列名与数值不确定情况下的模式匹配问题研究

黄冬梅 冯 恺 赵丹枫 郭颖新
(上海海洋大学信息学院 上海 201306)

摘 要 模式匹配是数据集成领域的一个重要研究内容,列名与数据值不确定是模式匹配中的一种常见情况,当前较普遍的方法是基于互信息及欧式空间距离。但该方法没有解决因属性相似度相同或相近而引起的错误匹配问题。针对该问题,提出了多重迭代筛选方法,首先确定两个关系模式中能一次性正确匹配的部分属性对,再从中求出最优属性对,然后给出基于条件互信息的匹配方法,利用最优属性对计算未匹配属性的条件互信息,进一步计算各属性之间的欧氏距离,最终得到匹配结果,从而解决了错误匹配问题。实验结果表明所提算法正确、有效。

关键词 不确定性,模式匹配,条件互信息
中图法分类号 TP391.7 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.08.018

Study on Schema Matching with Uncertain Column Names and Data Values

HUANG Dong-mei FENG Kai ZHAO Dan-feng GUO Ying-xin
(College of Information Technology, Shanghai Ocean University, Shanghai 201306, China)

Abstract Schema matching is an important research in the field of data integration. The uncertainty of column names and data values is a common situation. The common method at present dealing with schema matching problem is based on mutual information and Euclidean distance. But this method does not solve the mistaken matching problem caused by the identity or the high similarity of the attributes. To solve this problem, this paper proposed multiple iterative screening method, which firstly, in two relation models, fixes some of the corrects attribute pairs in one time and then selects the best optimized attribute pair. Secondly, this paper lodged the method based on conditional mutual information, which utilizes the best optimized attribute pair to calculate the conditional mutual information of un-matched attributes and further calculates the Euclidean distance between each attribute. Finally, the matching result was acquired. The wrong matching problem was solved. The experiment result indicates the given algorithm is correct and effective.

Keywords Uncertainty, Schema matching, Conditional mutual information

1 引言

数据集成是不确定数据产生的一大根源[1]。模式匹配是数据集成中至关重要的一步。在模式匹配过程中减少数据不确定性对数据集成有重要意义。

传统的基于语义的模式匹配需要消耗大量的人力和时间,而且容易出错[2]。因此,现代人们开始使用半自动的模式匹配工具来辅助完成模式匹配。但是,因为依赖于特定领域上下文的模式匹配无法确定其是否正确或者相似的标准很难准确描述,具有不确定性,所以该不确定性是模式匹配过程所固有的、不可避免的。

在不确定性模式匹配情况中,会遇到不透明模式的匹配问题,即实际表中的列名(属性名)或值并没有直接给出,而是用一些代码表示,这便给原先的语义匹配方法带来了困难。本文主要研究不确定模式匹配中属性值差异性较大的情况。例如,"Four Wheel Car"用"FWC"代表,而在另一个表中"Four Wheel Car"用"4WC"表示。原先表的属性通过语义匹配便无法得到匹配结果。此外,这种情况下增大了语义匹配的不准确性,所以语义匹配在属性增加时难度加大。

目前有关不透明列名与数值的模式匹配方法主要有3种:1)Kang J, Naughton J F[3]提出非解释型匹配算法:采用计算属性(或属性对)之间的互信息,建立依赖图,再计算欧氏距离得到解决途径。2)Jaiswal A, Miller D J, Mitra P[4]提出植入数值匹配算法:在建立依赖图之前,植入数值,构造匹配方程并引入新的距离公式寻找正确匹配。3)Rabinovich B, Last M.[5]提出空白列匹配算法:引入空白列简化对不完全匹配情况进行改进。上述三者均在不透明模式匹配中有贡献,但均未能显著提高匹配效率及准确率。

下面通过实例来说明不确定模式匹配问题。

例1 给定两张表:表1和表2,其记录的是一家汽车厂

到稿日期:2013-07-15 返修日期:2013-07-20 本文受国家自然科学基金资助项目(61272098),科技部973项目(2012CB316200),南北极环境综合考察与评估专项(CHINARE2012-04-07)资助。

黄冬梅(1963—),女,硕士,教授,主要研究方向为空间数据库技术、海洋GIS及辅助决策技术空间数据库技术、海洋GIS及辅助决策技术等,E-mail:dmhuang@shou.edu.cn;**冯 恺**(1989—),男,硕士,主要研究方向为数据集成、模式匹配;**赵丹枫**(1982—),女,博士,讲师,主要研究方向为业务流程管理、数据库理论、云计算等;**郭颖新**(1988—),男,硕士,主要研究方向为ZigBee、数据集成。

中的一些汽车属性。由于工作疏忽，表 2 列名与部分数值被油墨污染，无法辨认出其中内容。使用者只能凭借自身经验恢复对先前表中数据的印象，尤其是对其中数据的重复次数等相关信息。

表 1　汽车厂各型汽车属性表

A	B	C	D
A1	b2	c1	d1
A3	b4	c2	d2
A1	b1	c1	d2
A4	b3	c2	d3

表 2　汽车厂各型汽车属性表

W	X	Y	Z
W2	x1	y1	z2
W4	x2	y3	z3
W3	x3	y3	z1
W1	x2	y1	z2

假设这两张表是一对一映射。由于两张表中的语义都无法确定，因此基于语义的匹配工具就无法完成匹配。

然而，使用之前解决不透明模式匹配的方法会得出匹配重复的情况，即 B 与 W 和 B 与 Z。此时，无法确定属性 B 映射属性 W 还是映射属性 Z。此外，其他属性的匹配有时虽然距离不同，但其距离差距较小，无法有效映射，使匹配准确性降低。

本文主要贡献如下：

(1)提出多重迭代筛选方法，筛选出最优正确匹配对。该方法提高了不确定模式匹配的效率。

(2)在多重筛选匹配对的基础上，提出了条件互信息方法。该方法有效地解决了属性匹配不准确问题。

2　定义

定义 1(互信息[6])　是信息论里的一种信息度量，它用于描述两个事件集合之间的相关性程度。令 X 与 Y 表示同一关系表中的两个属性，则互信息公式[6]为：

$$MI(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)P(y)}$$

其描述的是属性 X 和 Y 之间的关系。$P(x,y)$ 表示一个元组中两个属性值之间的条件概率。$P(x)$、$P(y)$ 则表示各个属性之间的独立概率。

以例 1 中的表 1 为例，则属性 A 与 B 的相关性为：

$$\begin{aligned}MI(A;B)&=\sum_{a\in A}\sum_{b\in B}p(a,b)\log\frac{p(a,b)}{p(a)P(b)}\\&=p(a1,b2)\log\frac{p(a1,b2)}{p(a1)p(b2)}+p(a3,b4)\log\frac{p(a3,b4)}{p(a3)p(b4)}+p(a1,b1)\log\frac{p(a1,b1)}{p(a1)p(b1)}+p(a4,b3)\log\frac{p(a4,b3)}{p(a4)p(b3)}\\&=0.5*1*\log\frac{0.5*1}{0.5*0.25}+0.25*1*\log\frac{0.25*1}{0.25*0.25}+0.5*1*\log\frac{0.5*1}{0.5*0.25}+0.25*1*\log\frac{0.25*1}{0.25*0.25}\\&=1.5\end{aligned}$$

定义 2(欧几里得距离[7])　指在 n 维空间内，两个点在空间中的最短的线段长度。其通用距离公式为：

$$D_M^N(A,B)=\sum_{i,j}(1-\theta\frac{|a_{ij}-b_{m(i)m(j)}|}{a_{ij}+b_{m(i)m(j)}})$$[7]

该公式用于描述属性间的欧氏距离。其中，$D_M^N(A,B)$ 表示两属性间的欧几里得距离。θ 控制普通距离结果的参数是一个正常数。a_{ij}，b_{ij} 分别表示各自关系表中任意一属性与其同一个关系表中其他属性之间的互信息值。

假设两个属性 X、Y，其属性个数分别为 m、n，令 A_i、B_i 分别表示 X、Y 中的任一属性，则 A_i/B_i 与其他 $m-1/n-1$ 个属性存在 $m-1/n-1$ 个互信息值。那么，属性 A_i 和 B_i 的距离为：其中 $MI(A_i;A_j)$ 为属性 A_i 与 A_j 互信息。即：

$$MI(A_i;A_j)=\sum_{x\in X}\sum_{y\in Y}p(ai,aj)\log\frac{p(ai,aj)}{p(ai)P(aj)}$$

如果 $m=n$，则为完全匹配。该公式可化简为：

$$D_M^N(A,B)=\sqrt{\sum_{i,j}(a_{ij}-b_{m(i)m(j)})^2}$$

其中，m 为从 A 表中 a 属性映射到 B 表中 b 属性的标志(例如：m(A 表中属性)$=B$ 中匹配属性)。例如，例 2 中 B 与 X 属性之间的欧氏距离为：

$$\sqrt{(1.5-1.5)^2+(1.0-0.5)^2+(1.5-1.0)^2}=0.25$$

即 B 与 X 之间的距离为 0.25。

定义 3(条件互信息)　指在确定一个正确匹配属性基础上，对剩余两个属性互信息的计算。

$$MI(X;Y|Z)=\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z}P(x,y,z)\log\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)}$$

X，Y，Z 为一个关系表中的 3 个属性。其中 Z 为已经确认的正确匹配属性。例 1 中，正确匹配属性为 C 的情况下 A 与 B 的条件互信息计算过程为：

$$\begin{aligned}&P(a1,b2,c1)*\log\frac{p(c1)*p(a1,b2,c1)}{p(a1,c1)*p(b2,c1)}+\\&P(a3,b4,c2)*\log\frac{p(c2)*p(a3,b4,c2)}{p(a3,c2)*p(b4,c2)}+\\&P(a1,b1,c1)*\log\frac{p(c1)*p(a1,b1,c1)}{p(a1,c1)*p(b1,c1)}+\\&P(a4,b3,c2)*\log\frac{p(c2)*p(a4,b3,c2)}{p(a4,c2)*p(b3,c2)}\\&=0+0+0.25+0.25=0.5\end{aligned}$$

即 A 与 B 的条件互信息值为 0.5。

3　多重迭代筛选

为了选出正确匹配属性，这里用到多重迭代筛选方法[7]。对筛选出的结果进行分析，选出最优属性作为正确属性。此外，该方法也通过了实例化证明。

3.1　多重迭代筛选方法

多重迭代筛选方法运用定点计算原理对初始的相似度进行迭代运算，这一步通过迭代运算，最后得到各个候选匹配的稳定相似度值。

该方法的第 3 步计算(迭代计算)[8]公式如下：

$$\partial^{i+1}(x,y)=\partial^i(x,y)+\sum_{(a_u,p,x)\in A,(b_u,p,y)\in B}\partial^i(a_u,b_u)w((a_u,b_u),(x,y))+\sum_{(a_v,p,x)\in A,(b_v,p,y)\in B}\partial^i(a_v,b_v)w((a_v,b_v),(x,y))$$

其中，$a(X,Y)\geqslant 0$ 定义为一个在 $A\times B$ 上的总体函数，表示测量两个节点 $x\in A$，$y\in B$ 的相似度。定点计算方法就是基于定值的迭代运算。∂^i 表示 A 和 B 之间映射的第 i 步迭代。映射 ∂^0 表示 A 和 B 与中间节点间的初始相似度。

3.2　多重迭代筛选算法

多重迭代筛选算法是一种基于图形结构的模式匹配方

法。该算法的输入是要匹配的两个模式，在算法的执行过程中首先将这个模式转换成两个有向标记图，再依据数据库模式中的列名和数据类型进行定点计算，多步迭代之后，最后输出图中对应节点间的映射，得到匹配的结果。

3.2.1 算法描述

算法 1 RealMatch 算法(RM)

```
Input:Sl,S2;
Output:TotalResult()两个模式中元素的匹配关系。
RM(S1,S2)
Begin
Step 1  G1=SQL2Graph(S1);G2=SQL2Graph(S2)
        /*将模式 Sl 和 S2 转换成有向标记图 G1 和 G2*/
Step 2  initialMap=StringMatch(G1,G2)
        /*利用串匹配函数计算初始相似度 initialMap*/
Step 3  semiresult=SFJoin(G1,G2,initialMap)
        /*用 SFJion 函数得到 G1 和 G2 中匹配节点对*/
Step 4  midresult=SelectThreshold(semiresult)
        /*对得到的结果进行过滤，得到较精确的结果 t*/
Step 5  S1'=All(S1)-domain(midresult),S2'=A11(S2)-(All
        (midresult)—domain(midresult))
        /*构造 S1'和 S2'，即未匹配的元素*/
Step 6  N1=Count(S1');N2=Count(S2')
        /*计算 S1'和 S2'中元素个数*/
Step 7  /*进一步的匹配操作*/
        if(N1>1) and (N2>1) then
        /*判断是否要接着进行匹配操作 4*/
        G1'=Table2DepGraph(S1');G2'=Table22DepGraph(S2')
        /*将 S1'和 S2'转换为依赖关系图*/
        {G1'(a),G2'(b)}=GraphMatch(G1',G2')
        /*得到匹配节点对*/
        totalresult=midresult+{G1'(a),G2'(b)}
        /*合并候选匹配*/
            return(totalresult)
        else
            return(midresult)
End
```

算法分析：n 表示属性个数。这里的每一步计算用到 3 次循环筛选计算可能匹配的对。所以算法时间复杂度为 $O(n^3)$。

3.3 实例分析

表 3 列举出一部分例 1 中属性相似度情况。其中可以得出部分可确认匹配结果，即 A 与 Z，B 与 W。由于经过上述步骤有大量不合理匹配集合，因此，仍需要其他方法对剩余未成功匹配属性进行进一步处理。

表 3 经过定点计算的相似度值(部分)

表 1 属性	表 2 属性	相似度
A	W	0.14
B	X	0.24
C	Y	0.36
D	Z	0.13
A	X	0.52
A	Y	0.26
A	Z	1.0
B	W	1.0
B	Y	0.31
B	Z	0.24
C	W	0.17

3.4 最优属性对的选取

经过多重迭代筛选方法，可以得到部分正确匹配属性，为了进一步确定其他属性的对应关系，需要找到其中一对最优属性对，才能应用第 4 节的方法。

定理 1 给定两个维度为 n 的关系 R^1、R^2，(A_i^1, A_j^2) 为经过多重迭代筛选方法筛选后的精确匹配对，其中，A_i^1，A_j^2 分别为 R^1，R^2 的一个属性，$i,j\in[1,n]$，若用 $f(A_i)$ 表示 A_i 中不同属性值的个数，则 $f(A_i)$ 越大，A_i 所在的属性对越优。

证明：

最优评判标准：

设 A_i 与 B_i 为两不同值分布的属性。其中，$f(A_i)>f(B_i)$。X,Y 为待匹配属性。则两种情况下得到 X 与 Y 匹配的结果分别为：

$$MI(X;Y|A_i)=\sum_{x\in X}\sum_{y\in Y}\sum_{ai\in A_i}P(x,y,ai)\log\frac{p(ai)p(x,y,ai)}{p(x,ai)p(y,ai)}$$

$$MI(X;Y|B_i)=\sum_{x\in X}\sum_{y\in Y}\sum_{bi\in B_i}P(x,y,bi)\log\frac{p(bi)p(x,y,bi)}{p(x,bi)p(y,bi)}$$

由两个公式的计算结果可得知，A_i 作为最优匹配属性所得到的结果比 B_i 要小。说明不相同属性个数越多，匹配结果更准确。证毕。

以下通过例 2 说明定理 1。

例 2 表 4 中有两个属性，这两个属性分别作为正确匹配属性，则根据条件互信息公式的对比，可以得出：

$$MI(X;Y|Z)=\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z}P(x,y,z)\log\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)}$$

$$MI(X;Y|C)=\sum_{x\in X}\sum_{y\in Y}\sum_{c\in C}p(x,y,c)\log\frac{p(c)p(x,y,c)}{p(x,c)p(y,c)}$$

这两个公式分别把 Z 与 C 作为不同的正确属性。其中 Z 与 C 属性的数值分布不同，其数值分布如表 4 所列。

表 4

Z	C
Z1	C1
Z2	C2
Z4	C1
Z3	C2

由表 3 可以得知：

$$P(x,y,z)=P(x,y)*P(z) \tag{1}$$

$$P(x,y,c)=P(x,y)*P(c) \tag{2}$$

由式(1)、式(2)可得：

$$P(x,y)*(P(z)-P(c))$$

由式(1)、式(2)可知：这两个式子的大小取决于 $P(z)$ 和 $P(c)$。通过计算可以得出 $P(z)$ 比 $P(c)$ 要小。因为数值的离散程度不同，离散程度越大，P 的值越小。这样便可以推出式(1)比式(2)值小。

$$P(x,z)=P(x)*P(z) \tag{3}$$

$$P(x,c)=P(x)*P(c) \tag{4}$$

由式(3)、式(4)可知：式(3)要小于等于式(4)。

同理可知：

$$P(y,z)=P(y)*P(z) \tag{5}$$

$$P(y,c)=P(y)*P(c) \tag{6}$$

式(5)小于等于式(6)

这样，可以得出式(1)比式(2)值要小。也就是通过条件互信息方法得出的正确属性，在其属性值分布不同的情况下，选取值离散程度越大的属性可以减小两个属性之间的距离。这样可以使联系更紧密的正确匹配属性的相互关系凸显，有效提高区分度，从而提高属性间匹配的准确度。

4 基于条件互信息的匹配

条件互信息匹配方法是在多重迭代筛选方法得出正确匹配属性的基础上，对剩余未匹配属性进行匹配。该方法也通过实验得到验证。

4.1 条件互信息匹配算法

条件互信息匹配算法主要由两部分构成：

1. 在通过已有多重迭代筛选方法得到正确匹配对的基础上进行 CMI 算法计算。

2. 在 CMI 算法所得结果的基础上，使用 OD 算法，最终求得属性之间的距离，以得到正确的匹配结果。

4.1.1 CMI 算法描述

通过算法 2 对条件互信息值的求取过程进行描述。

算法 2 条件互信息计算过程(CMI)

Iuput：L[n]为一个动态数组，即(L_0，L_1，L_2，…，L_{n-1})，其中令 L_2 为固定数组列。

P[n]为各列中的数值，即(P_0，P_1，P_2，…，P_{n-1})；P_2 为固定列中的数值。

Output：C[n]为输出的条件互信息值，即(c_0，c_1，…，c_{n-1})

CMI(L[n],P[n])

Begin

Step 1 Input L_0，L_1，L_2，…，L_{n-1}

/ * 输入动态数组各列 * /

Step 2 $L_0=P_0$，$L_1=P_1$，$L_2=P_2$，…$L_{n-1}=P_{n-1}$

/ * 把各列中的值赋给每个列 * /

Step 3 getCMI()

{

$$CMI() = MI(X;Y \mid Z) = \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} P(x,y,z)\log\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)}$$

}

/ * 用条件互信息公式求解相互信息 * /

Step 4 C_0=getCMI()，C_2= getCMI()，…，C_{n-1}=getCMI()

/ * 各两列之间条件互信息值 * /

End

算法分析：n 是关系表中匹配属性个数。该算法的时间复杂度为 $O(n^3)$。其正确性已经通过实例得到验证。

在通过算法 2 求得各个两组属性条件互信息之后，需要求解各个属性结点之间的欧氏距离值。下面将通过算法 3 描述结点间欧氏距离的求解方法。

4.1.2 OD 算法描述

通过算法 3 对欧氏距离的求取过程进行描述。

算法 3 欧氏距离求解过程(OD)

Input：C[n]为条件互信息值，即(C_0，C_1，C_2，…，C_{n-1})

Output：D[n][n]为两点之间的欧式距离，即(D_{00}，D_{01}，D_{02}，…，$D_{(n-1)(n-1)}$)

OD(C[n])

Begin

Step 1 getDis()

{getDis()=$D_M^U(A,B)=\sqrt{\sum_{i,j}(a_{ij}-b_{m(i)m(j)})^2}$

}

/ * 欧几里得距离 * /

Step 2 D_{00}=getDis()，D_{01}=getDis()，…，D_{nn}=getDis()

/ * 求得距离输出 * /

End

算法复杂度：由于引入了 RealMatch 与条件互信息算法，在使用条件互信息之前需要通过之前的算法求得至少一对匹配属性。接着，在得到确定属性之后，用条件互信息算法对剩余的属性进行匹配。n 为关系表匹配属性个数。算法需循环执行 3 次，所以算法的时间复杂度为 $O(n^3)$。算法的正确性也已经通过实例得到证明。

4.2 实例分析

这里，Z 为确定的正确匹配属性，$P(x,y,z)$为三者的联合概率。$P(x,z)$及 $P(y,z)$是两者之间的联合概率。从公式可以知道，确定的正确属性对后面属性匹配的互信息取值有很大影响。由之前的多重迭代筛选方法得到的一对正确匹配属性为 C 和 Y。这样便把 C 与 Y 分别作为表 1 与表 2 的条件互信息。通过条件互信息所求得的每个表的 3 个属性的条件互信息如图 1 所示。

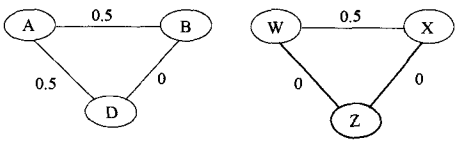

图 1 属性间的条件互信息值

通过欧几里得距离公式计算得到 A 到 Z 距离最小，B 到 W 距离最小，D 到 X 距离最小。这样便确定了 A 映射 Z，B 映射 W，D 映射 X。通过条件互信息算法，得到两属性距离与原先两属性之间的距离有明显不同的结果。如例 1 中，A 与 W，A 与 Z 两对属性之间的距离用以前的方法求得皆为 0.5，这样的结果无法区分最正确匹配结果。而用条件互信息方法求得 A 与 Z 的值为 0.25。这样便能找到准确的匹配对，从而提高匹配结果的准确度，很好地解决了之前匹配不准确的问题。

5 实验

5.1 实验设置

(1)实验环境

操作系统：Win 7 Professional

处理器：Intel Core 2

RAM：2.00GB

算法运行环境：Visual Studio 2010

实验语言：VB

(2)实验数据设置

本文的实验数据来自某海洋监测点的浮标采集数据，关系模式中属性个数为 2 至 20 个。使用条件互信息方法与文献[3]中的方法在匹配准确度上进行对比。实验分为一对一匹配、满射匹配与部分匹配两部分进行，并根据实验结果讨论属性个数对匹配结果精度的影响。

5.2 实验结果及分析

这里对两类匹配情况进行实验分析，一种情况为一对一匹配与满射匹配，另一种情况为部分匹配。一对一匹配与满

射匹配描述的是两张关系表中每个属性都可以在对应的关系表中寻找到应匹配属性。部分匹配为两张关系表中一部分属性无法匹配对应属性。

5.2.1 一对一匹配与满射匹配

每个统计图中横轴为属性个数，纵轴为匹配准确度(单位:%)。其中准确度用 P 表示，c 为通过条件互信息得到的成果匹配结果，n 为两匹配表之间实际正确匹配对数量，则准确匹配度公式为:$P=c/n$。

实验中，两张关系表中的属性一样，属性值随机产生。通过多重迭代筛选方法及条件互信息匹配方法得到匹配结果与原先匹配结果准确度情况，如图 2 所示。

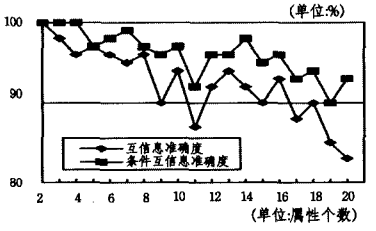

图 2 一对一匹配 & 满射匹配

图 2 中 20 个属性值皆随机取得。由图 2 可知，条件互信息求得的值，匹配准确度都在 90%以上。这样的准确度比互信息准确度都要高。同时，条件互信息匹配准确度在属性数为奇数、偶数的时候，准确度会有所波动。随着属性个数的增多，匹配的精度会略有下降。满射匹配情况在经过实验后可得到的匹配准确度与一对一匹配略有下降。整体的准确度匹配水平与一对一匹配较相似。

5.2.2 部分匹配

每个统计图中横轴为属性个数，纵轴为匹配准确度(单位:%)。其中准确度用 P 表示，c 为通过条件互信息得到的成果匹配结果，n 为两匹配表之间实际正确匹配对数量，则准确匹配度公式为:$P=c/n$。

实验中，两张关系表中的属性不完全一样，属性值随机产生。通过多重迭代方法及条件互信息匹配方法得到匹配结果与原先匹配结果准确度情况，如图 3 所示。

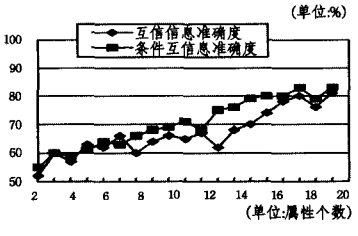

图 3 部分匹配

图 3 为部分匹配准确度情况。由于部分匹配的特殊情况，有一些属性是没有对应匹配属性的。同时，参数 θ 的取值是不确定的。这里，在给定参数 θ 数值的情况下统计条件互信息与互信息的准确度。图 3 便是参数 θ 取 1.0 时，条件互信息与互信息匹配的情况。由图可知，条件互信息的准确度仍然比互信息的准确度要高。在属性个数越来越多的情况下，匹配准确度在 80%左右徘徊。属性个数少与属性个数多的准确率区别明显。此外，θ 参数取值的异同也会影响最后的匹配结果。如果参数 θ 越大，则相对距离就越小，匹配效果也就越好。相反，如果参数 θ 取值越小，距离值便越大。匹配效果越差。这里参数 A 取值的大小是依照不同匹配效果而定的。如果匹配结果需要数量不多但匹配可靠度高的结果，那么参数 θ 则应相应取大一些。相反，参数 θ 取值应相应小一些。

结束语 本文针对模式匹配不准确问题，提出了两步解决方法，即多重迭代筛选方法及条件互信息匹配方法，解决了模式匹配不准确的问题，并提高了匹配的效率。

下一步研究工作是研究多对最优匹配属性对匹配结果的影响及效率，并需要对部分匹配距离公式模型进行优化，使得匹配的准确度能更高，并简化算法。此外，今后需进一步研究不确定模式匹配中，同一表属性值相似度分布相同情况下的特殊模式匹配问题。

参 考 文 献

[1] 翁年凤，刁兴春，曹建军，等. 不确定模式匹配研究综述[J]. 计算机科学，2011，38(12)：1-5

[2] Doan A H，Halevy A Y. Semantic integration research in the database community：A brief survey [J]. AI magazine，2005，26(1)：83

[3] Kang J，Naughton J F. On schema matching with opaque column names and data values[J]. International Conference on Management of Data：Proceedings of the 2003 ACM SIGMOD international conference on Management of data，2003，9(12)：205-216

[4] Jaiswal A，Miller D J，Mitra P. Schema matching and embedded value mapping for databases with opaque column names and mixed continuous and discrete-valued data fields [J]. ACM Transactions on Database Systems (TODS)，2013，38(1)：2

[5] Rabinovich B，Last M. Uninterpreted Semi-Automatic Schema Matching Approach Using Inter-Attribute Dependencies[C]// NATO Workshop on Semantic Interoperability Framework. Oslo. Norway. 2011

[6] 吕锋，王虹，刘皓春. 信息理论与编码[M]. 北京：人民邮电出版社，2004：1-200

[7] 王尊芳，石生明. 高等数学(第三版)[M]. 北京：高等教育出版社，2003

[8] Chen W，Guo H，Zhang F，et al. Mining schema matching between heterogeneous databases[C]// 2012 2nd International Conference on Consumer Electronics，Communications and Networks (CECNet). IEEE，2012：1128-1131

(上接第 59 页)

[4] 张德富，彭煜，朱文兴，等. 求解三维装箱问题的混合模拟退火算法[J]. 计算机学报，2009，32(11)：2147-2156

[5] He Kun，Huang Wen-qi. An efficient placement heuristic for three-dimensional rectangular packing[J]. Computers & Operations Research，2011，38(1)：227-233

[6] 何琨，黄文奇. 三维矩形 Packing 问题的拟人求解算法[J]. 中国科学(F 辑)，2010，40(12)：1586-1595

[7] Ford L R，Fulkerson D R. Maximal flow through a network [J]. Canadian Journal of Mathematics，1956，8：399-404

[8] Andrew V G，Robert E T. A new approach to the maximum flow problem [J]. Journal of the ACM，1988，35(4)：921-940

[9] Prüfer H. Neuer Beweis eines Satzes über Permutationen [J]. Archiv für. Mathematik und Physik，1918，27：742-744