma_{L_t}$

$$\because\ \frac{\mathrm{d}F}{\mathrm{d}\omega_{L_t}}=G\frac{\mathrm{d}}{\mathrm{d}\omega_{L_t}}\Big(\frac{\overset{m}{\underset{i=1,i\notin\{k\}}{T}}(s(\alpha^i_{V_L(V)},\omega^2_{L_i}))}{\varepsilon\cdot g(V,V_L)}\Big)-\lambda=2\gamma_{L_t}\frac{G}{\beta_{V_L}}(\omega_{L_t}-\omega_{L_t}\alpha^t_{V_L(V)})-\lambda=0$$

$$\therefore\ \omega_{L_t}=2\lambda\gamma_{L_t}\frac{G}{\beta_{V_L}}(1-\alpha^t_{V_L(V)})$$

$$\because\ \sum_{i=1,i\notin\{k\}}^{m}\omega_{L_i}=1$$

$$\therefore\ \omega_{L_t}=\frac{1}{\gamma_{L_t}(1-\alpha^t_{V_L(V)})\sum\limits_{i=1,i\notin\{k\}}^{m}\frac{1}{\gamma_{L_i}(1-\alpha^i_{V_L(V)})}}\tag{7}$$

□

定理 1 给出了相似度公式中权重 $\boldsymbol{\Omega}$ 的计算公式. 计算式 (7), 每个知识网都会获得一个优化的权重向量, 其反映了知识网的局部特征, 即较高的权重分量值意味着相应 $v_s$ 的满意度值也较高. 将 $\boldsymbol{\Omega}$ 代入 $Q(L)$, 找出 $Q(L)$ 为最大的知识网, 即为聚类中心.

### 2.3　检索空间

按照第 2.1 节和第 2.2 节方法得到的聚类中心, 是在某些需求下具有最佳状态的知识网. 为了使用户增加选择的灵活性, 实现知识网的检索问题由需要 $m$ 次检索的细粒度空间转化为 $A$ (分类数) 次检索的粗粒度空间, 构造以聚类中心为中心的检索空间. 设 $V$ 为某个聚类中心, $\boldsymbol{\Omega}$ 为对应的权重向量, 计算 $V$ 与所有知识网的平均相似度:

$$q=\frac{1}{\rho\cdot N}\sum_{j=1}^{N}sim(V,W_j,\boldsymbol{\Omega})$$

$\rho$ 为平均相似度的调节系数, $\rho>1$. $\rho$ 的作用是调节检索空间的大小, 决定空间中知识网的数量, 避免因为 $q$ 值过大而使得检索空间中只包含聚类中心.

将 $q$ 值作为确定检索空间的指标, 寻找一个与知识网集合无关的满意度向量 $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_m)$ 满足 $q=\dfrac{\overset{m}{\underset{i=1}{T}}(s(\alpha_i,\omega_i^2))}{\varepsilon g(V,W)}$, 其中 $g(V,W)$ 取为 1. $\boldsymbol{\alpha}$ 一旦确定, 即可判断位于检索空间内知识网的满意度的取值范围. 给出 $\boldsymbol{\alpha}$ 的精确解是非常困难的, 将该问题转化为标准均方误差的近似问题:

$$P=\left(\frac{\overset{m}{\underset{i=1}{T}}(s(\alpha_i,\omega_i^2))}{\varepsilon g(V,W)}-q\right)^2\to\min(\boldsymbol{\alpha})\tag{8}$$

采用梯度法对 $\boldsymbol{\alpha}$ 进行修正: $\boldsymbol{\alpha}(new)=\boldsymbol{\alpha}-\gamma\cdot\nabla_{\boldsymbol{\alpha}}P$, 其中 $\gamma$ 代表正学习率.

$$\alpha_k(new)=\alpha_k-\gamma\frac{\partial P}{\partial\alpha_k}=\alpha_k-2\gamma\left(\frac{\overset{m}{\underset{i=1}{T}}(s(\alpha_i,\omega_i^2))}{\varepsilon g(V,W)}-q\right)\times\frac{\overset{m}{\underset{i=1,i\neq k}{T}}(s(\alpha_i,\omega_i^2))}{\varepsilon g(V,W)}(1-\omega_k^2)$$

聚类中心 $V$ 对应的满意度为 $\boldsymbol{\alpha}_V=(\alpha_V^1,\alpha_V^2,\cdots,\alpha_V^m)$, 式 (8) 的解为 $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_m)$, 则位于该检索空间的知识网满意度分量的取值范围 $[x_i^-,x_i^+]$ 由下式确定:

$$x_i^-=\max\{\alpha_V^i\cdot\alpha_i,0\},\ \ x_i^+=\min\left\{\frac{\alpha_V^i}{\alpha_i},1\right\}\tag{9}$$

### 2.4 算法步骤

知识网的模糊分类与检索的算法步骤可概括为:

**步骤 1.** 将用户需求映射为目标知识网 $V$.

**步骤 2.** 列出知识网库中知识网关于目标知识网的完善度、匹配度和复杂性系数, 其中按照式 (1) 和式 (3) 计算复杂性系数. 选取满意度分量不全小于 $\tau$ 的知识网参与分类与检索.

**步骤 3.** 按照式 (7) 计算每个知识网的权重向量. 取定 $\varepsilon$, 计算式 (4), 确定第 1 个聚类中心 $V_1$. 取定阈值 $\theta$, 确定第 1 个类特征 $\{v_{s_1}\}$.

**步骤 4.** 计算式 (5) 和式 (7), 获得第 2 个聚类中心 $V_2$ 和类特征 $\{v_{s_2}\}$. 计算式 (6) 和式 (7), 依次获得其他聚类中心. 当所有类特征 $\{v_{s_1}\}$, $\{v_{s_2}\}$, $\cdots$, $\{v_{s_A}\}$ 能够构成 $V$ 时, 分类数 $A$ 确定. 否则, 调低 $\theta$ 值重新聚类.

**步骤 5.** 设定 $\rho$ 和 $\gamma$, 按照式 (8) 和式 (9) 计算对应于每个聚类中心的检索空间, 确定检索空间中的知识网.

## 3 实例

上述知识网的模糊聚类算法已在.Net 平台上采用 C# 语言实现, 并利用数据库管理系统 SQL Server 2000, 开发了知识网的模糊分类和检索的使能工具. 下面结合某公司简化 CIMS/MIS 工程的实例来说明知识网的模糊分类和检索过程.

假设企业用户对管理信息系统提出了如下的功能需求: 财务管理, 生产管理, 质量管理, 设备管理. 具体子功能见表 1. 在知识网库中有 15 个知识网 $W_1, W_2, \cdots, W_{15}$. 根据第 1.3 节, 需要给出目标知识网以及 15 个知识网关于目标知识网的完善度、匹配度和复杂性系数. 这些数据可通过专家评分、匹配法则、复杂度计算等方法获得. 这里由于篇幅所限, 不给出它们的具体计算过程, 而是直接给出 15 个知识网的满意度, 见表 1. 复杂性系数均取为 1. 但复杂性系数将通过一个简单的例子给以说明. 由表 1 可知 15 个知识网均为不能完全满足用户需求的知识网, 需要进行自重构以获得满意度较高的新知识网, 现在从中选出进行自重构的知识网.

### 3.1 复杂性系数的计算

企业用户的需求为表 1 的功能需求, 无法映射为完整的目标知识网, 所以知识网复杂度的计算只考虑功能复杂度 (若用户为可以详细描述其具体需求的高素质用户, 能够将需求映射为完整的目标知识网, 则复杂度的计算为整个知识网所有元素的复杂度). 为了说明完全相同的知识网的复杂度未必相同, 这里同时给出在功能需求下知识网的匹配度. 另外, 知识网的表示也只显示知识点, 而忽略知识网表示中的其他元素, 并且知识点的功能若与用户需求的功能相同, 则知识点用相同的字母代替. $W_i$ 表示知识网, $W_{M_i}$ 表示知识网 $W_i$ 对应的多重集, $W_{i+j}$ 和 $W_{i+j+k}$ 表示经过并集运算得到的新知识网[2], $W_{M_{i+j}}$ 和 $W_{M_{i+j+k}}$ 分别表示 $W_{i+j}$ 和 $W_{i+j+k}$ 对应的多重集. $i, j, k = 1, 2, 3$. 目标知识网 $W_0 = W_{M_0} = \{a, b, c, d, e, b_1, b_2, c_1, c_2, c_3, d_1, d_2, d_3, e_1, e_2\}$. 另有知识网 $W_1 = W_{M_1} = \{a, b, c, d, b_1, c_1, c_2, d_1, d_2, d_3\}$. $W_2 = W_{M_2} = \{a, b, c, e, b_1, b_2, c_1, c_2, c_3, e_1, e_2\}$. $W_3 = W_{M_3} = \{a, c, d, c_1, c_2, c_3, d_1, d_2, d_3\}$. $W_{M_{1+2}} = \{2a, 2b, 2c, d, e, 2b_1, b_2, 2c_1, 2c_2, c_3, d_1, d_2, d_3, e_1, e_2\}$. $W_{M_{2+3}} = \{2a, b, 2c, d, e, b_1, b_2, 2c_1, 2c_2, 2c_3, d_1, d_2, d_3, e_1, e_2\}$. $W_{M_{1+2+3}} = \{3a, 2b,$

表 1 知识网的基本数据

Table 1 Basic data of knowledge mesh

| | | | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_5$ | $W_6$ | $W_7$ | $W_8$ | $W_9$ | $W_{10}$ | $W_{11}$ | $W_{12}$ | $W_{13}$ | $W_{14}$ | $W_{15}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | b 财务管理 | b1 入帐报表 | 0.9638 | 0.8011 | 0.6571 | 0.6461 | 0.0669 | 0.7560 | 0.5579 | 0.8903 | 0.4875 | 0.5875 | 0.7697 | 0.3326 | 0.5195 | 0.5784 | 0.5277 |
| | | b2 凭证管理 | 0.8634 | 0.7472 | 0.7887 | 0.7697 | 0.0326 | 0.7413 | 0.6495 | 0.8525 | 0.4875 | 0.6875 | 0.8195 | 0.8784 | 0.6440 | 0.5530 | 0.6423 |
| a 管理信息系统 | c 生产管理 | c1 生产监控 | 0.6178 | 0.6011 | 0.6087 | 0.8995 | 0.0784 | 0.8277 | 0.7963 | 0.6842 | 0.5625 | 0.4625 | 0.8440 | 0.6530 | 0.6940 | 0.6659 | 0.6124 |
| | | c2 计划调度 | 0.7550 | 0.5099 | 0.5455 | 0.9440 | 0.0530 | 0.8423 | 0.8963 | 0.7540 | 0.6625 | 0.5625 | 0.8940 | 0.7659 | 0.5989 | 0.5274 | 0.5491 |
| | | c3 物料管理 | 0.8204 | 0.4582 | 0.6091 | 0.9440 | 0.0659 | 0.9124 | 0.6205 | 0.7093 | 0.5776 | 0.5364 | 0.8989 | 0.6274 | 0.5684 | 0.6281 | 0.6349 |
| | d 质量管理 | d1 计量管理 | 0.5010 | 0.6611 | 0.6255 | 0.6989 | 0.0274 | 0.6491 | 0.5679 | 0.5326 | 0.4413 | 0.3495 | 0.6525 | 0.9697 | 0.6427 | 0.7660 | 0.6511 |
| | | d2 质量检查 | 0.7010 | 0.5611 | 0.6251 | 0.6684 | 0.0281 | 0.6349 | 0.6195 | 0.6784 | 0.5277 | 0.4963 | 0.6842 | 0.7195 | 0.8881 | 0.8277 | 0.9044 |
| | | d3 过程控制 | 0.6776 | 0.5236 | 0.7404 | 0.7427 | 0.0660 | 0.6511 | 0.7440 | 0.6530 | 0.5423 | 0.4963 | 0.7540 | 0.8440 | 0.9017 | 0.8365 | 0.9273 |
| | e 设备管理 | e1 设备档案 | 0.6545 | 0.6057 | 0.7792 | 0.7881 | 0.0277 | 0.7444 | 0.6940 | 0.6659 | 0.9124 | 0.8205 | 0.6093 | 0.5940 | 0.5426 | 0.6319 | 0.6020 |
| | | e2 预修预检 | 0.7545 | 0.6057 | 0.7375 | 0.8071 | 0.0365 | 0.8673 | 0.7989 | 0.7274 | 0.9491 | 0.8010 | 0.6525 | 0.7989 | 0.6697 | 0.6285 | 0.7361 |

$3c$, $2d$, $e$, $2b_1$, $b_2$, $3c_1$, $3c_2$, $2c_3$, $2d_1$, $2d_2$, $2d_3$, $e_1$, $e_2\}$. $W_0 = W_{1+2} = W_{2+3} = W_{1+2+3}$.

取权重 $\lambda_i = 1$, log 取以 2 为底, 计算第 1.2 节中的式 (1) 和第 1.3 节中的式 (3). 知识网功能匹配度 $f(W_0, W) = (2 \times$ 满足 $W_0$ 的功能数$)/(W_0$ 的总功能数 $+ W$ 的总功能数). 表 2 列出各知识网的复杂度、匹配度和复杂性系数. 由表 2 可知: $G(W_{1+2+3}) > G(W_{1+2}) > G(W_{2+3}) > G(W_0)$, 而匹配度全为 1, 说明 $W_{1+2+3}, W_{1+2}, W_{2+3}$ 虽然满足用户的功能需求, 但元素多重数大于 1 导致复杂度系数大于 1. 而 $W_1, W_2, W_3$ 复杂度比 $W_0$ 的复杂度值小, 但匹配度小于 1, 说明它们为不能完全匹配的知识网. 就表 2 数据而言, $W_{2+3}$ 为最适合用户需求的知识网. 满足用户需求的知识网应该是具有尽可能高的匹配度, 复杂性系数越接近 1 的知识网.

### 3.2　利用文中方法对知识网进行分类和检索

取 $\tau = 0.7800$, $W_5$ 的所有满意度分量均小于 $\tau$, 所以 $W_5$ 并不参与其后的计算. 相似度的调节系数 $\varepsilon = 10^{-4}$. 计算第 2.1 节中的式 (6) 和第 2.2 节中的式 (7), 不同 $\theta$ 值将得到不同的聚类结果, 分别见表 3 ∼ 5.

由表 5, 当阈值 $\theta = 0.8700$ 时, 聚类中心依次为 $W_4$, $W_{12}$, $W_9$, $W_1$, $W_{15}$. $W_4$ 确定的类为生产管理: 生产监控, 计划调度, 物料管理; $W_{12}$ 确定的类为凭证管理, 计量管理; $W_9$ 确定的类为设备管理, 预修预检; $W_1$ 确定的类为入帐报表; $W_{15}$ 确定的类为质量检查, 过程控制. 聚类中心将用户的 10 个需求分成 5 大类. 各聚类中心在相应需求对应的满意度分量上具有较高的数值, 而且其相似度是随着聚类次数的增加而减少, $Q(W_4) > Q(W_{12}) > Q(W_9) > Q(W_1) > Q(W_{15})$. 计算第 2.3 节中的式 (8) 和式 (9), 取 $\rho = 10$, $\gamma = 8 \times 10^{-8}$, 当 $\theta = 0.8700$ 时, 各聚类中心对应的检索空间范围见表 6. 表 6 中的数据为满意度的三元组表示方法: 上界, 聚类中心的数值, 下界. 由表 6 可得位于各检索空间中的知识网集

表 2　各知识网的复杂度、匹配度和复杂性系数

Table 2　Complexity degree, matching degree and complexity coefficient of each knowledge mesh

| | $W_{1+2+3}$ | $W_{2+3}$ | $W_{1+2}$ | $W_3$ | $W_2$ | $W_1$ | $W_0$ |
|---|---|---|---|---|---|---|---|
| 复杂度 | 117.3258 | 78.5175 | 82.4176 | 29.8974 | 38.0537 | 34.5943 | 59.6000 |
| 匹配度 | 1.0000 | 1.0000 | 1.0000 | 0.75 | 0.8462 | 0.8000 | 1.0000 |
| 复杂性系数 | 1.9686 | 1.3174 | 1.3828 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

表 3　$\theta = 0$ 聚类结果

Table 3　Clustering results when $\theta = 0$

| 聚类 | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_6$ | $W_7$ | $W_8$ | $W_9$ | $W_{10}$ | $W_{11}$ | $W_{12}$ | $W_{13}$ | $W_{14}$ | $W_{15}$ | 中心 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 第 1 次 | 381.3701 | 63.3542 | 182.2824 | 893.7749 | 632.6256 | 240.9258 | 327.6538 | 57.4021 | 33.7003 | 586.5985 | 260.4690 | 152.7988 | 154.9141 | 180.0372 | $W_4$ |

表 4　$\theta = 0.8$ 聚类结果

Table 4　Clustering results when $\theta = 0.8$

| 聚类 | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_6$ | $W_7$ | $W_8$ | $W_9$ | $W_{10}$ | $W_{11}$ | $W_{12}$ | $W_{13}$ | $W_{14}$ | $W_{15}$ | 中心 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 第 1 次 | 381.3701 | 63.3542 | 182.2824 | 893.7749 | 632.6256 | 240.9258 | 327.6538 | 57.4021 | 33.7003 | 586.5985 | 260.4690 | 152.7988 | 154.9141 | 180.0372 | $W_4$ |
| 第 2 次 | 299.8684 | 141.1103 | 2.2429 | 0.0000 | 28.8980 | 48.5514 | 198.1941 | 143.4518 | 122.2015 | 55.9445 | 424.7442 | 282.3961 | 212.8395 | 294.7367 | $W_{12}$ |
| 第 3 次 | 73.0290 | 46.8342 | 0.0644 | 0.0000 | 3.5098 | 3.2495 | 32.4272 | 40.0245 | 13.5379 | 12.6753 | 0.0000 | 43.5224 | 14.3549 | 32.0015 | $W_1$ |
| 第 4 次 | 0.0000 | 0.1840 | 0.0015 | 0.0000 | 0.0028 | 0.0150 | 0.0007 | 3.2376 | 1.5094 | 0.0005 | 0.0000 | 1.6192 | 0.0915 | 0.5704 | $W_9$ |

表 5　$\theta = 0.8700$ 聚类结果

Table 5　Clustering results when $\theta = 0.8700$

| 聚类 | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_6$ | $W_7$ | $W_8$ | $W_9$ | $W_{10}$ | $W_{11}$ | $W_{12}$ | $W_{13}$ | $W_{14}$ | $W_{15}$ | 中心 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 第 1 次 | 381.3701 | 63.3542 | 182.2824 | 893.7749 | 632.6256 | 240.9258 | 327.6538 | 57.4021 | 33.7003 | 586.5985 | 260.4690 | 152.7988 | 154.9141 | 180.0372 | $W_4$ |
| 第 2 次 | 356.5447 | 185.1562 | 18.7258 | 0.0000 | 84.0204 | 76.5802 | 267.8350 | 173.1894 | 122.9895 | 150.8244 | 426.0107 | 305.7628 | 278.9412 | 336.4904 | $W_{12}$ |
| 第 3 次 | 221.9136 | 245.5751 | 4.7872 | 0.0000 | 64.8634 | 15.1586 | 167.1228 | 258.9033 | 123.9391 | 98.5210 | 0.0000 | 223.0637 | 141.2152 | 198.9729 | $W_9$ |
| 第 4 次 | 86.7043 | 44.3770 | 0.1960 | 0.0000 | 11.4600 | 1.3227 | 45.4957 | 0.0000 | 11.9714 | 4.8012 | 0.0000 | 61.5481 | 19.9896 | 84.6767 | $W_1$ |
| 第 5 次 | 0.0000 | 0.1750 | 0.0022 | 0.0000 | 0.0245 | 0.0032 | 0.0066 | 0.0000 | 1.0606 | 0.0002 | 0.0000 | 4.3249 | 0.3263 | 8.1757 | $W_{15}$ |

合分别为 $W_4, W_{11}, W_{12}, W_9, W_1, W_{15}, W_{13}$. 并且数据显示表现类特征的检索区间数值会更高一些, 如 $W_4$ 对应的类为生产管理: 生产监控, 计划调度, 物料管理, 对应的检索区间分别为: [0.8091, 0.8995, 1.0000], [0.8911, 0.9440, 1.0000], [0.8911, 0.9440, 1.0000], 它们明显高于其他区间数值, 说明检索空间的数值是与其类的特征是一一对应的. 利用自重构运算[2], 对这 5 类知识网进行组合优化[3], 可获得满足用户需求的最佳知识网. 图 3 为 $\theta = 0.8700$ 知识网的模糊分类和检索的功能界面.

在实例中, 若对用户的 10 个需求分别进行检索, 得到每个需求下满意度最高的知识网, 依次为 $W_1, W_{12}, W_4, W_4, W_4, W_{12}, W_{15}, W_{15}, W_9, W_9$, 很明显这里有重复的检索过程, 而接下来进行的自重构"并" 运算虽然能得到各需求满意度均为最高的知识网, 但也同样会进行重复的运算过程. 也就是说用这种方法获得新知识网的过程既冗余又复杂. 而运用基于粒度分析原理的算法, 当 $\theta = 0.8700$ 时, 用户进行 5 次检索就可得到 $W_4, W_{12}, W_9, W_1, W_{15}$, 这里没有重复的检索过程. 检索空间中存在的其他知识

表 6 $\theta = 0.8700$ 各聚类中心对应的检索空间

Table 6 Search space corresponding to each clustering center when $\theta = 0.8700$

| | $W_4$ | $W_{12}$ | $W_9$ | $W_1$ | $W_{15}$ |
|---|---|---|---|---|---|
| 入帐报表 | 0.4174,0.6460,1 | 0.1106,0.3326,1 | 0.2377,0.4875,1 | 0.9289,0.9638,1 | 0.2785,0.5277,1 |
| 凭证管理 | 0.5924,0.7697,1 | 0.7716,0.8784,1 | 0.2377,0.4875,1 | 0.7455,0.8634,1 | 0.4125,0.6423,1 |
| 生产监控 | 0.8091,0.8995,1 | 0.4264,0.6530,1 | 0.3164,0.5625,1 | 0.3817,0.6178,1 | 0.3750,0.6124,1 |
| 计划调度 | 0.8911,0.9440,1 | 0.5866,0.7659,1 | 0.4389,0.6625,1 | 0.5700,0.7550,1 | 0.3015,0.5491,1 |
| 物料管理 | 0.8911,0.9440,1 | 0.3936,0.6274,1 | 0.3336,0.5776,1 | 0.6731,0.8204,1 | 0.4031,0.6349,1 |
| 计量管理 | 0.4885,0.6989,1 | 0.9403,0.9697,1 | 0.1947,0.4413,1 | 0.2510,0.5010,1 | 0.4239,0.6511,1 |
| 质量检查 | 0.4468,0.6684,1 | 0.5177,0.7195,1 | 0.2785,0.5277,1 | 0.4914,0.7010,1 | 0.8179,0.9044,1 |
| 过程控制 | 0.5516,0.7427,1 | 0.7123,0.8440,1 | 0.2941,0.5423,1 | 0.4591,0.6776,1 | 0.8599,0.9273,1 |
| 设备档案 | 0.6211,0.7881,1 | 0.3528,0.5940,1 | 0.8325,0.9124,1 | 0.4284,0.6545,1 | 0.3624,0.6020,1 |
| 预修预检 | 0.6514,0.8071,1 | 0.6382,0.7989,1 | 0.9008,0.9491,1 | 0.5693,0.7545,1 | 0.5418,0.7361,1 |

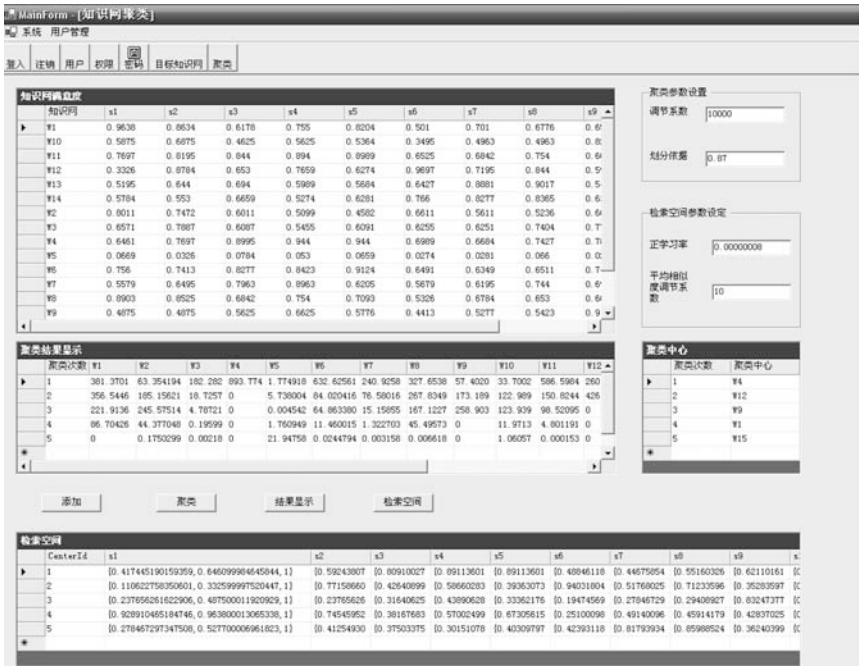


图 3 知识网的模糊分类和检索的功能界面

Fig. 3 The function interface of fuzzy classification and searching for knowledge mesh

网, 如 $W_{11}$, $W_{13}$, 也是具有较高满意度的知识网, 这给用户提供了更多的选择, 便于下一步自重构的组合优化问题的讨论.

### 3.3 文中方法与 FCM 和 FRC 算法的比较分析

因为知识网库中知识网分成几类以及每一类的特征并不确定, 知识网关于目标知识网的数据也并不能完全用确定的数值表示, 所以要实现知识网的分类需要采用模糊聚类的方法, 下面对文中方法、模糊 C-均值 (FCM)[16] 算法和模糊关联聚类 (FRC)[17] 算法进行比较分析.

利用 FCM 对 14 个知识网 ($W_5$ 除外) 进行聚类, 距离函数为欧几里德距离, 模糊化因子 $m = 1.25$, 分类数为 5, 迭代 10 次, 误差为 0.01, 初始隶属度矩阵为随机赋值的矩阵, 结果见表 7. 由表 7, 类中知识网隶属度均都大于 0.9900, 14 个知识网得到很好的划分. 但要从类中找出部分需求下具有较高满意度的知识网却很困难. 因为每个类的特征可能并不明确, 如第 1 类 (满意度分量位于 [0.3495,0.8205]). 即使类特征明确, 如第 2 类 (生产管理方面的满意度分量位于区间 [0.8277,0.9440]), 还需要进一步确定各知识网关于目标知识网满意度的排序. 而文中方法是以知识网库中知识网为聚类中心, 类特征以及聚类数都可以以此确定, 这是 FCM 无法解决的问题.

与 FCM 不同, 建立模糊关联矩阵的模糊关联聚类 (FRC) 算法能够挖掘类内知识网的特征, 采用 FRC 对 14 个知识网 ($W_5$ 除外) 进行聚类. 将表 1 中两两知识网的数据进行比较, 转化为 $14 \times 14$ 的模糊关联矩阵 $R$. 求 $G_{14\times5}$, 满足 $R = G \circ G^{\mathrm{T}}$. 取算子为模糊集合的 $s-t$ 卷积, $t(x,y) = xy$, $s(x,y) = x + y - xy$. 采用基于梯度的方法分解 $R$. 取学习速率为 0.047, 分类数为 5, 迭代 500 次, 结果见表 8. 取隶属度大于 0.8500, 聚类结果与文中检索空间的知识网集合类似. 但若取每个类中隶属度最高的知识网, 依次为 $W_4, W_8, W_{10}, W_{12}, W_{15}$, 并不都是在相应需求下总满意度最高的知识网. 因此 FRC 算法仍旧需要对类中知识网满意度进行排序. 另外, FRC 需要对知识网的基本数据进行两两比较, 因此比文中方法计算复杂.

## 4 结论

随着各种先进制造模式的不断涌现, 知识化制造系统知识网库中的知识网数量会越来越多. 要实现对若干种先进制造模式的优势互补, 知识网的检索问题是必须解决的问题. 本文从粒度原理的角度, 提出一种知识网的模糊分类与检索方法. 该方法的实质是按照用户需求, 将知识网库中的知识网进行聚类, 从而达到分类和排序的目的. 用户是在类中选择知识网, 而不是对每种需求都进行选择, 检索次数和自重构运算次数的减少使得最终新知识网的复杂性降低, 更符合实际需求. 构造的相似度不仅考虑量上的匹配, 还考虑更深层次的质和复杂程度的匹配. 提出的知识网复杂度也可以应用到知识网其他优化问题的研究中. 最终得到的新知识网不一定是各需求下满意度均为最高的知识网, 但却是在某些需求下总满意度最高的知识网. 检索空间的构造, 使得不同知识网之间的关系更加清晰,有利于新知识网的冗

表 7　FCM 聚类结果

Table 7　FCM clustering results

| 类 | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_6$ | $W_7$ | $W_8$ | $W_9$ | $W_{10}$ | $W_{11}$ | $W_{12}$ | $W_{13}$ | $W_{14}$ | $W_{15}$ | 结果 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0000 | 1.0000 | 0.9998 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9959 | 0.0000 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | $W_2, W_3, W_{10}$ |
| 2 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | 0.0022 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | 0.0003 | 0.0000 | 0.0000 | 0.0000 | $W_4, W_6, W_{11}$ |
| 3 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9975 | 0.0000 | 0.9999 | 0.0040 | 0.0000 | 0.0002 | 0.0000 | 0.0000 | 0.0000 | $W_7, W_9$ |
| 4 | 0.0000 | 0.0000 | 0.0001 | 0.0000 | 0.0000 | 0.0001 | 0.0000 | 0.0001 | 0.0001 | 0.0000 | 0.9994 | 1.0000 | 1.0000 | 1.0000 | $W_{12}, W_{13}, W_{14}, W_{15}$ |
| 5 | 0.9999 | 0.0000 | 0.0001 | 0.0000 | 0.0000 | 0.0001 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | $W_1, W_8$ |

表 8　FRC 聚类结果

Table 8　FRC clustering results

| 类 | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_6$ | $W_7$ | $W_8$ | $W_9$ | $W_{10}$ | $W_{11}$ | $W_{12}$ | $W_{13}$ | $W_{14}$ | $W_{15}$ | 结果 | 隶属度 > 0.8500 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.1702 | 0.5637 | 0.5736 | 0.0742 | 0.0183 | 0.3699 | 0.2054 | 0.4594 | 0.4704 | 0.1343 | 0.5890 | 0.9297 | 0.9032 | 0.9458 | $W_{13}, W_{14}, W_{15}$ | $W_{13}, W_{14}, W_{15}$ |
| 2 | 0.5175 | 0.7183 | 0.7667 | 0.9258 | 0.8855 | 0.7891 | 0.5776 | 0.2087 | 0.2832 | 0.9197 | 0.5983 | 0.7276 | 0.7372 | 0.7222 | $W_4, W_6, W_7, W_{11}$ | $W_4, W_6, W_{11}$ |
| 3 | 0.9944 | 0.7752 | 0.7730 | 0.4286 | 0.5813 | 0.5787 | 1.0000 | 0.4921 | 0.5379 | 0.6803 | 0.6921 | 0.6691 | 0.6775 | 0.6944 | $W_1, W_2, W_8$ | $W_1, W_8$ |
| 4 | 0.6268 | 0.0000 | 0.4837 | 0.7069 | 0.6211 | 0.7860 | 0.6215 | 0.5449 | 0.4105 | 0.6208 | 0.8651 | 0.6702 | 0.6344 | 0.6731 | $W_{12}$ | $W_{12}$ |
| 5 | 0.5537 | 0.7187 | 0.8152 | 0.6840 | 0.7109 | 0.7767 | 0.6047 | 0.9434 | 0.9645 | 0.5717 | 0.5042 | 0.6169 | 0.6197 | 0.6442 | $W_3, W_9, W_{10}$ | $W_9, W_{10}$ |

余、复杂性等问题的讨论. 实例显示, 该方法效果明显, 与其他模糊聚类方法相比较, 文中方法不需要指定类的个数, 并且可以得到关于目标知识网满意度的排序, 而其他聚类方法需要给定分类数, 隶属度显示的是对类的归属情况, 不能达到关于满意度的排序, 这是本文方法优于其他模糊聚类方法的地方. 但对知识网间的完善度和匹配度的获得还需要更详细的讨论.

## 附录

**命题 1.** $G(W)$ 是元素个数和多重数的单增函数; 当知识网的元素多重数非负, 知识网的交和差运算不会使得复杂度增加, 知识网的并运算不会使得复杂度减少.

**证明.** 1) 先证 $G(W)$ 是元素个数的单增函数. 设知识网 $W = \{x_1, x_2, \cdots, x_n\}$ 和知识网 $\tilde{W} = \{x_1, x_2, \cdots, x_n, x_{n+1}\}$ 的元素 $x_i$ 的多重数为 $\alpha_i$ ($\alpha_i > 0$), $\lambda_i$ 表示权重, $i = 1, 2, \cdots, n+1$. 设 $\lambda_i\alpha_i = c_i$, $\sum_{i=1}^{n} c_i = m$.

$$G(W) - G(\tilde{W}) =$$

$$\sum_{i=1}^{n} c_i \log\left(\frac{m}{c_i} + 1\right) - \sum_{i=1}^{n+1} c_i \log\left(\frac{m + c_{n+1}}{c_i} + 1\right) =$$

$$\log \prod_{i=1}^{n} \left(\frac{m + c_i}{m + c_{n+1} + c_i}\right)^{c_i} \times \left(\frac{c_{n+1}}{m + c_{n+1} + c_{n+1}}\right)^{c_{n+1}}$$

由于等式 $\prod$ 中的每一项 $\dfrac{m + c_i}{m + c_{n+1} + c_i} \leq 1$, 且 $\dfrac{c_{n+1}}{m + c_{n+1} + c_{n+1}} \leq 1$, 所以 $G(W) - G(\tilde{W}) \leq 0$.

2) 再证 $G(W)$ 是多重数的单调函数. 设知识网 $W = \{x_1, x_2, \cdots, x_n\}$ 和知识网 $\tilde{W} = \{x_1, x_2, \cdots, x_n\}$ 的元素 $x_i$ 的多重数分别为 $\alpha_i$, $\tilde{\alpha}_i > 0$, 且 $\tilde{\alpha}_1 = k\alpha_1$ ($k \geq 1$), $\tilde{\alpha}_i = \alpha_i$, $i = 2, \cdots, n$, $\lambda_i\alpha_i = c_i$, $\sum_{i=1}^{n} c_i = m$, $\tilde{m} = \sum_{i=1}^{n} \tilde{c}_i = m + (k-1)c_1$,

$$G(W) - G(\tilde{W}) = \sum_{i=1}^{n} c_i \log(\frac{m}{c_i} + 1) - \sum_{i=1}^{n} \tilde{c}_i \log(\frac{\tilde{m}}{\tilde{c}_i} + 1) =$$

$$\log \prod_{i=2}^{d} \left(\frac{m + c_i}{m + (k-1)c_1 + c_i}\right)^{c_i} \times \frac{(m + c_1)^{c_1}(kc_1)^{kc_1}}{c_1^{c_1}[m + (2k-1)c_1]^{kc_1}}$$

$$\because \frac{m + c_i}{m + (k-1)c_1 + c_i} \leq 1, \ \frac{(m + c_1)^{c_1}(kc_1)^{kc_1}}{c_1^{c_1}[m + (2k-1)c_1]^{kc_1}} =$$

$$\left\{\frac{1 + \dfrac{m}{c_1}}{\left[\dfrac{m}{kc_1} + \dfrac{2k-1}{k}\right]^k}\right\}^{c_1} \leq \left\{\frac{1 + \dfrac{m}{c_1}}{\left[\dfrac{m}{kc_1} + 1\right]^k}\right\}^{c_1} \leq 1$$

$$\therefore G(W) - G(\tilde{W}) \leq 0$$

3) 由 $G(W)$ 是元素个数和多重数的单增函数可知, 当多重数非负时, 知识网的交和差运算不会使得复杂度增加和知识网的并运算不会使得复杂度减少, 这两个结论均是成立的. □

**命题 2.** 1) 当 $\varepsilon = 1$, $0 \leq sim(\cdot) \leq 1$. 当且仅当对于 $\forall i$, $\mu_{W(v_i)} = f(v_i, W) = g(V, W) = 1$ 时, $sim(\cdot) = 1$. 当且仅当 $\exists i$, $\mu_{W(v_i)} f(v_i, W) = 0$ 时, $sim(\cdot) = 0$.

2) $sim(\cdot)$ 分别是 $\alpha^i_{W(V)}$, $\omega_i^2$ 的单调递增函数, 是 $g(V, W)$ 的单调递减函数.

3) 设知识网 $U$, $V$ 和 $W$, 若 $W \subseteq U$, $g(V, U) = 1$, 则 $sim(V, W, \mathbf{\Omega}) \leq sim(V, U, \mathbf{\Omega})$; 若 $g(V, U + W) = 1$, 则 $\max\{sim(V, W, \mathbf{\Omega}), sim(V, U, \mathbf{\Omega})\} \leq sim(V, U + W, \mathbf{\Omega})$, 当且仅当 $W \subseteq U$, 不等式中的 "=" 成立; 若 $g(V, U) = 1$, 则 $sim(V, U \cdot W, \mathbf{\Omega}) \leq sim(V, U, \mathbf{\Omega})$, 当且仅当 $W = U$, 不等式中的 "=" 成立.

**证明.** 1) 将 $T(\cdot)$, $s(\cdot)$ 代入第 1.3 节中式 (2),

$$sim(V, W, \mathbf{\Omega}) =$$

$$\frac{\prod_{i=1}^{m} (\mu_{W(v_i)} f(v_i, W) + \omega_i^2 - \mu_{W(v_i)} f(v_i, W)\omega_i^2))}{\varepsilon g(V, W)}$$

其中, $0 \leq \mu_{W(v_i)} \leq 1$, $0 \leq f(v_i, W) \leq 1$, $g(V, W) \geq 1$.

$$\because \ 0 \leq \mu_{W(v_i)} f(v_i, W) + \omega_i^2 - \mu_{W(v_i)} f(v_i, W)\omega_i^2 =$$

$$\mu_{W(v_i)} f(v_i, W) + (1 - \mu_{W(v_i)} f(v_i, W))\omega_i^2 \leq$$

$$\mu_{W(v_i)} f(v_i, W) + (1 - \mu_{W(v_i)} f(v_i, W)) = 1$$

$$\therefore \text{当 } \varepsilon = 1 \text{ 时, } 0 \leq sim(\cdot) \leq 1 \tag{10}$$

当且仅当对于 $\forall i$, $\mu_{W(v_i)} = f(v_i, W) = g(V, W) = 1$ 时, 有 $\mu_{W(v_i)} f(v_i, W) + \omega_i^2 - \mu_{W(v_i)} f(v_i, W)\omega_i^2 = 1$ 和 $sim(\cdot) = 1$. 当且仅当 $\exists i$, $\mu_{W(v_i)} f(v_i, W) = 0$ 时, 有 $\omega_i^2 = 0$, $\mu_{W(v_i)} \times f(v_i, W) + \omega_i^2 - \mu_{W(v_i)} f(v_i, W)\omega_i^2 = 0$, 则 $sim(\cdot) = 0$.

2) 因为 $\mu_{W(v_i)} f(v_i, W) = \alpha^i_{W(V)}$, $0 \leq \alpha^i_{W(V)} \leq 1$, $0 \leq 1 - \alpha^i_{W(V)} \leq 1$, $0 \leq 1 - \omega_i^2 \leq 1$, $sim(V, W, \mathbf{\Omega})$ 分别是 $\alpha^i_{W(V)}$, $\omega_i^2$ 和 $g(V, W)$ 的单调函数.

3) $W \subseteq U$, $(W \cap V) \subseteq (U \cap V)$, $\alpha^i_{W(V)} \leq \alpha^i_{U(V)}$. 又 $g(V, U) = 1$, 有 $g(V, W) = 1$. 由命题 2.2, 可知 $sim(V, W, \mathbf{\Omega}) \leq sim(V, U, \mathbf{\Omega})$. 又因为 $W \subseteq U + W$, $U \subseteq U + W$, $U \cdot W \subseteq U$, 所以 $\max\{sim(V, W, \mathbf{\Omega}), sim(V, U, \mathbf{\Omega})\} \leq sim(V, U + W, \mathbf{\Omega})$, $sim(V, U \cdot W, \mathbf{\Omega}) \leq sim(V, U, \mathbf{\Omega})$. □

## References


1 Yan Hong-Sen, Liu Fei. Knowledgeable manufacturing system — a new kind of advanced manufacturing system. *Computer Integrated Manufacturing Systems*, 2001, **7**(8): 7−11
(严洪森, 刘飞. 知识化制造系统 — 新一代先进制造系统. 计算机集成制造系统, 2001, **7**(8): 7−11)

2 Yan H S. A new complicated-knowledge representation approach based on knowledge meshes. *IEEE Transactions on Knowledge and Data Engineering*, 2006, **18**(1): 47−62

3 Xue Chao-Gai, Yan Hong-Sen. Research of automatic construction of the knowledge mesh based on the user's functional requirement. *Control and Decision*, 2005, **20**(9): 996−1001
(薛朝改, 严洪森. 基于用户功能需求的知识网的自动生成研究. 控制与决策, 2005, **20**(9): 996−1001)

4 Yan H S, Xue C G. Decision-making in self-reconfiguration of a knowledgeable manufacturing system. *International Journal of Production Research*, 2007, **45**(12): 2735−2758

5 Qiu Tao-Rong, Liu Qing, Huang Hou-Kuan. A granular computing approach to knowledge discovery in relational database. *Acta Automatica Sinica*, 2009, **35**(8): 1071−1079

6 Sun Liang, Han Chong-Zhao, Shen Jian-Jing, Dai Ning. Generalized rough set method for ensemble feature selection and multiple classifier fusion. *Acta Automatica Sinica*, 2008, **34**(3): 298−304
(孙亮, 韩崇昭, 沈建京, 戴宁. 集成特征选择的广义粗集方法与多分类器融合. 自动化学报, 2008, **34**(3): 298−304)

7 Chen Jie, Wu Di, Zhang Juan. Distributed simulation system hierarchical design model based on quotient space granular computation. *Acta Automatica Sinica*, 2010, **36**(7): 923−930
(陈杰, 吴狄, 张娟. 分布式仿真系统层次设计商空间粒计算模型. 自动化学报, 2010, **36**(7): 923−930)

8 Norvag K. Granularity reduction in temporal document databases. *Information Systems*, 2006, **31**(2): 134−147

9 Gao Q, Li M, Vitanyi P. Applying MDL to learn best model granularity. *Artificial Intelligence*, 2000, **121**(1−2): 1−29

10 Fan Z P, Yang L. A method for group decision-making based on multi-granularity uncertain linguistic information. *Expert Systems with Applications*, 2010, **37**(5): 4000−4008

11 Deshmukh A V, Talavage J J, Barash M M. Complexity in manufacturing systems, part I: analysis of static complexity. *IIE Transactions*, 1998, **30**(7): 645−655

12 Frizelle G, Woodcock E. Measuring complexity as an aid to developing operational strategy. *International Journal of Operations and Production Management*, 1995, **15**(5): 26−39

13 Zhang Z F, Xiao R B. Empirical study on entropy models of cellular manufacturing systems. *Progress in Natural Science*, 2009, **19**(3): 389−395

14 Yang Ren-Zi, Yan Hong-Sen. Structure of knowledge mesh in knowledge-oriented manufacturing system. *Computer Integrated Manufacturing Systems*, 2008, **14**(3): 595−601
(杨人子, 严洪森. 知识化制造系统中知识网的结构研究. 计算机集成制造系统, 2008, **14**(3): 595−601)

15 Ding Xue-Feng, Yan Hong-Sen, Xue Chao-Gai. Self-reconfiguration of knowledgeable manufacturing system based on approximate match. *Control and Decision*, 2008, **23**(1): 70−74
(丁雪峰, 严洪森, 薛朝改. 基于近似匹配的知识化制造系统自重构研究. 控制与决策, 2008, **23**(1): 70−74)

16 Bezdek J C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981

17 Pedrycz W. Classification of relational patterns as a decomposition problem. *Pattern Recognition Letters*, 1996, **17**(1): 91−99

**杨人子** 东南大学自动化学院博士研究生. 主要研究方向为知识化制造系统. 本文通信作者. E-mail: yrz@seu.edu.cn
(**YANG Ren-Zi** Ph. D. candidate at the School of Automation, Southeast University. Her research interest covers knowledgeable manufacturing systems. Corresponding author of this paper.)

**严洪森** 东南大学自动化学院教授. 主要研究方向为知识化制造系统, 计算机集成制造系统和生产计划与调度.
E-mail: hsyan@seu.edu.cn
(**YAN Hong-Sen** Professor at the School of Automation, Southeast University. His research interest covers knowledgeable manufacturing systems, computer integrated manufacturing systems, and production planning and scheduling.)