模式匹配中的依赖冲突

杜小坤 李艳红 涂 韬
(中南民族大学计算机科学学院 武汉 430074)

摘 要 通过分析已有匹配方法的缺陷,提出了一种利用依赖冲突选取匹配关系的新方法。首先为目标模式中每个元素选取候选匹配,然后计算每个全局匹配方案的冲突值,最后选取冲突值最小的匹配方案作为最终结果。实验表明,该方法能够显著提高匹配结果的准确率,并使得后续数据映射结果的优化操作更省时。

关键词 模式匹配,依赖冲突,数据映射
中图法分类号 TP301.131 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.4.048

Dependency Conflict in Schema Matching

DU Xiao-kun LI Yan-hong TU Tao
(College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China)

Abstract Through analyzing the drawback of existing matching method, a new method was proposed to select matching relation between elements based on the dependency conflict. At first, the candidate match of elements in target schema is selected, and then the value of conflict is calculated for each overall matching solution, at last the solution with minimum value of conflict is selected as the final result. Experimental results show that the precision of matching result is increased and the time for the optimization procedure of the mapping data is reduced with this strategy.

Keywords Schema matching, Dependency conflict, Data mapping

1 绪论

模式映射主要研究异构数据源间的数据转换问题,是信息广度扩展的重要手段,广泛应用于数据空间、商业智能、P2P数据库等热点研究领域。已有的研究成果将模式映射过程分为模式匹配和数据映射两个步骤分别进行,模式匹配主要研究如何获取模式间高准确率的元素匹配关系,根据元素匹配关系进行数据转换并优化则是数据映射研究的主要内容。已有的研究成果将模式匹配和数据映射两个操作独立地进行研究,取得了较为丰富的成果[1-4]。目前已有的成果虽然能够获取较高质量的映射结果,但也存在较为明显的缺点。一方面,虽然模式匹配能够得到较高准确率的匹配结果,但距离自动模式匹配的目标仍然有较大的差距,匹配结果的准确率仍然有待提高。另一方面,数据映射结果的优化是一个十分耗时的过程,尤其是当模式中包含较多的数据或者错误数据较多时,在当前大数据环境下该问题显得尤为突出。

已有的模式匹配研究都基于这样一个策略:尽可能挖掘更多的元素相互匹配的证据以证明元素是相互匹配的,并将这些证据量化成相似度指标,最终根据相似度指标选取匹配关系[5-8]。由于建模语言自身的缺陷,从模式中挖掘出的信息并不能够完全准确地反映元素语义,甚至还会出现误导的情况,因此当该策略将匹配准确率提高到一定程度后,继续提高将变得不现实。但匹配结果的准确与否对数据映射至关重要,个别匹配错误即能够生成大量的错误数据(一个错误匹配对可导致其中元素对应的所有元组的数据错误),对这些错误数据进行优化是数据优化过程最耗时的部分(一些错误甚至是不可能被优化的)。

本文通过对数据映射结果中的常见问题进行分析,提出了匹配结果中依赖冲突的概念,并给出了通过消除或减少匹配结果中的依赖冲突提高匹配结果准确率并进而减少数据优化执行时间的方法。本文的主要创新点如下:

(1)提出依赖冲突的概念。依赖冲突是元素匹配的负面信息,一个元素匹配对若与另一个匹配对存在冲突,则二者相互匹配的概率下降,这是一种与已有方法完全不同的角度。

(2)给出一种利用依赖冲突提高匹配准确率的新方法。实验结果表明,该方法能够提高匹配结果准确率。

本文第2节介绍了相关工作;第3节介绍了依赖冲突及分类检测算法;第4节介绍了利用依赖冲突选取匹配关系的新方法;第5节是实验对比分析;最后是结论与展望。

2 相关工作

2.1 模式匹配

模式匹配的主要目标是获取高准确率的元素匹配关系,目前已有的研究成果基本都从如下3个方面着手:①获取新

到稿日期:2014-09-29 返修日期:2014-11-03 本文受国家自然科学基金(61173049),湖北省自然科学基金(2014CFB915),中央高校基本业务专项经费(CZQ14015)资助。

杜小坤(1980-),男,博士,讲师,主要研究方向为数据集成、商业智能,E-mail:hustdxkun@163.com;李艳红(1973-),女,博士,副教授,主要研究方向为路网查询;涂 韬(1984-),女,硕士生,主要研究方向为数据集成。

的辅助信息；②优化辅助信息的使用方式；③提供高效的用户干预手段。下面分别对其中的典型算法进行介绍。

①获取新的辅助信息。在获取新的辅助信息方面，Erhard Rahm 和 Philip A. Bernstein 等人在文献[1,2]中分别对2001 年前及 2001 至 2011 年间的研究进行了总结。其中多数研究成果[5,6,8,9]利用最直观的信息如模式自身信息、结构信息、数据实例信息等获取元素间的匹配关系。典型的元素自身信息有元素名称、数据类型、说明信息等，对元素名称、说明信息等文本类信息首先进行一致性处理(消除其中缩写、简写，发现同义、近义词等)，然后利用启发式方法计算文本间的相似度，并以此作为元素自身信息相似度，对数据类型则利用各种数据类型间的兼容性表示其相似性[9]。PFD_based 方法[8]利用元素间的函数依赖关系描述结构信息，并且通过对元素数据进行分析发掘隐含的依赖关系以丰富模式的结构信息。该方法虽然利用了较多元素间的关联信息，但由于仍然以元素为基础描述结构关联，存在信息量大、处理耗时，且描述不够准确等问题。

除了上述较为直观的信息外，研究人员还发现一些其它类型的信息可用以辅助匹配。例如：申德荣等提出的 SKM 模型[10]利用模式结构信息以及同领域其它模式间的匹配信息获取元素匹配关系。Hazem Elmeleegy 等人[11]提出了一种利用查询日志辅助获取元素匹配的 Usage_based 方法，该方法以元素在查询语句中出现的位置信息为依据辅助匹配。Pinkel C[12]提出一种利用本体辅助获取较复杂模式间元素匹配关系的方法。

②优化信息使用方式。除了挖掘新的辅助信息外，还可以对已有信息的使用方式进行优化。为了综合多种信息提高匹配准确率，Aumueller D 人[13]提出了一种可动态配置各种不同类型信息匹配器的 COMA＋＋模型。如何对 COMA＋＋模型中的各种匹配器进行配置对匹配结果至关重要，同时几乎所有的匹配器都需要由用户设定各种参数，参数设置的优劣会对匹配结果产生较大影响。Peukert E 等[14]阐述了上述问题并给出了相应的自动配置(Self-configuring)模型，其取得了较好的效果。

③高效的用户干预手段。在一些对匹配准确率有较高要求的应用环境中，用户对匹配结果进行人工干预不可避免。高效、方便的用户干预方式也是目前模式匹配研究的重点。Li Qian 等[15]提出由数据驱动的 Sample-driven 映射模型，其通过对用户在目标模式中输入的数据进行分析直接获取源、目标模式间的数据映射关系。Chen Jason Zhang 等[16]提出借助用户对现实世界的敏感直觉来辅助映射的思路。针对匹配过程中的不确定因素生成一些简单问题，通过 Crowd-Sourcing平台向用户提问，再根据用户的回答对映射结果做相应的修正并继续提问，直至获取满足要求的映射结果。

2.2 数据映射

数据映射研究如何根据元素匹配关系对数据进行转换。Renee J. Miller 等[17]首先提出 Clio 原型实现了自动映射的功能，但其映射结果中却存在如下问题：①大量数据冗余；②数据关联丢失。后续的研究围绕这两个问题展开。

对于映射结果中的数据冗余问题，Ronald Fagin 等[18]进行了较为系统的分析，提出了核心(Core)映射(是所有相互同构的映射结果中包含冗余数据最少的映射)的概念，给出了将 Clio 原型的映射结果转换核心映射的优化算法。核心映射虽然消除了映射过程中产生的冗余，但元组数较多时，优化算法耗时较长。Giansalvatore Mecca 等[3]提出的＋spicy 系统在生成映射语句时就综合考虑获取核心映射的因素，去除了耗时的优化算法，减少了运行时间，但并不能保证获取的一定是核心映射。

模式映射必须保持数据间的关联，映射结果中的数据关联可能是直接的，也可能是间接的。间接的关联需要通过其它数据来间接实现，当该数据为空值时，我们必须为其生成合适值来实现数据关联的保持。Clio 原型中使用能够准确描述数据关联的 Skolem 函数填充空值，但却未对空值间的关联予以考虑。Bogdan Alexe 等[19]提出了一种通过对映射前数据关联进行分析来建立映射后空值间关联关系的启发式方法，很好地解决了空值关联问题，但这是一个十分耗时的过程。

3 依赖冲突

错误的匹配关系经常会在数据映射时引发异常，导致后续的优化算法十分耗时且难以获取好的优化结果。本节对数据映射结果中最常见的两种问题产生的原因进行分析，并给出依赖冲突的定义。

图 1 的左部为模式匹配算法获取的元素匹配关系及其对应的数据映射表达式，右部为实际的数据映射结果。根据元素匹配关系，目标模式中关系 StuInfo 的各个元素分别与源模式中不同关系的元素匹配，根据文献[17]，若源模式中两个元组的数据能够合并到目标模式的同一元组中，则对应的两个关系间必须有外键连接。由于图 1 中关系 StudentInfo 和 TeacherInfo 之间没有外键关联，因此数据映射结果中这些无法关联的属性值必须存入不同的元组，从而导致关系 T. StuInfo 的映射结果中存在大量的空值，且部分元组的主键对应元素为空，违反了关系数据库的实体完整性规则，无法将其正确存入到数据库中。

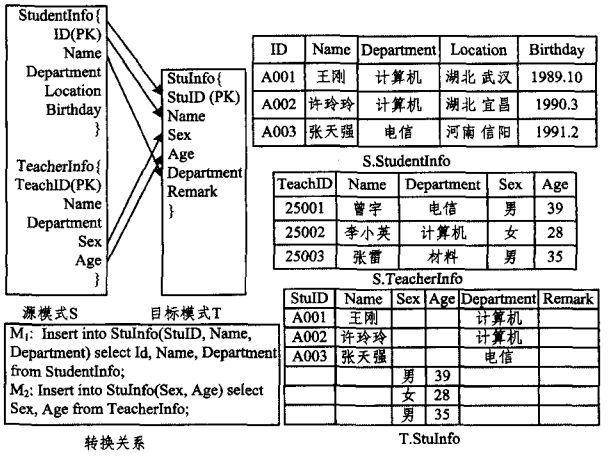
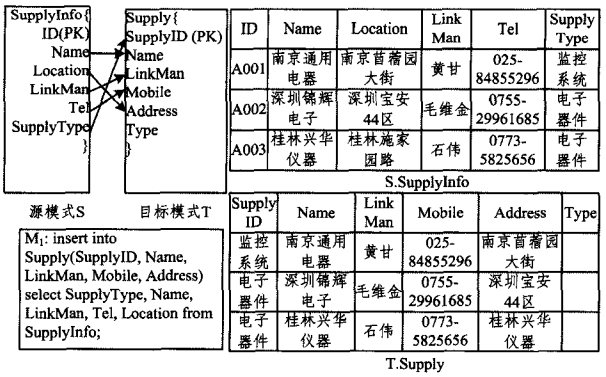


图 1 源模式 S 和目标模式 T 间的映射结果(1)

图 2 的模式匹配结果中，目标模式中的元素 T. Supply. SupplyID 与源模式中元素 S. SupplyInfo. SupplyType 匹配，由于前者是关系的主键而后者不是，因此图 2 右下方的映射数据中，主键的唯一性约束被破坏。上述两种情况是数据映射结果中经常出现的两种异常情况，并且在优化过程中难以对其进行有效的处理。下面我们从函数依赖关系的角度分析上述异常产生的原因。



ID	Name	Location	Link Man	Tel	Supply Type
A001	南京通用电器	南京苜蓿园大街	黄甘	025-84855296	监控系统
A002	深圳锦辉电子	深圳宝安44区	毛维金	0755-29961685	电子器件
A003	桂林兴华仪器	桂林施家园路	石伟	0773-5825656	电子器件

S.SupplyInfo

Supply ID	Name	Link Man	Mobile	Address	Type
监控系统	南京通用电器	黄甘	025-84855296	南京苜蓿园大街	
电子器件	深圳锦辉电子	毛维金	0755-29961685	深圳宝安44区	
电子器件	桂林兴华仪器	石伟	0773-5825656	桂林兴华仪器	

T.Supply

图 2　源模式 S 和目标模式 T 间的映射结果(2)

图 1 中映射数据的主要错误为元素 Sex 和 Age 对应的数据与其它元素对应的数据无法关联；从函数依赖的角度来看，即目标模式中存在的依赖关系 StuID→Sex 在其分别对应的元素 StudentInfo. ID 和 TeacherInfo. Sex 之间并不存在(对依赖关系 StuID→Age 也是如此)。图 2 中映射数据的主要错误为主键函数决定其它元素的依赖关系在映射数据中并不存在；从函数依赖的角度来看，即目标模式中主键函数决定其它元素的依赖关系在其对应的匹配元素间也不存在(与图 1 中的情况有所区别的是，这些元素能够通过共同的元素 ID 相互关联)。我们将上述目标模式依赖关系在源模式对应元素间的依赖关系存在差异的问题称为匹配关系间的依赖冲突，下面给出依赖冲突的定义。

定义 1　待匹配模式 S 和 T 中，对目标模式 T 中任一依赖关系 $A\rightarrow B$，设 A,B 在源模式中的匹配元素为 A'、B'，若源模式中不存在函数依赖关系 $A'\rightarrow B'$，则称匹配对(A',A)和(B',B)间存在依赖冲突。

对图 1 中的匹配对(StudentInfo. ID, StuInfo. StuID)和(TeacherInfo. Sex, StuInfo. Sex)，模式 T 中存在依赖关系 StuInfo. StuID→StuInfo. Sex，而其匹配元素在源模式 S 中却并不存在依赖关系，所以上述匹配对间存在依赖冲突。图 2 中的匹配对(SupplyInfo. SupplyType, Supply. SupplyID)和(SupplyInfo. Name, Supply. Name)也存在依赖冲突。这两个依赖冲突对映射结果产生的影响却并不相同，图 1 中由于元素 StudentInfo. ID 和 TeacherInfo. Sex 间不存在任何关联关系，从而导致大量的元素违反实体完整性规则；而图 2 中由于元素 SupplyInfo. SupplyType 和 SupplyInfo. Name 以及其它元素间虽然没有直接的函数依赖关系，但却可以通过元素 ID 进行间接关联，所以对映射数据造成的实际影响远小于前者。

定义 2　同一模式中任意两个元素 A、B，若 A、B 间满足如下两个条件之一，则元素 A、B 间的数据能够相互关联，称元素 A、B 间存在数据关联。

①A、B 间存在函数依赖关系 $A\rightarrow B$ 或 $B\rightarrow A$；

②模式中存在另一个元素 C 且 $C\rightarrow\{A,B\}$。

根据定义 2，若两个元素间存在函数依赖关系或存在共同的决定元素(即同属于某个元素的决定元素闭包)，则其对应的数据存在关联，例如图 2 中源模式关系 SupplyInfo 中元素 SupplyType 和 Name 之间存在数据关联，因为对任意一个元组中元素 SupplyInfo 对应的数据都能唯一确定一个(通过共同的决定元素 ID)元素 Name 对应的数据与其关联。对存在依赖冲突的两个匹配对来说，若目标模式中的元素 A'、B' 间存在数据关联，则对映射数据的影响较小，据此我们将匹配元素间的依赖冲突进行如下分类。

定义 3　匹配对(A,A')和(B,B')间存在依赖冲突，若源模式中元素 A、B 间存在数据关联，则称(A,A')和(B,B')间存在相对依赖冲突，否则称(A,A')和(B,B')间存在绝对依赖冲突。

对任意给定的匹配结果 M，根据定义 3 可分别获取其中存在的绝对和相对依赖冲突，具体如算法 1 所示。

算法 1　依赖冲突检测分类算法

```
ConflictInspect (M,FDS,FDT)
{
  Input:匹配算法输出的匹配结果 M,模式 S 和 T 中的函数依赖关系
        集 FDS,FDT
  OutPut:匹配结果中存在绝对依赖冲突的映射关系对集合 Abs-
         ConSet(M)和存在相对依赖冲突的映射关系对集合 Rel-
         ConSet(M)
1. foreach M 中的匹配关系(A,A') {
2.   foreach FDT 中的 A'为决定因素的依赖关系 A'→B' {
3.     if((M 中存在匹配关系(B,B'))&&(不存在依赖关系 A→B)){
4.       if((FDS 中存在依赖关系 B→A) ‖ (A,B 同时出现在某个元
         素的决定元素集中))
5.         insert conflict[(A,A'),(B,B')] into RelConSet(M);
6.       else
7.       insert conflict[(A,A'),(B,B')] into AbsConSet(M);
}}}}
```

本节对映射数据中的两种常见问题进行分析，给出了依赖冲突的定义；由于不同类型的依赖冲突对映射结果的影响不同，因此进一步对依赖冲突进行了分类；最后给出了依赖冲突的检测算法。

4　利用依赖冲突选取匹配结果

已有匹配方法尽可能地寻找模式相互匹配的正面依据，并采用启发式的方法将其量化成相似度指标，最后为每个元素选取相似度较高的元素作为匹配元素。这种匹配策略具有如下两个缺点：

①由于建模语言自身的缺陷，从模式中获取的信息并不能完全准确地描述元素语义，因此通过挖掘更多匹配信息的方式提高匹配准确率的策略在准确率达到一定程度后难以继续提高。

②已有的相似度计算方法都是基于启发式的方法，得到的相似度值大小并无实际意义，单纯地根据相似度值选取的最终匹配结果不够准确。

依赖冲突是元素相互匹配的负面证据，若两个匹配对间存在依赖冲突，则意味着两个匹配对同时存在于匹配结果中的可能性减小。结合已有策略的缺点及依赖冲突的性质，本文提出一种利用依赖冲突概念选取匹配结果的方法。首先利用已有方法计算待匹配模式元素间的相似度，然后为目标模式中每个元素选取候选匹配，最后通过对匹配结果中的依赖冲突进行评价以选取最优的匹配结果。

本文使用 COMA++算法计算源、目标模式元素间的相似度，然后采用 MaxDelta 和 Threshold 策略结合为目标模式中每个元素选取候选匹配。对于目标模式中任意元素 A，得到的候选匹配元素集合为 $CAND(A)$。得到每个元素的候选匹配集合后，从每个候选匹配集合中任意选取一个元素即可得到一个全局模式匹配关系 M。对任意全局模式匹配关系

M,有如下定义:

定义 4 对模式匹配关系 M 中的任一匹配对(A,A'),所有与其存在绝对依赖冲突的匹配对数目称为匹配对(A,A')的绝对依赖冲突数,记为 $AbsConfNum_M(A,A')$;所有与匹配对(A,A')存在相对依赖冲突的匹配对数目称为匹配对(A,A')的相对依赖冲突数,记为 $RelConfNum_M(A,A')$。

定义 5 与目标模式 T 中任意元素 A'存在函数依赖关系的元素数目称为元素 A'的关联数,记为 $ConnNum(A')$。在模式匹配关系 M 中,匹配对(A,A')的绝对依赖冲突数与元素 A'的关联数的比值称为匹配对(A,A')的绝对冲突率,记为 $AbsConfRatio_M(A,A')$,相对依赖冲突数与元素 A'的关联数的比值称为相对冲突率,记为 $RelConfRatio_M(A,A')$。

对任意模式匹配关系 M,根据定义 4 和定义 5 可计算每个匹配对(A,A')的绝对和相对冲突率。匹配对的冲突率越高,则该匹配对对其它匹配对的影响越大,所以匹配结果应尽量降低其中匹配对的冲突以有效减少最终映射数据中的相关问题。一种简单的策略是统计其中冲突的数量,选取冲突数较少的匹配关系作为最终匹配,但这种统计方式存在两个显著的问题。首先,绝对冲突和相对冲突对转换数据产生的影响不同,所以不能简单地等同对待;其次,元素的关联数越大,说明与该元素关联的元素越多,该元素对相关元素的影响越大,所以对不同元素的依赖冲突应分别对待。根据上述分析,本文用式(1)－式(3)计算模式匹配关系 M 的冲突值 $CV(M)$。

$$Abs\_CV(M)=\sum_{(A,A')\in M}^{(A,A')}(w(A')\times AbsConfRatio_M(A,A')) \tag{1}$$

$$Rel\_CV(M)=\sum_{(A,A')\in M}^{(A,A')}(w(A')\times RelConfRatio_M(A,A')) \tag{2}$$

$$CV(M)=\alpha\times Abs\_CV(M)+\beta\times Rel\_CV(M),\alpha+\beta=1 \tag{3}$$

式(1)和式(2)中的 $w(A')$为目标模式中元素 A'的依赖冲突对整个匹配结果的影响值。该值与元素的关联数相关,元素的关联数越大,即与该元素关联的元素越多,则其匹配关系对整个匹配结果的依赖冲突值影响也越大。为使 $w(A')$能够较为准确地描述影响值和关联数间的关系,分别用如表 1 所列的 3 个函数关系式描述其定量关系,通过 5.1 节的实验结果可知,抛物线函数比线性函数和对数函数具有更好的效果。

表 1 依赖冲突影响值 $w(A)$和关联数 $ConnNum(A)$间的定量关系

序号	名称	函数关系式
1	线性函数	$w(A)=a+b\times ConnNum(A)$
2	抛物线函数	$w(A)=a+b\times(ConnNum(A))^2$
3	对数函数	$w(A)=a+b\times\log_2(ConnNum(A))$

依据上述公式可对任意模式匹配结果 M 计算其冲突值 $CV(M)$,并选取冲突值最小的匹配结果作为最终的匹配结果。由于在匹配结果中已尽可能地避免了其中的依赖冲突,因此依据选取的匹配结果进行转换可得到较高质量的数据,大幅缩短了后续优化操作的时间(见 5.3 节实验结果)。

5 实验评价

本文介绍了一种利用匹配结果中的依赖冲突选取高准确率的匹配结果并进而缩短数据优化时间的新方法。为验证该策略的有效性,本节将该策略与一些常用方法相结合,并对结合前后的结果进行对比。实验中的模式分别选取两个不同学校的教务管理系统,其中包含的关系和属性的基本情况如表 2 所列,数据实例采用 DTM Data Generator[1)] 生成。具体测试时将模式导入到 MySql5.2 中,使用 ODBC 连接数据库以获取模式信息。主机硬件采用 Intel Core i3 双核 2.27G 处理器,4G 内存;操作系统为 Windows 7。

表 2 待匹配模式中的关系及属性数目

模式	关系数目	属性数目
S	37	289
T	31	274

5.1 元素关联数与元素冲突影响值的定量关系实验

本文第 4 节中使用 3 种不同的函数关系式对元素关联数 $ConnNum(A)$与依赖冲突影响值 $w(A)$的定量关系进行了描述。为选取准确的定量关系描述,本小节分别使用这 3 种函数关系式描述的定量关系进行匹配结果的选取(元素相似度值的计算采用 COMA++方法进行),最终结果使用查准率、查全率、全面性 3 个指标进行评价[1]。实验结果如图 3 所示。

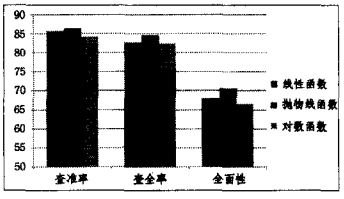


图 3 不同的定量关系获取最终结果的准确率对比

通过图 3 中的数据可知,使用抛物线函数描述的定量关系能够获取更准确的匹配结果,所以在后续的实验中均采用抛物线函数对定量关系进行描述。

5.2 模式匹配结果准确率比较

为验证该策略能提高匹配结果的准确率,我们选取 COMA++方法和 PFD_BASED 方法,将使用依赖冲突策略选取的匹配结果与方法的原始匹配结果进行比较。实验数据如图 4 所示。

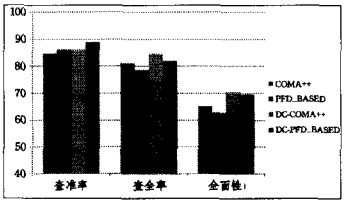


图 4 利用依赖冲突策略和未使用的匹配结果对比

图 4 显示了原始的 COMA++和 PFD_BASED 方法与利用依赖冲突策略选取最终结果的方法在匹配准确率上的对比。从对比结果可知,在两种方法中采用依赖冲突策略选取的匹配结果准确率都比原始方法选取结果的准确率有一定程度的提高。所以利用依赖冲突策略选取匹配结果的方法优于未使用该策略的方法。

5.3 数据优化的时间性能对比

由于本文提出的匹配结果选取策略尽可能地消除匹配结果中的依赖冲突,因此获取的数据映射结果中相应的数据错误较少,数据优化过程也更省时。本节通过实验来验证该性

1) http://www.sqledit.com/dg/download.html

质。为验证对不同规模数据的性能，分别向模式 S 的每个关系中注入 100、500、2000、10000 个元组（分别称为 D1/D2/D3/D4）分别进行对比测试。对比算法使用 COMA++算法，首先使用原始 COMA++算法获取匹配关系 *M1*，然后利用 COMA++与依赖冲突策略相结合获取匹配关系 *M2*；再分别使用 M1 和 *M2* 对 D1/D2/D3/D4 进行数据转换（采用文献[17]中的 Clio 算法）；最后使用文献[18]中方法分别对数据转化结果进行优化以得到核心数据转换方案。对不同转换结果进行优化的时间如表 3 所列。

表 3　优化时间对比

单位：秒	D1	D2	D3	D4
M1	56.7	217	864	3524
M2	12.4	38	146	563

从表 3 可知，在匹配结果选取时考虑其中的依赖冲突能够显著提高映射数据的质量，并缩短优化算法的执行时间。

结束语　本节提出了模式匹配结果中依赖冲突的新概念，给出了依赖冲突的定义以及冲突检测分类算法；最后提出一种对候选匹配结果中的依赖冲突进行分析以选取最终匹配结果的新策略，该策略能够在一定程度上提高匹配结果的准确率。实验结果表明：通过在已有匹配方法中结合新的匹配结果选取策略能够有效提高匹配结果准确率，同时数据映射结果的优化算法所耗费的时间会大幅减少。目前已有的模式匹配方法仅使用元素相互匹配的正面信息，在匹配准确率达到一定程度后再进一步提高会变得很困难，挖掘元素相互匹配的负面信息（证明元素不能匹配）不仅对进一步提高准确率有一定帮助，同时还可有效缩短匹配算法的执行时间（确定某些元素间不能匹配后可不计算其相似度），这将是我们进一步的研究方向。

参 考 文 献

[1] Rahm E, Bernstein P A. A Survey of approaches to automatic schema matching[J]. VLDB Journal, 2001, 10(4): 334-350

[2] Bernstein P A, Madhavan J, Rahm E. Generic schema matching, ten years later[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 695-701

[3] Mecca G, Papotti P, Raunich S. Core schema mappings: Scalable core computations in data exchange[J]. Information Systems, 2012, 37(7): 677-711

[4] Calvanese D, De Giacomo G, Lenzerini M, et al. On simplification of schema mappings[J]. Journal of Computer and System Sciences, 2013, 79(6): 816-834

[5] Sorrentino S, Bergamaschi S, Gawinecki M, et al. Schema label normalization for improving schema matching[J]. Data & Knowledge Engineering, 2009, 69(12): 1254-1273

[6] Bilke A, Naumann F. Schema matching using duplicates[C]// Proceedings of 21st International Conference on Data Engineering. 2005: 69-80

[7] Elmeleegy H, Elmagarmid A, Lee J. Leveraging query logs for schema mapping generation in U-MAP[C]// Proceedings of the 2011 International Conference on Management of Data. Athens Greece, 2011: 121-132

[8] 李国徽，杜小坤，杨兵，等. 基于部分函数依赖的结构匹配方法[J]. 计算机学报，2010，33(2)：240-250

[9] Madhavan M J, Bernstein P A, Rahm E. Generic schema matching with cupid[C]// Proc. of VLDB. 2001: 49-58

[10] 申德荣，余恩运，张旭，等. SKM：一种基于模式结构和已有匹配知识的模式匹配模型[J]. 软件学报，2009，20(2)：327-338

[11] Elmeleegy H, Elmagarmid A, Lee J. Leveraging query logs for schema mapping generation in U-MAP[C]// Proceedings of the 2011 International Conference on Management of Data. Athens Greece, 2011: 121-132

[12] Pinkel C. Interactive Pay as You Go Relational-to-Ontology Mapping[M]// The Semantic Web-ISWC. 2013: 456-464

[13] Aumueller D, Do H H, Massmann S, et al. Schema and ontology matching with COMA++[C]// Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. Chicago, IL, USA, 2005: 906-908

[14] Peukert E, Eberius J, Rahm E. A self-configuring schema matching system[C]// Proceedings of 28st International Conference on Data Engineering. Washington DC, USA, 2012: 306-317

[15] Qian L, Cafarella M J, Jagadish H V. Sample-driven schema mapping[C]// Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. Scottsdale, USA, 2012: 73-84

[16] Zhang C J, Chen L, Jagadish H V, et al. Reducing uncertainty of schema matching via crowdsourcing[J]. Proceedings of the VLDB Endowment, 2013, 6(9): 757-768

[17] Popa L, Velegrakis Y, Hernández M A, et al. Translating web data[C]// Proceedings of the 28th international conference on Very Large Data Bases. VLDB Endowment, 2002: 598-609

[18] Fagin R, Kolaitis P G, Popa L. Data exchange: getting to the core[J]. ACM Transactions on Database Systems (TODS), 2005, 30(1): 174-210

[19] Alexe B, Hernández M, Popa L, et al. MapMerge: Correlating independent schema mappings[J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 81-92

（上接第 212 页）

参 考 文 献

[1] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[C]// ACL'04. 2004

[2] Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis[C]// CIKM'05. NewYork, NY, USA, 2005: 625-631

[3] 中国计算机学会中文信息技术专业委员会 2013 年会评测[OL]. 2013-03-10[2013-09-30]. http://tcci.ccf.org.cn/conference/2013/pages/page04_eva.html

[4] 第五届中文倾向性分析评测（COAE2013）大纲[OL]. 2013-08-01[2013-09-30]. http://www.cipsc.org.cn/hytx/13.html#23

[5] 朱嫣岚，闵锦，周雅倩，等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报，2006，20(1)：14-20

[6] 董振东，董强.《知网》情感分析用词语集[OL]. http://www.keenage.com/html.com/html/c_bulletin_2007.htm

[7] 赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2010，21(8)：1834-1848

[8] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 情报学报，2008，27(2)：180-185

[9] 张华平. NLPIR 汉语分词系统[OL]. http://ictclas.nlpir.org