

文章编号: 1002-0411(2001)02-129-06

分布式知识系统中基于粗糙集合的查询策略研究

马志锋¹ 邢汉承¹ 郑晓妹²

(1. 东南大学计算机科学与工程系 南京 210096; 2. 南京航空航天大学计算机科学与工程系 南京 210016)

摘要: 传统数据库中的数据查询通常忽视了对于知识的应用. 随着网络技术的发展, 数据被分布在不同的物理节点上, 用户已不太可能对所有数据均有全面的了解. 本文讨论了分布式知识系统中基于粗糙集合的查询策略, 强调了系统中知识的重要性. 它可大大地降低网络节点间的通信代价, 从而提高查询效率. 粗糙集合理论的应用使得查询不必局限于传统的精确匹配式查询, 而可以是一种粗糙、模糊匹配式的查询.*

关键词: 分布式知识系统; 查询策略; 粗糙集合

中图分类号: TP18

文献标识码: A

RESEARCH ON ROUGH SET BASED QUERY STRATEGY IN DISTRIBUTED KNOWLEDGE SYSTEM

MA Zhi-feng¹ XING Han-cheng¹ ZHENG Xiao-mei²

(1. Dept. of Computer Science and Engineering, Southeast University, Nanjing 210096;

2. Dept. of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016)

Abstract: It has usually ignored the usage of knowledge when querying data in a traditional database. By the rapid development of network, data are distributed to different physical sites. It is almost impossible for the users to know the data wholly in such complicated cases. In this paper, we present a rough set-based query strategy in a distributed knowledge system and concentrate on the role of knowledge. It can reduce the communication costs among the different sites of network, and enhance the query efficiency accordingly. Rough set theory makes the query rough or fuzzy, not being precise as in the conventional case.

Keywords: distributed knowledge system, query strategy, rough sets

1 引言(Introduction)

近年来, 随着计算机网络的推广和普及, 特别是 Internet 网络在全球范围内的兴起, 使得网络在人们的日常生活中几乎无所不及. 伴随着网络的迅速崛起, 知识系统(Knowledge System, KS)的应用也因此有了新的特点: 分布式知识系统(Distributed Knowledge System, DKS). DKS 是以分布式数据库(Distributed Database, DDB)为其核心组成部分的, 对于 DKS 的查询强调了在 DDB 中知识的应用. DDB 把分布在计算机网络的不同节点上的数据逻辑上看成属于同一个数据库系统^[1]. 网络中的各个节点应该具有自治的(Autonomous)处理能力, 并且能够执行本地的应用, 而对于用户来说这种操作却是透明的(Transparent), 系统完全有能力确定哪些查询操作是本地无法执行而需提交给其它节点处理的. 当前, DKS/DDB 已经发展成为信息处理中的一

个重要分支, 它的作用正随着全球信息一体化进程的不断深入推进而得到进一步的巩固和加强.

分布式查询策略一直是 DKS/DDB 领域中的一个研究热点之一^[1]. 现有的分布式查询策略大多是针对数据库中元组的“精确”匹配式查询, 而在实际应用中, 往往并不一定需要完全地“精确”匹配, 人们可能只需要得到一个“差不多”、“模糊”的结果. 本文无意于介绍现有分布式查询策略的具体实现, 也不准备过多地评述其性能优劣, 而是要建立起一种新型的基于粗糙集合(Rough Sets, RS)理论的智能型查询策略. 采用这一新策略可大大降低通信代价并能有效地提高查询执行效率.

2 分布式知识系统(DKS)(Distributed knowledge system)

DKS 是一组自治的且能与其它节点上的 KS

协同工作的KS集合^[2,3]. 每一个KS包含了: 本地数据库、知识库(Knowledge Base, KB)以及基于Client/Server结构的智能查询系统(Intelligent Query System, IQS). 其中, KB又由领域知识库(Domain KB)、本地知识库(Local KB)和相邻节点知识库(Neighboring KB)所组成. 为了充分利用好决策知识在查询中所起的作用, IQS除了常规的数据库查询之外, 还要特别强调的是各个节点本身以及与其相邻节点的知识库中所存放的决策信息.

定义1 知识库 $KB = DK \cup LK \cup IK$. 其中, DK 为领域知识(Domain Knowledge), LK 为本地知识(Local Knowledge), IK 为与相邻节点间的交互知识(Interactive Knowledge).

定义2 信息系统(Information System, IS)是由一个四元组 (U, A, V, ρ) 组成的. 其中, U 为对象的集合, A 为非空属性集合, $V = \bigcup_{a \in A} V_a$ 称作属性值域, 且对 $\forall a, b \in A, a \neq b$ 有 $V_a \cap V_b = \Phi$; $\rho: U \times A \rightarrow V$ 称为信息函数, 对 $\forall x \in U, a \in A$, 满足 $\rho(x, a) \in V_a$.

本地数据库可以看作是一种信息系统. 知识库中存放的是一组由IS所生成的决策规则, 每条规则有相对应的准确度(Accuracy)和覆盖率(Coverage)因子以确定其可用性. 根据准确度因子, 可将规则划分为确定与可能规则两大类. 波兰学者 Z Pawlak 提出了一种新型的处理不确定性问题的数学工具RS理论^[4], 它具有很强的定性分析能力, 且不需象其它理论那样事先给出某些特征的数量描述. J W Grzymala-Busse 开发出了基于RS的事例学习系统LERS^[5], A Hrm 根据RS基本算法研制了一个可用于决策分类以及对信息表中数据间相互依赖性进行分析的软件包ROSETTA^[6].

定义3 分布式知识系统 $DKS = (\{KS_i\}_{i \in I}, L)$, 见图1所示, 其中:

$KS_i = IS_i \cup KB_i \cup IQS_i$. $IS_i = (U_i, A_i, V_i, \rho_i)$ 表示第 $i \in I$ 个节点的本地数据库(信息系统).

L 是一个定义在集合 I 上的对称二元关系, 表示了相互间有联结关系的网络节点.

I 是 DKS 中的网络节点集合.

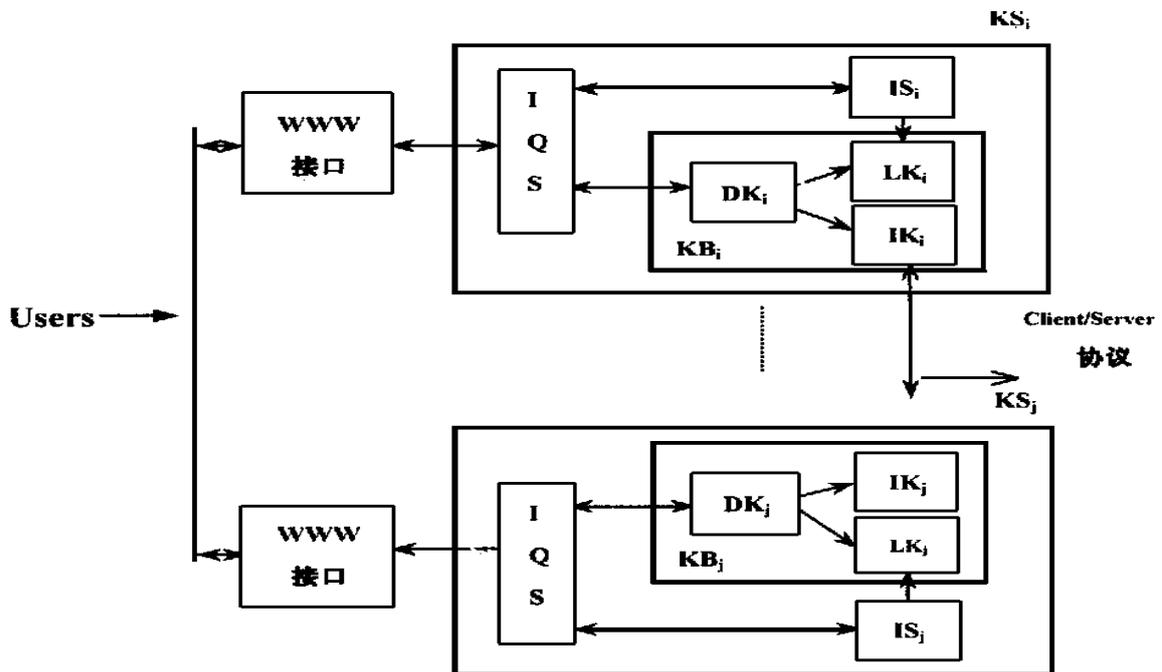


图1 分布式知识系统DKS示意图

Fig. 1 An illustration of distributed knowledge system (DKS)

定义4 两知识系统 KS_i 与 $KS_j (i, j \in I)$ 称为相邻的, 当且仅当 $(i, j) \in L$.

设 $DKS = (\{KS_i\}_{i \in I}, L)$ 是一分布式知识系统, L 在集合 I 中的传递闭包记作 L^+ . 若有 $G_i \subseteq I$ 且 $G_i \in I/L^+$ 是由传递闭包所确定的关于节点 i 的等价

类, 则 G_i 是与 i 相邻的节点组. 相应地把 $\{DKS_{ki}\}_{k \in G_i}$ 称为由节点 k 所生成的DKS中的知识系统组, 事实上它对应了节点 k 与 i 间的交互知识. 因此有 $KB_i = \bigcup \{DKS_{ki} | k \in G_i\} \cup R_i \cup DK_i$, 其中 $R_i = LK_i$ 为存放于节点 i 的本地规则集合.

定义 5 若 $(\forall i)(\forall j)(\forall x \in U_i \cap U_j)(\forall a \in A_i \cap A_j)\{[(x, a) \in \text{Dom}(\rho_i) \cap \text{Dom}(\rho_j)] \rightarrow [\rho_i(x, a) = \rho_j(x, a)]\}$, 则称 DKS 是相容的, 否则为不相容的.

3 粗糙集合(RS)理论准备(Rough sets theory preliminaries)

RS 理论是智能信息处理领域中新近发展起来的一种有效数学工具. 目前, 它已在专家系统、智能信息检索、机器学习、知识获取等多个领域中取得了成功的应用^[4,5,7-9]. 从有关数据库中提取重要知识, 并将进行必要约简后的知识存储于知识库中是 RS 理论的核心内容之一.

设 $U \neq \Phi$ 为关于查询对象的论域, $x \in U$ 为论域中的对象, π 为 U 上的二元等价关系, $[x]_{U/\pi}$ 称作 x 关于 π 的等价类. 若 $X \subseteq U$, 则有:

定义 6 $\pi^- X = \bigcup \{[x]_{U/\pi} \mid [x]_{U/\pi} \subseteq X\}$ 称为关于 X 的 π^- 下近似集(π^- Lower Approximation); $\pi^+ X = \bigcup \{[x]_{U/\pi} \mid x \in X\}$ 称为关于 X 的 π^+ 上近似集(π^+ Upper Approximation); $\text{POS}_\pi(X) = \pi^- X$, $\text{NEG}_\pi(X) = U - \pi^+ X$ 和 $\text{BN}_\pi(X) = \pi^+ X - \pi^- X$ 分别称作 X 的 π 正区域、负区域和边界域(π -Positive Region, π -Negative Region, π -Boundary Region).

定义 7 二元组 $(\pi^- X, \pi^+ X)$ 称为 U 上的基于 π 的 RS.

4 基于 RS 的知识库生成(RS-Based knowledge base generation)

知识库中所存放的主要是从 IS 中提取出来的决策规则. 由于知识库中的规则要用来为智能查询服务, 所以这些规则必须具有较强的适应性, 为此本节对生成知识库的传统 RS 从两方面作了一些改进. 首先, 将传统 RS 所基于的等价关系推广到相似关系(自反、对称), 这样可以使得知识的“颗粒”不至于太小. 其次, X Hu 等在文献[10]中提出了采用概念树的方法来爬升概念, IS 中的较为特殊的概念随着爬升而更具普遍性, 所生成的规则也更为通用. 概念树及相似关系在分布式环境下特别重要, 因为这样的规则不致于过份具体化. 它们的建立均需要有 DK 的支持.

4.1 采用相似关系 RS 生成决策规则

设在 $U \times U$ 空间上, 定义任意两个对象之间的相似关系 $\mathbf{P}_a \in V_a \times V_a$, 其中 $a \in A$, \mathbf{P}_a 满足:

- 自反性: 对于 $\forall v \in V_a$, 均有 $\mathbf{P}_a(v, v)$ 成立.

- 对称性: 对于 $\forall v, v' \in V_a$, 如果 $\mathbf{P}_a(v, v')$, 那么 $\mathbf{P}_a(v', v)$ 也同时成立.

定义 8 称二个对象 x, y 在属性 $a \in A$ 上相似 $\Leftrightarrow \mathbf{P}_a(\rho(x, a), \rho(y, a))$ 成立.

定义 9 称 x, y 相似 $\mathbf{P}_A(x, y) \Leftrightarrow f(A, \lambda_a(\rho(x, a), \rho(y, a)))$ 满足给定的约束条件. 其中 $\lambda_a(\cdot, \cdot)$ 为相似度量函数, $f(\cdot, \cdot)$ 为对象相似的整体约束条件^[11].

定义 10 $\mathbf{P}^- X = \bigcup_{x \in U} \{[x]_{\mathbf{P}} \mid [x]_{\mathbf{P}} \subseteq X\}$ 称为集合 X 的基于相似关系 \mathbf{P}_A 的下近似集.

$\mathbf{P}^+ X = \bigcup_{x \in U} \{[x]_{\mathbf{P}} \mid [x]_{\mathbf{P}} \cap X \neq \Phi\}$ 称为集合 X 的基于相似关系 \mathbf{P}_A 的上近似集.

$\text{BN}_{\mathbf{P}} X = \mathbf{P}^+ X - \mathbf{P}^- X$ 称为集合 X 的基于相似关系 \mathbf{P}_A 的边界区域.

如同在采用等价关系中一样, 基于相似关系的决策规则也采用最为通用且易于理解的“if $\sim \alpha$ then $\sim \beta$ ”形式, 记作“ $\sim \alpha \rightarrow \sim \beta$ ”. 其中 \sim 表示近似于的含义, 即近似 α 则近似 β . 设 $B \subseteq A$, $x \in U$, 有 $\bigwedge_{b \in B} \rho(x, b) = m_b \Leftrightarrow \rho(x, d) = m_d^0 \vee \rho(x, d) = m_d^1 \vee \dots \vee \rho(x, d) = m_d^k$. 如果式中 $k=0$, 即规则的右端仅含一个析取项, 则称此规则为确定规则, 否则称作可能规则.

为了更加精确地描述基于相似关系决策规则的不确定性, 可以分别从条件概率 $P(\sim \beta \mid \sim \alpha)$ 及 $P(\sim \alpha \mid \sim \beta)$ 的角度来对不确定性进行分析. 通常, 某条规则“ $\sim \alpha \rightarrow \sim \beta$ ”的满足因子(Satisfaction)是指信息系统中同时满足条件 $\sim \alpha$ 及决策 $\sim \beta$ 的决策对象个数, 即 $\text{satisfaction}(\sim \alpha \rightarrow \sim \beta) = \text{count}(\sim \alpha \wedge \sim \beta)$. 于是有关于相似规则的准确度(Accuracy)及覆盖率(Coverage)概念的描述如下:

$$\begin{aligned} \text{accuracy}(\sim \alpha \rightarrow \sim \beta) &= \frac{\text{satisfaction}(\sim \alpha \rightarrow \sim \beta)}{\text{satisfaction}(\sim \alpha)} \\ &= P(\sim \beta \mid \sim \alpha); \\ \text{coverage}(\sim \alpha \rightarrow \sim \beta) &= \frac{\text{satisfaction}(\sim \alpha \rightarrow \sim \beta)}{\text{satisfaction}(\sim \beta)} \\ &= P(\sim \alpha \mid \sim \beta). \end{aligned}$$

定理 1 设集合 $X \subseteq U$, π 和 \mathbf{P} 分别为等价与相似关系, \mathbf{P} 为 π 的扩展, 即满足:

- $\forall u \in U, [u]_{\pi} \subseteq [u]_{\mathbf{P}}$,
- $\forall u \in U$, 若 $\forall v \in [u]_{\pi}$, 则 $[v]_{\pi} \subseteq [u]_{\mathbf{P}}$.

那么, 下列各式同时成立:

$$\textcircled{1} \mathbf{P}^- X \subseteq \pi^- X; \textcircled{2} \mathbf{P}^+ X \supseteq \pi^+ X;$$

$$\textcircled{3} \text{BN}_{\pi} X \subseteq \text{BN}_{\mathbf{P}} X.$$

证明: 参见文献[11].

4.2 采用概念树生成决策规则^[10]

概念树是 IS 中属性取值上的一组多对一的关系集合. 每个多对一关系则定义为一组可穷举值与一个更一般的概念间的联系. 例如, $\{A_1, A_2, \dots, A_m\} \rightarrow G$ 表示 G 在概念树中是比 $\{A_1, A_2, \dots, A_m\}$ 更高一层的父概念, 即 G 是 $\{A_1, A_2, \dots, A_m\}$ 的泛化.

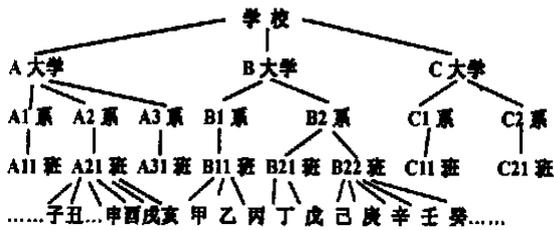


图 2 概念树举例

Fig. 2 An example of hierarchical concept tree

表 1 学生成绩举例

Table 1 An illustration of student reports

学生	年龄	生源	成绩	学生	年龄	生源	成绩
1 甲	30	新疆	差	22 子	31	云南	差
4 乙	32	贵州	一般	30 丑	30	新疆	一般
6 丙	20	新疆	好	35 寅	26	上海	一般
8 丁	20	江苏	好	38 卯	26	上海	一般
9 戊	21	浙江	好	44 辰	20	北京	好
10 己	21	江苏	好	48 巳	21	北京	好
12 庚	26	浙江	一般	49 午	21	江苏	好
14 辛	20	江苏	好	52 未	26	浙江	一般
17 壬	20	江苏	好	54 申	30	新疆	差
18 癸	21	江苏	好	58 酉	27	云南	一般

表 1 给出了关于学生的成绩信息表, 其中学生属性标明了学号. 图 2 是表 1 中有关概念根据领域知识的概念树举例. 很显然, 若直接从学生个人上来发现规则, 极有可能得不到任何有意义的规则, 然而在本例中如果进行概念爬升 2 层, 学生 {子, 丑, 寅, ..., 申, 酉} 用“ A 2 系” 来取代, 学生 {甲, 乙, 丙} 用“ B 1 系” 来替代, 而学生 {丁, 戊, 己, 庚, 辛, 壬, 癸, } 用“ B 2 系” 来替换, 于是可以有类似于这样的知识: “ B 2 系的学生成绩好 (with accuracy = ★ and coverage = ☆)”. 另外, 根据领域常识, 信息系统中的学生学号在分析时是不可能提供任何知识的, 为此, 在进一步分析之前可以将其略去.

为了更好地利用 RS 理论的有关算法对信息系统进行有效地分析, 一般还需把年龄这样的实值属性离散化为若干子区间. 假设 $a \in A$ 的值域为 $[\alpha, \beta]$, 将 $[\alpha, \beta]$ 划分成 m 个子区间: $\{[\mathcal{Y}_0, \mathcal{Y}_1), [\mathcal{Y}_1, \mathcal{Y}_2),$

$\dots, [\mathcal{Y}_{m-1}, \mathcal{Y}_m]\}$, 其中 $\mathcal{Y}_0 = \alpha$ 且 $\mathcal{Y}_m = \beta$, 每个子区间对应了一个语义值描述 (某些方法允许各子区间存在交叠). 较为常用的区间划分策略有:

区间等分策略. 将输入空间等分成 m 个子区间, 特别适用于输入空间的值为均匀分布的情况.

等对象区间策略. 用户首先指出需要划分的区间个数, 使每个区间包含平均对象个数, 然后据此去确定区间边界. 若对象分布均匀时, 本策略就是上面的区间等分法.

信息熵策略. 首先找出划分点 p 使得信息熵 $E(A, U, p) = |S_1| * Ent(S_1) / |U| + |S_2| * Ent(S_2) / |U|$ 为最小, 然后根据 p 将原始区间一分为二, 划分成 U_1 和 U_2 子区间, 再对 U_1 和 U_2 分别寻找最佳划分点 p_1 与 p_2 . 比较 $E(A, U_1, p_1)$ 和 $E(A, U_2, p_2)$, 若前者较大, 则表示 U_1 划分较差, 有待进一步划分 U_1 , 否则划分 U_2 . 递归地使用上述方法找出其他 $k-2$ 个划分点, 这样总共就有了 k 个区间.

假设年龄字段经过上述某种方法分析后, 划分子区间为: $[20, 24)$: 小; $[25, 29)$: 中; $[30, 35]$: 大. 于是根据 RS 理论可有如下确定规则与可能规则:

确定规则:

- (1) 年龄小的学生成绩好;
- (2) 来自江苏的学生成绩好;
- (3) 年龄中等的学生成绩一般;
- (4) 来自贵州的学生成绩一般;
- (5) B 1 系的来自新疆的年龄大的学生成绩差;
- (6) 来自云南的年龄大的学生成绩差;
- (7) 来自上海的学生成绩一般;
- (8) 来自北京的学生成绩好;
- (9) 来自浙江的 A 2 系学生成绩一般.

可能规则:

- (10) A 2 系的来自新疆的学生成绩一般;
- (12) A 2 系的来自新疆的学生成绩差.

注: (1)~ (9) 的确定规则显然也是可能规则.

另外, 由上也可导出否定规则: A 2 系的来自新疆的学生成绩不好.

以上各规则的准确度 (ac) 及覆盖率 (cv) 可分别计算如下:

$$ac(1) = 1; ac(2) = 1; ac(3) = 1; ac(4) = 1; ac(5) = 1; ac(6) = 1;$$

$$ac(7) = 1; ac(8) = 1; ac(9) = 1; ac(10) = 0.5; ac(11) = 0.5;$$

$$cv(1) = 1; cv(2) = 0.6; cv(3) = 0.71; cv(4) = 0.14; cv(5) = 0.33; cv(6) = 0.33;$$

$$cv(7) = 0.29; cv(8) = 0.2; cv(9) = 0.14; cv(10) = 0.14; cv(11) = 0.33.$$

5 基于知识的查询策略(Knowledge-based query strategy)

在分布式知识系统中,从语法意义上讲,一个查询实际上是建立在 $V = \cup \{V_i | i \in I\}$ 之上的.因此除了本地查询之外还会出现全局查询(Global Query),这里把不仅仅使用了 V_i 的、对于 i 节点是不可达的(Locally Unreachable)查询称作全局查询.

IQS 在基于知识的智能查询系统中协调着用户与知识系统间的关系.在具体介绍查询算法之前,有必要对知识库的存放结构再作一说明.逻辑上,它是一个层次结构,即:如果本地知识库不能解决客户端所提交的查询,则将它递交给 $G_i \in I/L^-$ 中其它节点存放在本地的相邻知识库来解决,只有在 G_i 组内仍无法查询到所需信息时,才将查询请求通过网络传送到组内的协调节点上来,每个组内均设有一协调节点,它们之间则是通过网络协议完全互联的,可以相互交换各组内的知识.这样做所带来的好处是大大地减轻了网络通信负担.能够在本地信息系统或知识系统中解决的查询不必推到网上.另外需要说明的是:对于一个用户所提交的查询可以有两种类型,一种是选择满足条件的对象集合,另一种是对一些属性作出某种预测或判断.为了使得查询能适用于第一种情况,知识库还需作一点补充:对于每一条规则均记录信息系统中的哪些对象是能够得到满足的.

基于知识的智能查询算法可描述如下:

Step 1: 根据查询中所用到的属性确定其查询类型.若是全局查询则转 Step 3.

Step 2: 执行本地服务端应用程序,搜索本地知识库 LK_i 中的规则.

2.1: 若有合适的规则,则给出查询结果及准确度、覆盖率因子,并合并有关结果.转 Step 5.

2.2: 若无可用规则,对于本地 IS_i 进行查询操作,并合并有关结果.转 Step 5.

Step 3: 判定查询中是否有属性是本地可达节点 k ,若是则重复执行:

3.1: 运行本地服务程序,查找交互知识库 IK_i 中的有关规则.

3.2: 若 $\forall (k, i) \in L^-$ 且规则的前件属性包含于 $T_k \cap T_i$, 后件项 $d \in V_k \cup V_i$, 则由查询结果取代未知的后件项 d , 并给出此代换的准确度及覆盖率因子,其中 T_k, T_i 为查询的项.若不同的相邻节点所得

到的查询结果不一致,则还要解决冲突问题.转 Step 2 完成其它本地属性的查询.

3.3: 若 IK_i 中找不出合适的规则,则将有关的查询请求通过网络发送给近邻节点 k 的 IS_k .

3.4: 启动 IS_k 端的查询服务应用,并将查询结果发回给节点 i .

3.5: 若 i 收到了从不同节点返回的内容相左结论,则还需进行冲突分析.

Step 4: 判定查询中是否仍有本地不可达节点的属性,若没有则转 Step 2.

4.1: 提交有关查询请求给组内协调节点.

4.2: 由协调节点与其它协调节点通信,其它协调节点在其组内请求有关节点执行并返回结果.

4.3: 转 Step 2.

Step 5: 结束.

另外,由于在查询中引进了 RS 理论,所以在查询中具有很强的柔性,具体表现在以下两方面:

可有上、下近似集供选择.

① DECLARE

```
x_record, P . X ISP.  $\prod(x)$ @type; /*  $\prod(x) = [x]_p$ 
*/
CURSOR x_cursor IS select  $\prod(x)$  IN X; /* 下近似集
*/
```

BEGIN

```
P . X :=  $\Phi$ ;
FOR x_record IN x_cursor LOOP
P . X := P . X UNION x_record;
END LOOP;
```

END; 若是上近似集则将上述 select 语句改成如下:

```
select  $\prod(x)$  from ISP where ( $\prod(x)$  INTERSECTION X
< >  $\Phi$ );
```

② 查询条件式中可以引进粗糙等价以及粗糙包含等关系.

1. 粗糙等价

- (1) $P . X = P . Y$ 查询条件 X 与 Y 下等价;
- (2) $P^+ X = P^+ Y$ 查询条件 X 与 Y 上等价;
- (3) $P . X = P . Y$ 且 $P^+ X = P^+ Y$ 查询条件 X 与 Y 上、下等价.

2. 粗糙包含

- (1) $P . X \subseteq P . Y$ 或 $P . X \supseteq P . Y$ 查询条件 X 与 Y 下包含;
- (2) $P^+ X \subseteq P^+ Y$ 或 $P^+ X \supseteq P^+ Y$ 查询条件 X 与 Y 上包含;
- (3) ($P . X \subseteq P . Y$ 且 $P^+ X \subseteq P^+ Y$) 或 ($P . X \supseteq P . Y$ 且 $P^+ X \supseteq P^+ Y$) 查询条件 X 与 Y 上、下包含.

下面以表 1 中的学生情况为例来说明上述分布式智能查询策略.假定表 1 存放在节点 1, 另有学生家庭经济状况(表 2)存放在节点 2.

表2 学生家庭经济状况表举例

Table 2 An illustration of student financial condition

学生	年龄	生源	经济	学生	年龄	生源	经济
4乙	32	贵州	差	30丑	30	新疆	一般
6丙	20	新疆	差	35寅	26	上海	好
8丁	20	江苏	好	38卯	26	上海	好
9戊	21	浙江	好	44辰	20	北京	好
10己	21	江苏	好	48巳	21	北京	好
12庚	26	浙江	好	49午	21	江苏	好
14辛	20	江苏	好	52未	26	浙江	一般
17壬	20	江苏	好	56戌	32	新疆	差
18癸	21	江苏	好	59亥	26	云南	一般

如果在节点1需要查询是否存在成绩优秀而且家庭经济状况又好的学生,首先可将这一查询转化成形式化语言:“好_{成绩}*好_{经济}”。若节点1与2相邻,则节点1中存放了来自节点2的交互知识,其中我们发现确定规则:“来自江苏的学生家庭经济状况好”,而在 LK_1 中知道“来自江苏的学生成绩好”,于是在不查询 IS_2 的情况下即可作出肯定的判断。同样若有对象查询:查找所有成绩差经济好的学生,我们发现 LK_1 中的规则(5)所涉及的“1甲”同学为所要查询的结果。

6 结论(Conclusion)

本文讨论了一种分布式环境下的基于粗糙集合的智能查询策略,给出了分布式知识系统的结构框图及其查询实现算法,介绍了相似关系与概念树的RS知识库生成方法。它强调了知识在查询中所起的重要作用:减少查询时的网络通信流量,并能有效地提高查询效率,不必是传统的精确式查询,而可以是一种粗糙的查询方式。另外,关于系统知识库的维护这一重要问题将有待进一步研究。

参考文献(References)

- 1 王以和,涂小平. 分布式数据库系统. 电子工业出版社, 1988
- 2 Z W Ras. Collaboration Control in Distributed Knowledge-Based Systems. Information Sciences Journal, 1997, 96(3/4): 193~ 205
- 3 Z W Ras. Handling Queries in Incomplete CKBS through Knowledge Discovery. In: Proceedings of Rough Sets and Current Trends in Computing'98, Springer Verlag, 1998, 194~ 201
- 4 Z Pawlak. Rough Sets-theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, 1991
- 5 J W Grzymala-busse. Managing Uncertainty in Expert Systems. Kluwer Academic Publishers, 1991
- 6 A H rn. ROSETTA Technical Reference Manual: [Research Report]. Knowledge Systems Group, Norwegian University of Science and Technology, Trondheim, Norway, 1998
- 7 Z F Ma, H C Xing, X M Zheng. A Multiple Rough Set Approach for Information Retrieval. Journal of Southeast University (English Edition), 1999, 15(1): 61~ 66
- 8 马志锋, 邢汉承, 郑晓妹. 决策表中规则获取的不确定性研究. 控制与决策, 2000, 15(6)
- 9 马志锋, 邢汉承, 郑晓妹. 不完整 Vague 决策表中的近似集学习方法. 计算机研究与发展, 2000, 37(9): 1050~ 1057
- 10 X Hu, N Cercone. Learning in Relational Databases: a Rough Set Approach. Computational Intelligence, 1995, 11(2): 323~ 338
- 11 马志锋, 邢汉承, 郑晓妹. 基于不分明与相似关系的 Rough 集的超图描述. 计算机科学, 1999, 26(9): 35~ 39

作者简介

马志锋(1970-),男,博士生.研究领域为知识系统,数据库中知识挖掘,粗糙集合.

邢汉承(1938-),男,博士生导师.研究领域为人工智能与模式识别,图象处理, GIS.

郑晓妹(1973-),女,硕士生.研究领域为数据库系统,操作系统,人工智能.