集合特征码及相关算法研究

王树西
(对外经济贸易大学信息学院 北京 100029)

摘 要 在传统的集合运算过程中,集合内的元素一般通过自然语言表示,而不是形式化表示,这就在计算机处理时影响了集合运算的效率。为了解决这个问题,将二进制引入到集合运算过程中,提出集合特征码的概念,并定义了一系列的集合特征码运算规则,从而形成了一个较为完备的、形式化的集合特征码运算体系。在上述集合特征码的理论体系下,提出一系列的相关算法,从而在算法的层次上实现了集合特征码的理论体系。实验结果表明,通过集合特征码理论及相关算法,集合运算可以通过0、1运算快速实现,并且其成功实现了数据库中的查询操作。

关键词 集合特征码,集合运算,二进制,算法,数据库,查询
中图法分类号 TP392 **文献标识码** A

Research of "Set Attribute Code" and Related Algorithms

WANG Shu-xi
(Information Academy of the University of International Business and Economics, Beijing 100029, China)

Abstract In the traditional process of set operations, the elements of set are usually represented through natural language rather than formalized representation, and current representation has affected the efficiency of set operations. In order to solve this problem and improve the efficiency of set operations, this paper innovatively introduced the binary into the process of set operations, proposed the concept of "Set Attribute Code", and defined a series of operation rules about "Set Attribute Code". All the above formed a relatively complete and formal computing system of "Set Attribute Code". Based on above theory of "Set Attribute Code", this paper proposed a series of related algorithms, which demonstrated the correctness of the theory of "Set Attribute Code". Experimental results show that the process of set operations can be achieved through 0,1 operation in the theory of "Set Attribute Code" and related algorithms. And the target of query operation was successfully achieved, which is very important in the database through the mechanism of "Set Attribute Code".

Keywords Set attribute code, Set operations, Binary, Algorithm, Database, Query

1 引言

集合中的元素一般通过自然语言表示,而不是形式化表示。例如集合 $S=${王同,李军,曹秀兰,杜小刚,陈红,周小虎,张海波,吴天人,赵秉宏,王海,刘松涛}中,所有的元素(如"王同"、"李军"等)都是用自然语言(汉语)表示的,而不是形式化表示的。

在计算机进行集合运算的过程中(如集合并运算∪、集合交运算∩、集合补运算~、集合对称差运算⊕等),必须将集合中的元素形式化,这就影响了集合运算的效率。

为了解决这个问题,提高集合运算的效率,本文将二进制引进到集合运算过程中,创新性地提出了集合特征码的概念,并定义了一系列的集合特征码运算规则,从而形成了一个较为完备的、形式化的集合特征码运算体系。在上述集合特征码的理论体系下,提出一系列相关算法,从而在算法的层次上实现了集合特征码的理论体系。

本文首先介绍集合特征码的国内外研究现状,然后介绍集合的基本概念、基本运算,并指出集合运算需要改进的地方。在此基础上,将二进制引入到集合运算过程中,提出集合特征码的基本概念,并定义了一系列的集合特征码运算规则,从而形成了一个较为完备的、形式化的集合特征码运算体系。在上述集合特征码的理论体系下,提出一系列相关算法,从而在算法的层次上实现了集合特征码的理论体系,并对集合特征码理论进行了实验。实验结果表明,通过集合特征码,集合运算可以在计算机中通过0、1运算实现,并且可以成功实现关系数据库的查询操作。

2 集合特征码的国内外相关研究

国内外对集合相关问题的研究有很多,目前的研究倾向于理论研究,应用研究相对较少。

文献[1]研究了 P-集合与 F-记忆信息特性的应用;文献[2]研究了 $P(\rho,\sigma)$-集合与它的随机特性;文献[3]研究了基于集值决策属性的集值信息系统;文献[4]研究了函数 s-粗集、函数粗集与信息系统规律拆分-合成;文献[5]研究了特性

到稿日期:2011-08-30 返修日期:2011-11-22 本文受"对外经济贸易大学学术创新团队"和"对外经济贸易大学'211工程'三期建设"资助。
王树西(1976—),男,博士,主要研究方向为商务智能等,E-mail:wangshuxi2006@sina.com。

关系粗糙集下属性值粗化、细化时近似集增量更新方法；文献[6]研究了集值信息系统的知识约简与属性特征；文献[7]研究了 P-集合及其应用特征；文献[8]研究了关于主范式的下标集合及其应用；文献[9]对集合多覆盖问题的乘性权重更新分析进行了研究；文献[10]研究了测试集问题的集合覆盖贪心算法的深入近似；文献[11]研究了集合覆盖问题的启发函数算法；文献[12]对集合覆盖问题的数据约简进行了研究；文献[13]提出了一种最优特征集的选择算法。

在上述工作的基础上，本文提出集合特征码的理论体系；并在上述集合特征码的理论体系下，提出一系列相关算法，从而在算法的层次上实现了集合特征码的理论体系。

3 集合的基本概念和基本运算

集合的应用范围非常广泛。首先给出集合的基本概念和基本运算，然后将集合的概念应用到二维表中。

3.1 集合的基本概念

(1)集合：具有某些共性的或根据需要而规定的、个别的、不同的事物组成的群体。

(2)元素：集合中的个别事物。

(3)元素与集合的关系：属于($\in$)、不属于($\notin$)。

(4)二元有序组：若 A、B 是集合，$\forall a\in A$，$\forall b\in B$，则称 $\langle a,b\rangle$ 为二元有序组，其中 a 为第一元素，b 为第二元素。

(5)笛卡尔积：若 A、B 是集合，$C=\{\langle a,b\rangle \mid a\in A,b\in B\}$ 为 A 与 B 的笛卡尔积，记作 $C=A\times B$，符号“$\times$”表示笛卡尔积运算。

(6)n 阶笛卡尔积：若 $A_1,A_2,A_3,\cdots,A_n$ 是 n 个集合，称集合 $C=\{\langle a_1,a_2,a_3,\cdots,a_n\rangle \mid a_1\in A_1,a_2\in A_2,a_3\in A_3,\cdots,a_n\in A_n\}$ 为 $A_1,A_2,A_3,\cdots,A_n$ 的 n 阶笛卡尔积，记作 $C=A_1\times A_2\times\cdots\times A_n$。

3.2 集合的基本运算

集合的运算有很多，现在列举出其中一些基本运算，其中 A、B、C 是集合，E 是全集。

(1)并运算：称 $C=\{c\mid c\in A$ 或 $c\in B\}$ 是 A 与 B 的并集，记作 $C=A\cup B$。

(2)交运算：称 $C=\{c\mid c\in A$ 且 $c\in B\}$ 是 A 与 B 的交集，记作 $C=A\cap B$。

(3)补运算：若 A 是全集 E 的子集，称 $C=\{x\mid x\in E$ 且 $x\notin A\}$ 为 A 的补集，记作 $C=\sim A$。

(4)差运算：称 $C=\{c\mid c\in A$ 且 $c\notin B\}$ 为 A 与 B 的差集，记作 $C=A-B$。

(5)对称差运算：称 $C=(A-B)\cup(B-A)$ 为 A 与 B 的对称差集，记作 $C=A\oplus B$。

3.3 二维表中的集合概念

二维表是数据库系统中的操作对象。其中每一列可以看作是一个集合，列标题就是该集合的名称；每一行是一个若干数据的序列，称为一条记录。整个二维表是记录的集合。下面分析一张二维表(见表 1)。

表 1

学号	姓名	性别	年龄	籍贯
1001	王同	男	22	江苏
1002	李芳	女	21	陕西
…	…	…	…	…

表 1 中，第一列是“学号”集合，第二列是“姓名”集合，第三列是“性别”集合，第四列是“年龄”集合，第五列是“籍贯”集合。

每一列的标题是列数据的集合名，称其为字段名；每一行是一个 5 元有序组，称其为一条记录。

表 1 可以看作是“学号”、“姓名”、“性别”、“年龄”、“籍贯”这 5 个集合上的一个 5 元关系，也就是一个 5 阶笛卡尔积：学号×姓名×性别×年龄×籍贯。

表 1 是一个二维表，其中第一列和第二列给出从“学号”集合到“姓名”集合的映射，第一列和第三列给出从“学号”集合到“性别”集合的映射。

4 二维表中的集合运算举例

根据集合的基本概念和基本运算，可以在二维表中进行集合运算。由二维表组成的数据库称为关系型数据库。如果能够对二维表进行运算，那么就可以对关系型数据库进行运算。下面是一个二维表中集合运算的具体例子，见表 2。

表 2

姓名	年龄	文化程度	工作表现
王同	45	硕士	一般
李　军	23	大本	一般
曹秀兰	40	高中	良好
杜小刚	22	大本	优秀
陈　红	24	高中	良好
周小虎	31	大本	一般
张海波	36	硕士	一般
吴天人	47	大本	良好
赵秉宏	21	高中	优秀
王　海	32	博士	良好
刘松涛	33	大本	一般

4.1 二维表对应的集合

全集：$E=$｛王同，李军，曹秀兰，杜小刚，陈红，周小虎，张海波，吴天人，赵秉宏，王海，刘松涛｝

子集：

$A=\{x\mid x\in E$ 且 x 的年龄$\geqslant 40\}=$｛王同，曹秀兰，吴天人｝

$B=\{x\mid x\in E$ 且 $40>x$ 的年龄$\geqslant 30\}=$｛周小虎，张海波，王海，刘松涛｝

$C=\{x\mid x\in E$ 且 $30>x$ 的年龄｝$=$｛李军，杜小刚，陈红，赵秉宏｝

$I=\{x\mid x\in E$ 且 x 的文化程度高于大本｝$=$｛王同，张海波，王海｝

$J=\{x\mid x\in E$ 且 x 的文化程度为大本｝$=$｛李军，杜小刚，周小虎，吴天人，刘松涛｝

$K=\{x\mid x\in E$ 且 x 的文化程度不及大本｝$=$｛曹秀兰，陈红，赵秉宏｝

$P=\{x\mid x\in E$ 且 x 的表现一般｝$=$｛王同，李军，周小虎，张海波，刘松涛｝

$Q=\{x\mid x\in E$ 且 x 的表现良好｝$=$｛曹秀兰，陈红，吴天人，王海｝

$R=\{x\mid x\in E$ 且 x 的表现优秀｝$=$｛杜小刚，赵秉宏｝

4.2 运算要求和运算过程

选出年龄不足 40 岁、工作表现良好或优秀，且具有不低于大本学历的人组成的集合。

显然，该集合可以通过以下运算得到：

$$X=(\sim A)\cap(Q\cup R)\cap(\sim K)$$
$$=(\sim A)\cap(\sim P)\cap(\sim K)$$
$$=(\sim(A\cup P))\cap(\sim K)$$
$$=\sim(A\cup P\cup K)$$
$$=E-(A\cup P\cup K)=\{杜小刚，王海\}$$

5 集合特征码及其运算规则

为了提高集合运算的效率，需要将集合中的元素形式化。我们创新性地将二进制引进到集合运算过程中。首先提出集合特征码的概念，然后定义了集合特征码运算规则，从而形成了一个较为完备的、形式化的集合特征码运算体系。

5.1 集合特征码的定义

假设集合 A 中有 n 个元素，将其中的元素进行排序，每个元素占一个位。

全集 A 的任何一个子集 X，与一个 0、1 组成的串（称为布尔向量）一一映射。

映射规则如下：如果全集 A 的第 i 个元素在子集 X 中出现，那么 X 所对应的 0、1 串中，左起第 i 位为 1，否则为 0。

与子集 X 对应的 0、1 串，称为子集 X 的特征码，表示为 $C(X)$。

根据上述集合特征码的定义和表 2 可以得到，全集 E 以及子集 A、B、C、I、J、K、P、Q、R 的集合特征码如下：

$C(E)=11111111111$
$C(A)=10100001000$
$C(B)=00000110011$
$C(C)=01011000100$
$C(I)=10000010010$
$C(J)=01010101001$
$C(K)=00101000100$
$C(P)=11000110001$
$C(Q)=00101001010$
$C(R)=00010000100$

5.2 集合特征码的运算规则定义及相关算法

确定集合代码的运算规则（$\cup'$、$\cap'$、$\sim'$），使集合特征码的运算规则与集合的运算规则（$\cup$、$\cap$、$\sim$）一一对应。假设 P、Q 是两个集合特征码。

（1）集合特征码并运算（$\cup'$）规则

如果 P、Q 的第 i 位都是 0，那么“并”运算结果的第 i 位为 0，否则为 1。运算表见表 3。

表 3

P_i	Q_i	$P_i\cup'Q_i$
0	0	0
0	1	1
1	0	1
1	1	1

（2）集合特征码并运算（$\cup'$）算法

下面是单位集合特征码之间并运算的算法：

```
Code codeUnion(Code P[i],Code Q[i])
{
  if(P[i]==0 && Q[i]==0)
    return 0;
  else if(P[i]==0 && Q[i]==1)
    return 1;
  else if(P[i]==1 && Q[i]==0)
    return 1;
  else if(P[i]==1 && Q[i]==1)
    return 1;
}
```

下面是两串集合特征码之间并运算的算法：

```
Code[] codeUnion(Code P[],Code Q[],int& codeLength)
{
  Code[] resultCode="";
  for(i=0;i<codeLength;i++)
  {
    resultCode[i]=codeUnion(P[i],Q[i]);
  }
  return resultCode;
}
```

（3）集合特征码交运算（$\cap'$）规则

如果 P、Q 的第 i 位都是 1，那么“交”运算结果的第 i 位为 1，否则为 0。运算表见表 4。

表 4

P_i	Q_i	$P_i\cap'Q_i$
0	0	0
0	1	0
1	0	0
1	1	1

（4）集合特征码交运算（$\cap'$）算法

下面是单位集合特征码之间交运算的算法：

```
Code codeIntersect(Code P[i],Code Q[i])
{
  if(P[i]==0 && Q[i]==0)
    return 0;
  else if(P[i]==0 && Q[i]==1)
    return 0;
  else if(P[i]==1 && Q[i]==0)
    return 0;
  else if(P[i]==1 && Q[i]==1)
    return 1;
}
```

下面是两串集合特征码之间交运算的算法：

```
Code[] codeIntersect(Code P[],Code Q[],int& codeLength)
{
    Code[] resultCode="";
    for(i=0;i<codeLength;i++)
    {
      resultCode[i]=codeIntersect(P[i],Q[i]);
    }
    return resultCode;
}
```

（5）集合特征码补运算（$\sim'$）规则

如果 P 的第 i 位为 0（或者 1），那么“补”运算结果的第 i 位为 1（或者 0）。运算表见表 5。

表 5

P_i	$\sim'P_i$
1	0
0	1

(6)集合特征码补运算(～′)算法

下面是单位集合特征码之间并运算的算法：

```
Code codeComplement(Code P[i])
{
  if(P[i]==0)
    return 1;
  else if(P[i]==1)
    return 0;
}
```

下面是一串集合特征码求补运算的算法

```
Code[] codeComplement(Code P[],int& codeLength)
  {
    Code[] resultCode="";
    for(i=0;i<codeLength;i++)
    {
      resultCode[i]= codeComplement(P[i]);
    }
    return resultCode;
  }
```

(7)求集合特征码的算法

```
Code[] codeSet(Set P,Set totalSet,int setPLength)
{
  Code[] resultCode="";
  int setPLength=P. length();
  for(i=0;i<setPLength;i++)
  {
    if(P[i] appear in totalSet)
      resultCode[i]=1;
    else if(P[i] not appear in totalSet)
      resultCode[i]=0;
    }
    return resultCode;
}
```

(9)求并集特征码的算法

```
Code[] codeUnion(set P,set Q,Set totalSet,int& codeLength)
  {
    int setPLength=0;
    int setQLength=0;
    int codeLength=0;
    Code[] codeP=codeSet(P,totalSet;setPLength);
    Code[] codeQ=codeSet(Q,totalSet;setQLength);
    if(setPLength! =setQLength)
      return null;
    else if(setPLength==setQLength)
    {
      codeLength=setPLength;
      return codeUnion(codeP,codeQ,setLength);
    }
  }
```

(10)求交集特征码的运算规则

C(P∩Q)= C(P)∩′C(Q)

(11)求交集特征码的算法

```
Code[] codeIntersect(set P,set Q,Set totalSet,int& codeLength)
  {
    int setPLength=0;
    int setQLength=0;
    int codeLength=0;
    Code[] codeP=codeSet(P,totalSet;setPLength);
    Code[] codeQ=codeSet(Q,totalSet;setQLength);
    if(setPLength! =setQLength)
      return null;
    else if(setPLength==setQLength)
    {
      codeLength=setPLength;
      return codeIntersect(codeP,codeQ,setLength);
    }
  }
```

(12)求补集特征码的运算规则

C(～P)=～′C(P)

(13)求补集特征码的算法

```
Code[] codeComplement(set P,Set totalSet,int& codeLength)
  {
    int setPLength=0;
    int codeLength=0;
    Code[] codeP=codeSet(P,totalSet;setPLength);
    codeLength=setPLength;
    return codeComplement(codeP,codeLength);
  }
```

(14)根据特征码求集合的算法

```
String[] codeToSet(Code[] P,Set totalSet,int& setLength)
{
  String[] setResult="";
  setLength=0;
  codeLength=P. length();
  for(int i=0;i<codeLength;i++)
  {
    if(Code[i]==1)
    {
      setResult[setLength]=totalSet[i];
      setLength+=1;
    }
  }
  return setResult;
}
```

5.3 3种基本运算的实验

(1)并运算实验

$C(P)$=01011100101

$C(Q)$=10001101011

$C(P)\cup'C(Q)$=11011101111

(2)交运算实验

$C(P)$=01011100101

$C(Q)$=10001101011

$C(P)\cap'C(Q)$=00001100001

(3)补运算实验

$C(P)$=01011100101

$\sim'C(P)$=10100011010

5.4 集合特征码运算的综合实验

根据二维表(表2)以及这个二维表对应的全集、子集、运算要求,通过集合特征码的运算,得到符合要求的集合 X:

$$X=(\sim A)\cap(\sim P)\cap(\sim K)$$

$$C(X)=C(\sim A)\cap' C(\sim P)\cap' C(\sim K)$$

$$=\sim' C(A)\cap'\sim' C(P)\cap'\sim' C(K)$$

$$=01011110111\cap'00111001110\cap'11010111011$$

$$=00010000010$$

将该代码与全集人员姓名序列表进行对照,可以知道 X={杜小刚,王海}。

实验分析:此例说明,集合运算可以在计算机中通过0、1运算实现。通过集合运算,可以实现数据库中的查询操作。

结束语 本文创新性地将二进制引进到集合运算过程中,提出集合特征码的概念,并定义了一系列的集合特征码运算规则,从而形成了一个较为完备的、形式化的集合特征码运算体系。在集合特征码的运算体系下,提出一系列的算法来实现这个运算体系。实验结果表明,通过集合特征码,集合运算可以在计算机中通过0、1运算实现,可以成功地实现数据库中的查询操作。下一步的工作就是将这个集合特征码运算体系加以完善,并深化对集合特征码的应用研究。

参 考 文 献

[1] 汪洋,张冠宇,史开泉. P-集合与F-记忆信息特性-应用[J]. 计算机科学,2011,38(2):246-249,266

[2] 于秀清. $P(\rho,\sigma)$-集合与它的随机特性[J]. 计算机科学,2010,37(9):218-221

[3] 宋笑雪,张文修. 基于集值决策属性的集值信息系统[J]. 计算机工程与应用,2007,43(17):8-10

[4] 史开泉. 函数s-粗集,函数粗集与信息系统规律拆分-合成[J]. 计算机科学,2010,37(10):1-10

[5] 刘伟斌,李天瑞,邹维丽,等. 特性关系粗糙集下属性值粗化细化时近似集增量更新方法研究[J]. 计算机科学,2010,37(6):248-251

[6] 宋笑雪,李鸿儒,张文修. 集值信息系统的知识约简与属性特征[J]. 计算机工程,2006,32(22):26-27

[7] 史开泉. P-集合与它的应用特征[J]. 计算机科学,2010,37(8):1-8

[8] 亓正坤,王廷明. 关于主范式的下标集合及其应用[J]. 青岛理工大学学报,2009,30(4):205-208

[9] 崔鹏,钱丽艳. 集合多覆盖问题的乘性权重更新分析[J]. 计算机科学,2007,34(10):219-220

[10] 崔鹏,刘红静. 测试集问题的集合覆盖贪心算法的深入近似[J]. 软件学报,2006,17(7):1494-1500

[11] 权光日,洪炳熔,叶风,等. 集合覆盖问题的启发函数算法[J]. 软件学报,1998,9(2):156-160

[12] 陈韬,吴志勇,孙乐昌,等. 集合覆盖问题的数据约简研究[J]. 计算机应用研究,2010,27(9):3307-3311

[13] 朱明,王俊普. 一种最优特征集的选择算法[J]. 计算机研究与发展,1998,35(9):803-805

(上接第153页)

推理的扩展模糊时间Petri网(Backward Reasoning Extended Fuzzy Time Petri Net,BREFTN)模型,有效解决了上述种种问题。

结束语 本文在原有研究的基础上,结合数据起源的相关研究,针对EFTN特性进行了改进,提出了模糊时间Petri网逆向推理模型(BREFTN)。利用模型的算法,有效地解决了此模型的逆向推理以及模糊时间计算问题。但是本文模型在描述有环的Petri网存在薄弱环节,相关的模型以及算法尚待改进,这是未来工作的重点。

参 考 文 献

[1] Tan W C. Provenance in databases: Past, current, and future[J]. IEEE Data Eng. Bulletin, 2007, 30(4): 3-12

[2] Simmhan Y L, Plale B, Gannon D. A Survey of Data Provenance Techniques[R]. Bloomington, Bloomington IN 47405: Computer Science Department, Indiana University, 2005

[3] 高明,金澈清,王晓玲,等. 数据世系管理技术研究综述[J]. 计算机学报,2010,33(3):373-389

[4] 刘强,崔莉,陈海明. 物联网关键技术与应用[J]. 计算机科学,2010(6):1-4,10

[5] Zhou Y, Murata T. Modeling and Performance Using Extended Fuzzy-timing Petri Nets for Networked Virtual Environment [J]. IEEE Transactions on System, Man, and Cybernetics-part B: Cybernetics, 2000, 30(5): 737-755

[6] Zhou Y, Murata T. Modeling and Analysis of Distributed Multimedia Synchronization by Extended Fuzzy-timing Petri Nets[J]. Transaction of the SDPS, 2001, 15(4): 130-141

[7] Murata T. Temporal Uncertainty and Fuzzy-timing High-level Petri Nets[C]// Proc. of the 17th International Conference on Application and Theory of Petri Nets. New York, USA: Springer, 1996: 11-28

[8] Murata T, Suzuki T, Shatz S. Fuzzy-timing High-level Petri Nets (FTHNs) for Time-critical Systems[M]. Fuzziness in Petri Nets, Studies in Fuzziness and Soft Computing, Physica-Verlag, 1999: 88-114

[9] Zhou Y, Murata T. Petri Net Model with Fuzzy-timing and Fuzzy-metric Temporal Logic[J]. International Journal of Intelligent Systems, 1999, 14(8): 719-7467

[10] 张稳,张桂成. 改进的基于规则的逆向模糊推理算法[J]. 通信学报,2008,29(2):101-105

[11] Yang R, Heng P A, Leung K S. Backward Reasoning on Rule-based System Modeled by Fuzzy Petri Nets Through Backward Tree[J]. FSKD, 2002: 18-22

[12] Ram S, Liu J. A new perspective on semantics of data provenace [C]// Proc. First International Workshop on the Role of Semantic Web in Provenance Management. Washington D. C., 2009

[13] Chien E. Linux. Adore. Worm [EB/OL]. http://securityresponse. symantec. com/avcenter/venc/data/linux. adore. worm. html, 2008-08

[14] Hwang H U, Kim M S, Noh B N. Expert System Using Fuzzy Petri Nets in Computer Forensics[J]. Lecture Notes in Computer Science, 2007, 4413(2007): 312-322

[15] Russell S J, Norvig P. Artificial Intelligence: a modern approach (2nd edition)[M]. Prentice-Hall, 2003