

基于属性依赖关系的模式匹配方法

刘金岭,冯万利

(淮阴工学院 计算机工程学院, 江苏 淮安 223003)

摘 要: 结合数据实例信息和结构信息给出了一种基于属性依赖关系的模式匹配方法,该方法先是根据实例信息计算出各属性间的依赖度,进而建立关系的属性依赖图,依据属性依赖图计算出属性间的依赖结构相似度,最后确定了属性之间相对应的关系.在后面的实验中可以看出其查准率、查全率和一些性能指标与其它方法相比较优.

关键词: 属性依赖;模式匹配;相似度

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2011)12-0167-04

Pattern Matching Method Based on Attribute Dependence

LIU Jin-ling, FENG Wan-li

(Computer Engineering Faculty, Huaiyin Institute of Technology, Huaian 223003, China)

Abstract: In this paper, we introduce a method based on attribute dependence, which integrates the data instance information and structure information to supply matching. At first, we calculated the degree of attribute dependence according to the data instance information, and further establish the graph of attribute dependence and calculated the degree of the degree of structure similarity according to the graph of attribute dependence. At last, the paper gets the mapping between attributes according to the degree of structure similarity. Experiments are conducted and the results show that this algorithm is better than other algorithms in various performance metrics.

Key words: attribute dependence; schema matching; similarity

1 引言

模式匹配是对模式间的两个元素的操作,在关系模式中通常是两个关系的属性之间进行匹配并输出其结果.目前针对自动和半自动的模式匹配问题的研究已有许多成果^[1-2],在文献[3]中,先是对源模式和目标模式利用重复数据检测算法检测其相近数据,一般来讲,相同数据对应的元素相同,然后根据检测出的相近数据得到相互对应的元素对.这种方法对于确定数据的“相近”是很难的.文献[4]中先是利用图结构来表示源模式和目标模式,再给出源模式和目标模式名称的匹配及语义相似度的算法,而后对源模式和目标模式中的每一对节点进行验证,检测出所有的候选匹配对集合,根据候选匹配对

集合进一步调整为更优的相似度.由于这种方法利用了图结构,因此所表达的节点信息过多,这样就是时间复杂度很高.近几年来,研究者对于模式匹配的方法大都以关系模式属性自身的信息为主,其他信息做辅助作用进行匹配的.然而通常设计者对同一对象的描述其信息的差异也比较大,因此单独地利用属性自身信息进行匹配的效果很不理想.本文基于属性语义依赖关系的模式匹配方法(pattern matching of attribute dependence relation, PMADR),该方法首先根据关系模式中的数据信息来计算关系属性间的依赖程度,以得到该模式结构信息,然后根据模式结构信息建立该模式的属性依赖图,再根据属性依赖图计算属性间的结构相似度,最后得到模式属性间的映射关系.

算法 1:

$Q := \Phi, \text{depsim1} = 0$

DO While $X_A \neq \Phi$

$\forall x \in X_A Q := \text{CMS}(x) \cap Y_B$

If $Q \neq \Phi$ then

任取 $y \in Q, \min := \text{abs}(\text{dep}(T, B, y) - \text{dep}(S, A, x)),$

$t := y$

DO While $Q \neq \Phi$ // 根据命题 1 查找符合条件的属性

取 $y \in Q$

If $\text{abs}(\text{dep}(T, B, y) - \text{dep}(S, A, x)) < \min$ then

$\min = \text{abs}(\text{dep}(T, B, y) - \text{dep}(S, A, x))$

$t := y$

End If

$\text{depsim1} = \text{depsim1} + x \times$

$e^{-\text{abs}(\text{dep}(S, A, x) - \text{dep}(T, B, t)) \times \mu}$ [8]

$Q := Q - \{y\}$

End DO

End If

$X_A := X_A - \{x\}$

End DO

再对 Y_B 进行如上计算, 得 depsim2

$\text{Depsim} = \text{depsim1} + \text{depsim2}$

$\text{depsim}(S, A, T, B) = \text{depsim} / (|X_A| + |Y_B|)$

其中 $|X_A|$ 、 $|Y_B|$ 分别表示 X_A 、 Y_B 中所含属性的个数。

类似地, 可以计算属性匹配对 S, A, T, B 的确定相似度 $\text{detsim}(S, A, T, B)$ 。

根据文献[8], 有如下定义:

定义 4 设 A 是源关系 S 的属性, B 是目标关系 T 中与 S 相匹配的属性, X_A 、 Y_B 分别是属性 S, A 、 T, B 的依赖结构信息集, 则 X_A 、 Y_B 的依赖相似度 $\text{depsim}(X_A, Y_B)$ 定义如下^[8]:

$$\text{depsim}(X_A, Y_B) = \max(\sum_{m \in X_A} \sum_{n \in Y} \text{depsim}(S, A, T, B) \times e^{-|\text{dep}(S, A, m) - \text{dep}(T, B, n)| \times \mu}) \quad (1)$$

类似地, 可以定义确定相似度 $\text{detsim}(X_A, Y_B)$ 。

为了更真实地反映属性间结构相似程度的依赖结构相似度, 对所有候选匹配对的依赖结构相似度进行调整。算法如下:

输入: 源关系 S 和目标关系 T 的所有属性匹配对集合 SET , 预计调整次数 N 和阈值 λ

输出: 调整后的所有属性匹配对依赖相似度

算法 2:

Do While $SET \neq \Phi$

任取 $(S, A, T, B) \in SET$

If $\text{depsim}(S, A, T, B) > \lambda$ then

For $i = 1$ to N

$\text{depsim}(S, A, T, B) = \alpha \times \text{depsim}(S, A, T, B) + \beta \times \text{depsim}(X_A, Y_B)$ // 加权平均, $\alpha + \beta = 1$

Next i

输出 $\text{depsim}(S, A, T, B)$

End Do

该算法的目的是将候选匹配对的依赖结构相似度的变化都小于等于阈值 λ , 若经过 N 次调整仍未达到要求的匹配对 则在第 N 次调整后结束。

2.3 确定模式匹配的映射关系

模式匹配的第二个阶段是关于映射关系的构造, 这是模式匹配过程的一个关键步骤。对于模式匹配映射关系的确定, 文献[4]介绍了“稳定婚姻法”。一般地, 两个属性间的相似度不但与两个属性间的相似程度有关, 而且还与其他候选匹配对的相似程度有关, 另一方面, 两属性间相似度的具体数值一般不表达实际意义。这样就使得“稳定婚姻法”对于不同属性直接比较相似度的问题出现了困难。

命题 2 对于属性 $A \in$ 源模式, 设属性 A 候选匹配集 $\text{CMS}(S, A)$, 若值越大, 则说明 A 在目标模式中存在实际匹配的可能性就越大, 反之 A 在目标模式中存在实际匹配的可能性就越小。

本文采用类似于文献[8]的做法, 利用加权平均法进行合并到每个属性匹配对总的相似度:

$$\text{Sim}(S, A, T, B) = \alpha \times \text{Attrsim}(S, A, T, B) + \beta \times \text{depsim}(S, A, T, B) + \gamma \times \text{detsim}(S, A, T, B) \quad (2)$$

利用式(2), 由稳定婚姻法来确定最终的映射关系。

3 实验及结果分析

根据文献[9], 本文选择模式匹配研究中最常用的三个评价指标查准率(precision)、查全率(recall)和全面性(overall)来评判实验结果。

在关系模式匹配中常常要利用数据实例与结构信息这两类重要信息, 典型的方法有文献[3]中的 Duplicates 方法和文献[4]中的 SF 方法。试验数据是在 4Fang 财务管理系统 v2008 中取了 25 个关系及其它们的 224 个属性作为源模式信息, 记为 DB1; 在金手指财务管理系统 2005 企业版(第二版)中取了 27 个关系及其它们的 252 个属性作为目标模式信息, 记为 DB2。测试实例的数据信息为 5 000 条。

由于 Duplicates 方法和 SF 方法都是利用结构信息来进行匹配的方法, 而 PMADR 方法是利用实

例信息和结构信息相结合的方法,即利用数据实例信息计算得到的结构信息(属性依赖关系)来辅助匹配,因此 PMADR 方法要比 Duplicates 方法和 SF 方法更有效.为了验证本文方法的有效性,将本文方法 PMADR 与 Duplicates 方法^[3]、SF 方法^[4]在查准率、查全率和全面率上做一个对比试验.对于两种试验来说,数据量的大小对实验的结果都会有影响.在本文试验中取含有 50, 200, 1 000, 5 000 四种不同数据量大小的数据集进行对比试验.试验结果如图 3 所示.

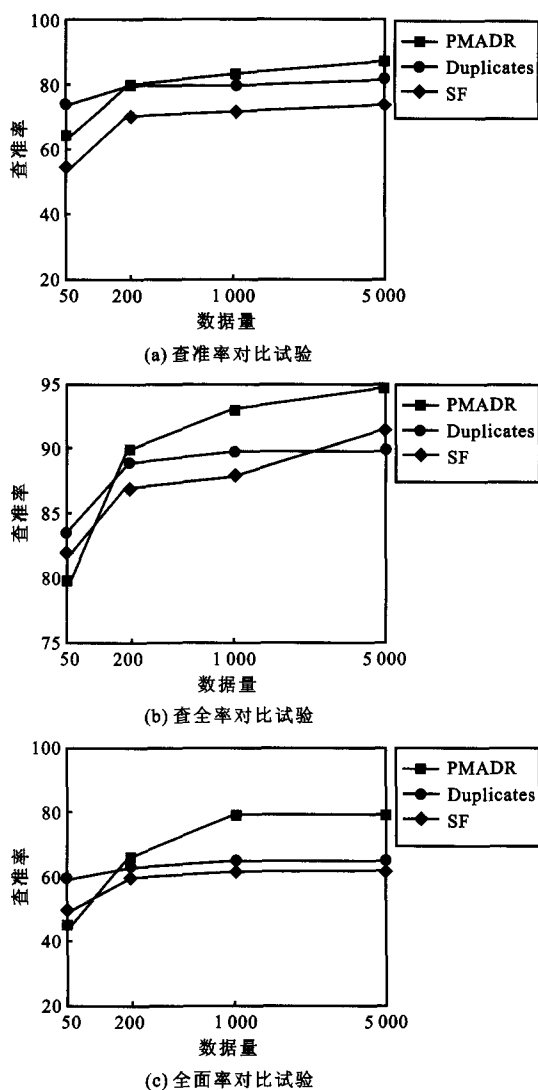


图 3 三种方法在不同数据集上的对比试验

由图 3(a)的查准率对比试验中可以看出, SF 方法比其它两种方法的查准率要低一些,而当数据量较小(≤ 200)时, PMADR 方法查准率低于 Duplicates 方法,当数据量较大(> 200)时, PMADR 方法查准率明显增高;在图 3(b)的查全率对比图中可以看出,当数据量较大(> 180)时 PMADR 方法查全

率明显高于其他两种方法;图 3(c)的全面性指标上, PMADR 方法也高于其他两种方法,即节省的工作量也比其它两种算法要多.因此,综合试验结果分析可以看出,当数据的数量超过 1 000 时,本文方法 PMADR 的查准率、查全率和全面率要比 Duplicates 方法和 SF 方法有大幅度提高,也就是说本文算法在很大程度上提高了模式匹配的精确度.但是,当数据量增大到一定程度时,其查准率的提高幅度并不太大,全面率基本上处于不变状态.事实上,随着试验实例数据量的增大,算法的复杂度也会随之增加.

4 结语

目前的模式匹配方法大多以属性自身信息为主导来进行匹配,由于关系模式的属性本身的信息有时存在的差异比较大,因此单纯的利用属性自身信息进行匹配达不到理想的效果,这也是由于有些属性信息在模式匹配过程中产生一些误导作用.而数据实例信息一般能够最真实地反映属性的语义信息,能够在模式匹配过程中发挥更重要的作用,因此,本文先是利用数据实例信息得到属性间的属性依赖度,得到关系属性依赖图,并进一步结合数据实例信息与结构信息这两类模式匹配中的重要信息构造了辅助匹配模式,试验表明,此方法在一定程度上能够提高模式匹配的精确度.

参考文献:

- [1] Huimin Zhao. Semantic matching across heterogeneous data sources[J]. Communications of the ACM, 2007 (50):45 - 50.
- [2] 张娜,张剑. 一个快速的字符串模式匹配改进算法[J]. 微电子学与计算机, 2007, 24(4):102-105.
- [3] Alexander Bilke, Naumann F. Schema matching using duplicates[C] // Proceeding of the 21st International Conference on Data Engineering (ICDE), Berlin, Germany: IEEE, 2005.
- [4] Sergey Melnik, Hector Garcia Molina, Erhard Rahm. Similarity flooding: a versatile graph matching algorithm and its application to schema matching. [C] // Proc of ICDE, San Jose, California, : IEEE, 2002.
- [5] 任永昌,邢涛,于忠党,等. 数据库规模估算数学模型研究[J]. 微电子学与计算机, 2009, 26(7):13-16.
- [6] 潘超,杨良怀,龚卫华,等. 模式匹配研究进展[J]. 计算机系统应用, 2010, 19(11):265-277.

(下转第 174 页)

4.3 Linux 操作系统内核移植

Linux 是一个开源的类 UNIX 操作系统,在嵌入式领域应用的很广泛. Linux 支持许多的嵌入式处理器结构,包括 ARM、PowerPC、MIPS 等. 由于 Linux 开放源代码,因此有很多软件开发工作都是基于 Linux 的.

Linux 不支持 ARCA3 体系结构,因此需要修改 Linux 的底层硬件驱动,将 Linux 移植到 ARCA3 体系结构. 需要修改的地方有 CPU 体系结构和外设的驱动.

用 ARCA3 交叉编译器重新编译 Linux,可以通过 U-Boot 来引导 Linux 操作系统. 图 4 所示为 Linux 在 FPGA 硬件原型的启动过程和 Linux shell 的执行,在 PC 主机通过串口控制台与 shell 交互.

```
Linux NET4.0 for Linux 2.4
Based upon Swansea University Computer Society NET3.039
Starting kswapd
Serial driver version 5.05c (2001-07-08) with no serial options enabled
ttyS00 at 0xe0001000x (irq = 7) is a Arcal6550
ttyS01 at 0xe0001400x (irq = 8) is a Arcal6550
ttyS02 at 0xe0001800x (irq = 9) is a 8250
ttyS03 at 0xe7ff03f8x (irq = 68) is a 8250
ttyS04 at 0xe7ff02f8x (irq = 67) is a 8250
block: 64 slots per queue, batch=16
RAMDISK driver initialized: 16 RAM disks of 8192K size 1024 blocksize
RAMDISK: Compressed image found at block 0
Freeing initrd memory: 384k freed
VFS: Mounted root (ext2 filesystem).
Freeing unused kernel memory: 40k freed
Welcome to ARCA3_FPGA Linux

-/bin/sh: /bin/hostname: not found
# ls
bin      home     lost+found  sbin      var
dev      lib      proc       tmp
etc      linuxrc  root       usr
# cd bin
# vi a_
```

图 4 Linux 在 FPGA 硬件原型的运行

5 结束语

在 ARCA3 和 AMBA 架构基础上添加 DDR2

控制器、串口、定时器、中断控制器等一些外设和 IP 核,构建一个简单的嵌入式 SoC. 在 Xilinx 公司的 ML510 FPGA 评估板上实现整个系统的软硬件协同验证环境,成功移植 U-Boot 和 Linux 内核到 ARCA3 体系结构,充分利用软硬件协同验证了 SoC 的功能. FPGA 验证其实是一种硬件加速的方法,优点是速度快,短时间内可以跑大量的验证程序^[3].

SoC 的 FPGA 原型验证系统实现对基于 ARCA3 的软件开发有重要的意义,程序人员可以在 Linux 操作系统上快速开发基于 ARCA3 的应用程序. 还可以为基于 ARCA3 的各种抽象层次的 IP 验证提供软硬件协同仿真环境,缩短系统设计、IP 集成和验证的时间. 本文的创新之处还在于编写一个 mini_boot 程序,用于加速 FPGA 的软件调试过程.

参考文献:

- [1] 赵刚, 侯立刚, 刘源. 基于 SoC 设计的软硬件协同验证方法学[J]. 微电子学与计算机, 2006, 23(6): 24-26.
- [2] 孔璐, 吴武臣, 侯立刚, 等. 基于 LEON3 的 SOC 平台设计与 SPI 嵌入[J]. 微电子学与计算机, 2010, 27(3): 172-176.
- [3] 虞致国, 魏敬和. 基于 JT AG 的 SoC 软硬件协同验证平台设计[J]. 微电子学与计算机, 2010, 27(10): 160-162.

作者简介:

陈达燕 男, (1987-), 硕士研究生. 研究方向为嵌入式系统、数字集成电路.

(上接第 170 页)

- [7] Rahm E. and Bernstein P A. A survey of approaches to automatic schema matching[J]. VLDB Journal, 2001, 10 (4): 334-350.
- [8] 李国徽, 杜小坤, 胡晓, 等. 基于函数依赖的结构匹配方法[J]. 软件学报, 2009, 20(10): 2667-2678.
- [9] David Aumuellner, Do Hong-Hai, Sabine Massmann, et al. Schema and ontology matching with COMA++[C] //Proc SIGMOD, Baltimore, Maryland: IEEE, 2005.

作者简介:

刘金岭 男, (1958-), 硕士生导师, 教授. 研究方向为数据库、数据仓库、数据挖掘.

倪晓红 女, (1974-), 硕士, 讲师. 研究方向为网络及数据库应用.

冯万利 男, (1973-), 硕士, 讲师. 研究方向为数据库、数据仓库、数据挖掘.