基于粗糙集的分布式数据库查询

Query for Distributed Database based on Rough Sets

李同英　（江苏淮阴师范学院 223001）

摘要：本文在分布式数据库技术和模糊集理论的基础上，应用作为数据挖掘的新方法的粗糙集理论，针对分布式数据库的复杂查询处理问题，构建一种基于粗糙集的分布式数据库系统，设计出粗糙集上的查询算法，不仅极大地降低了信息查询的时间复杂性，而且提高了信息查准率，又兼顾了查全率，从而可以更准确迅速地查到所需要的信息，为信息查询和科技查询提供决策支持。

关键词：分布式数据库 粗糙集 查询

1　分布式数据库的体系结构

分布式数据库技术是分布性与集中性的统一。分布性表现在网络中是跨结点物理存储的，集中性表现在用户逻辑上所见是一个简单的、同构的数据库。相比之下，集中式的数据库管理系统需要物理上和逻辑上的双重集中。

分布式数据库（Distributed Data Base，DDB）可以定义为物理上分布而逻辑上集中的共享数据的集合。分布式数据库管理系统是管理分布式数据库的软件，通过分布式数据库管理系统可以使分布式数据库的分布特性对用户透明。图1为分布式数据库的体系结构图：基于以上分布性和逻辑协调性的分布式数据库，是虚拟、逻辑的，即是由许多 LDB 逻辑组织而成的，它是针对于全体用户的，全局的数据库。

2　分布式查询处理技术

分布式查询处理技术是分布式数据库的关键技术之一，随着分布式数据库技术的不断发展、成熟及其在信息服务机构中的应用，分布式数据库查询日渐成为信息查询的一个重要需求。高查询速度和较高的查全率前提下的高查准率一直是信息查询的主题。在分布式数据库信息查询中，对每条信息抽取若干个描述标引词，用这些标引词的集合来代表原信息，近似表示原信息的语义，从而实现按原信息的语义内容特征进行查询。分布式数据库的信息数量虽然在急剧增加，但是总的信息标引词数量的增加却很缓慢。

粗糙集理论是由波兰的 Z. Pawlak 教授在 1982 年首先提出的，它是研究模糊性和不确定性的一种新的数学工具，是处理模糊空间的一种数学方法。它根据由属性派生的等价关系进行分类。粗糙集的主要特点是：具有严格的数学定义和较强的鲁棒性（robustness），并且粗糙集信息处理不需要附加任何先决条件。

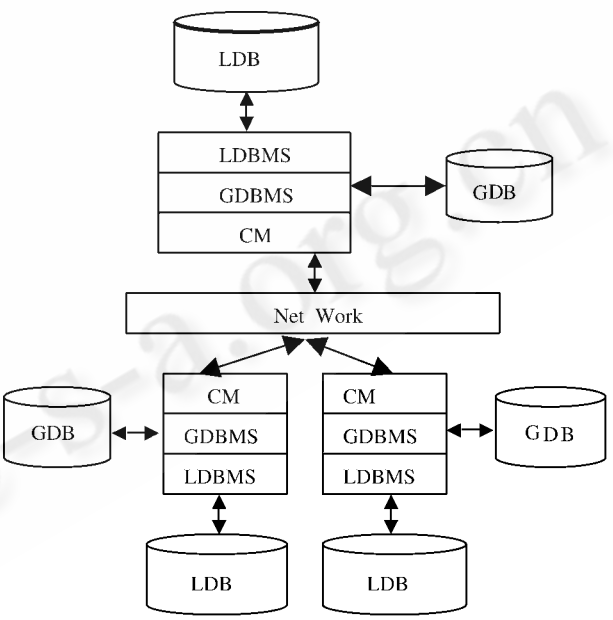

图 1　分布式数据库的体系结构

在不确定信息处理中，粗糙集和模糊集有类似之处，但是它们各自的着眼点不同。对于信息查询系统，粗糙集理论强调的是信息对象的不可分辨性，而模糊集理论强调信息对象的模糊性。两种方法相互补充，不能简单取代。

对于一个信息查询系统 $K=(D,P,A,f,R_a)$，设 R 是集合 D 上由属性集 A 确定的等价关系（A　P），R 将 D 划分为

$\{A_1, A_2, \cdots, A_k\}$，X 是 D 上的一个集合，由 R 表达的 X 的上近似空间 R－(X) 和下近似空间 R－(X) 都是 DIR 的模糊集合，R－(X) 和 R－(X) 称为粗糙模糊集。x 为自变量，则其隶属函数为：

$\mu_X^R(x) = \text{card}(X \cap R(x)) / \text{card}(R(x))$

当 $x \in R\_(X)$，$\mu_X^R(x) = 1$；

当 $x \in NEG_R(X)$，$\mu_X^R(x) = 0$；

当 $x \in BN_R(X)$，$0 < \mu_X^R(x) < 1$；

信息查询系统存储的数据记录即信息集合，每条记录的内容即是信息的表示。P 上的粗糙集 E 和查询分句集合 A 与标引词集合 P 之间的关系正是要探讨的，显然这是一种模糊关系，而对于某个粗糙集 Ek 和查询分句 Ak 与某标引词 p 而言，这种模糊关系反映了 Ak 和 Ek 与 p 之间的一种相关程度，这正是信息查询的实质。

3　粗糙集分布式数据库系统的生成

一个粗糙集分布式数据库系统的生成过程如图 2 所示。其中，“⇨”箭头表示粗糙模糊集信息系统生成时的数据流，也就是建造粗糙模糊集库的数据流。

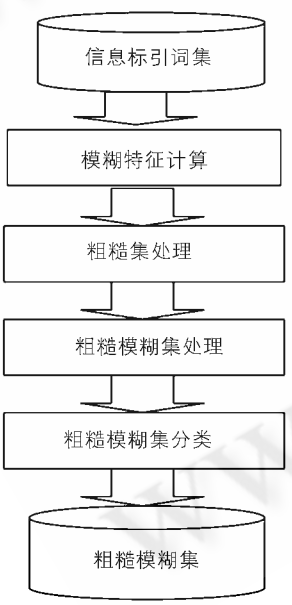

图 2　粗糙集分布式数据库系统的生成

模糊特征计算是将信息的标引词集合转换成模糊集合。假设标引词的设置是有序的（该顺序可由模糊统计得到），它们的模糊隶属度应该组成一个单调减序列。

设一条信息 D_k 可由 m 个描述标引词标引，即 $D_k = \{p_{k1}, p_{k2}, \ldots, p_{km}\}$，

如果我们取隶属度为：

$\mu_{D_k}(p_{k1}) = 1$，$\mu_{D_k}(p_{k2}) = (N-1)/N, \ldots, \mu_{D_k}(p_{km}) = 1/N$。

μ_{D_k} 显然是单调递减序列，那么利用这个算法可实现标引词模糊特征的自动化。

粗糙集处理是将标引词集合转换成粗糙集合。取每条信息的前 3 个标引词（一般情况下前 3 个标引词就可以近似代表该信息的语义内容特征）组成标引词集合 S，S　P，（P 为一个有限个标引词组成的属性集，$P = \{p_1, p_2, \cdots, p_m\}$ 表示分布式数据库中所有标引词的集合）。设 N = card(S)，这 3 个标引词称为该信息的主标引词。对 N 个元素做 3 个元素一组的组合，共有 N(N－1)(N－2)/6 种情况，形成 N(N－1)(N－2)/6 个集合。虽然信息的数量在急速增长，但是标引词的增长速度是缓慢的，所以 $N(N-1)(N-2)/6 << \text{card}(D)$，card(D) 为信息的条数，（D 为一个有限的信息对象集，也称论域，$D = \{D_1, D_2, \cdots D_k, \cdots, D_n\}$ 表示分布式数据库中经过标引的信息集合）。经过这样的数据整合后，粗糙模糊集信息系统就把对每条信息的海量查询转换成了对 N(N－1)(N－2)/6 个集合的粗糙模糊集查询，查询到隶属的粗糙模糊集后，再在所得到的粗糙模糊集查询得到所求信息。

粗糙模糊集处理是将粗糙集中的标引词和信息模糊化。标引词模糊化是指产生标引词隶属度。我们把在粗糙集处理中选取的主标引词的隶属度都设为 1，再找出所有由这 3 个主标引词代表的信息，把这些信息的标引词全部加入这个集合，它们的隶属度分别为：该标引词出现的次数/信息条数。

粗糙模糊集分类是按照字典序列将粗糙模糊集分类，建成树型层次结构，以便查找。最后把建成的粗糙模糊集取入粗糙模糊集系统。

模糊特征计算、粗糙集处理、粗糙模糊集处理和粗糙模糊集分类是粗糙模糊集信息系统的建造部分，经过上述数据综合归类后，粗糙模糊集分布式数据库系统的生成工作即告完成，为信息的模糊查询做好准备。

4　粗糙集信息系统的模糊查询过程

一个粗糙模糊集信息查询过程如图 3 所示。其中，“→”箭头表示信息查询时的数据流。

查询词预处理是将用户给出的查询关键词集进行模糊化。如果用户已经给出了每个关键词的隶属度，那就直接进

入下一步;否则,按关键词给出的先后次序给出递减的隶属度值。若用户给出的关键词不多于 3 个,那么每个关键词的隶属度都为 1;否则,前 3 个关键词的隶属度为 1,其余的 N 个关键词的隶属度以依次为:N/(N +1),(N -1)/(N +1),…,1/(N +1)。

粗糙模糊集上的模糊查询是按照下面 5.d 中粗糙模糊集的模糊查询法 RF - FQ,在粗糙模糊集信息系统上,求得最大的语义特征贴近度,找出最大的语义贴近度对应的粗糙模糊集合 E(可用模糊匹配法或根据三个主关键词),在粗糙模糊集 E 上进一步模糊匹配可得到用户所需的已经排好序信息 D。

上述 3、4 系统生成和查询实现算法具体如下。

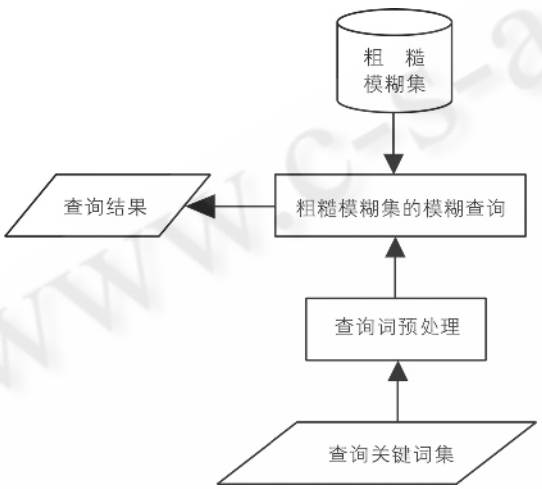

图 3　粗糙模糊集模糊查询过程

5　实现算法

5.1　粗糙集处理算法

输入:信息集合 D 及其标引词;

输出:信息集合 D 的以标引词为属性的粗糙集 π。

(1) 设信息条数为 N = card(D),j = 0,标引词集 S = ε,π = ε;

(2) 当 j≤N 时,取 D(+ +j)的前 3 个标引词存入 S;

(3) 设 M = card(S);

(4) for(i = 0;i < M;i + +)

for(j = i +1;j < M;j + +)

for(k = j +1;k < M;k + +)将{S(i),S(j),S(k)}组成的一个集合放入 π;

(5) 设 H = card(π);

(6) for(i = 0;i < H;i + +)

for(j = 0;j < N;j + +)

若 D(j)信息含有 π(i)中 3 个标引词,则将 D(j)存入 π(i)的信息域中;

(7) 输出粗糙集 π,结束。

粗糙集处理算法的时间复杂性主要花在第 6 句。因为 N > > M 并且 N > > H,其算法复杂性为 O(H × N)。

5.2　粗糙模糊集处理算法

输入:粗糙集 π;

输出:粗糙模糊集 E。

(1) 设 E = ε,粗糙集 π 的集合数 H = card(π);

(2) for(i = 0;i < H;i + +)

设 π(i)的标引词数 K = card(π(i)),数组 a[K] = 0,π(i)的信息条数 T = ‖π(i)‖;前 3 个标引词的隶属度为 1,a[0] = a[1] = a[2] = 1;

for(j = 3;j < K;j + +)

for(t = 0;t < T;t + +)

if 标引词 π(i)(j)出现在信息 π(i)[t]中 then a[j] + + ;

for(j = 3;j < K;j + +)a[j] = a[j]/T;

将集合 π(i)及其标引词的隶属度 a[K]存入 E;

(3) 输出 E,结束。粗糙模糊集处理算法的时间复杂性显然为 O(H × K × T)。由于 H > > K,H > > T,所以该算法的时间复杂性为 O(H)。

5.3　粗糙模糊集分类算法

输入:粗糙模糊集 E;

输出:树型粗糙模糊集 T。

(1) 建造一个空树 T;

(2) 建造以 T 为根的平衡二叉排序树;

(3) 存储树 T,结束。

树 T 的高度 h = O(log 2H),其中 H 是粗糙模糊集合数,这就保证了信息查询的时间复杂性是 O(log 2H)。

5.4　粗糙模糊集查询算法 RF - FQ

输入:模糊查询关键词集合 A 和粗糙模糊集 T;

输出:查询到的信息集合 D。

(1) 设 E = ε,信息粗糙模糊集的根 t = T;

(2) 取前 3 个查询词 K = First Three(A);

(3) 若 t = φ,没有查到 3 个查询词,调用 2 个查询词查询算法 TwoKey;若 TwoKey 成功,转 9;否则,没有查到,退出;

(4) 若 K→key = = First Three(t)→key,就是有 3 个关键词相同,转 7;

（5）若 K→key < First Three(t)→key，t = t→lchild，转 3；

（6）若 K→key > First Three(t)→key，t = t→rchild，转 3；

（7）求出查询到的信息的匹配模糊值；

（8）将查询到的信息按其模糊值进行递减排序，存入 D；

（9）输出 D，结束。

6 分布式数据库查询系统实现与结果分析

考虑到数据库的分布式特性和存储数据量的庞大，在创建分布式数据库查询系统时我们采用了 Oracle 数据库作为后台数据库。前台开发环境采用了 PowerBuild 软件工具。PowerBuild 软件工具提供了与 Oracle 数据库的专用接口，可以方便地实现与数据库的连接。图 4 给出了 Oracle 数据库作为后台数据库的分布式数据库查询系统结构图。

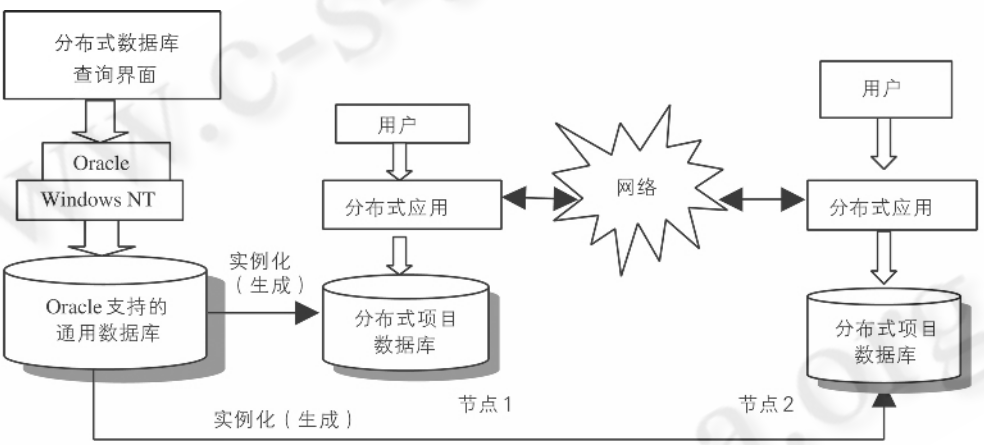

图 4　分布式数据库查询系统结构图

粗糙模糊集上的信息查询与传统的信息查询相比，时间复杂性有实质性的改进，从 O(log 2N) 降为 O(log 2H)。其中 N 是信息的数量，H 是粗糙模糊集上集合的数量，H << N。设有 1000 个主要标引词，将建成 1000 × 999 × 998/6 = 166167000 ≈ 1.7 亿个粗糙模糊集合。设有 10 亿条信息，有人要查询 50 条信息。按照粗糙模糊集上的查询算法只需要查找 (log 2 166167000) < 28 次，也就是说最多只需要 28 次查找。按照传统的信息查询算法，直接对信息本身查找，找一条信息就需要 (log 21000000000) ≈ 30 次查找，要查询 50 条信息就需要进行 30 × 50 = 1500 次查找。粗糙模糊集查询算法的时间复杂性为原来的 1/53。更重要的是粗糙模糊集上的查询算法的时间复杂性不会随着信息条数的增加而增加，它只随着主要标引词数量的变化而变化。

7 结语

随着分布式数据库技术的不断发展、成熟及其在信息服务机构中的应用，分布式数据库查询日渐成为信息查询的一个重要需求。高查询速度和较高的查全率前提下的高查准率一直是信息查询的主题。现代科技的飞速发展一方面生产出大量的信息资源，另一方面也给科技工作者在海量的信息库中发掘和使用资源上带来困难。分布式数据库技术既为我们提供了容纳大量信息的场所，又为我们对信息资源进行适时分析和深层挖掘提供支持；运用分布式数据库技术，使我们能够从大量繁杂的数据记录中发现有价值的信息和知识。可以预言，随着分布式数据库技术的不断发展、成熟及其在信息服务机构中的应用，必将使信息服务机构的信息服务能力、决策能力和信息服务机构信息整体应用效能得到进一步的改善和提高，同时也将为信息机构数字化建设和发展奠定基础。

参考文献

1 Pawlak Z. Rough Sets – Theoretical Aspects of Reasoning about Data. Kluwer：Academic Publishers，1992。

2 吉根林、杨明、赵斌、孙志挥，基于 DDMINER 分布式数据库系统中频繁项目集的更新，计算机学报，2003(10)[J]。

3 阳国贵、满家巨编，Oracle 数据库管理与使用教程，国防科技大学出版社，1998[M]。

4 袁松，PowerBuilder8.0 高级应用与开发，中国水利水电出版社，2002.3[M]。