Hans Journal of Data Mining 数据挖掘, 2020, 10(1), 76-89
Published Online January 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.101008

Hans汉斯

Web Data Exchange Schema Mapping Optimization Method

Yuhang Ji, Gui Li, Zhengyu Li, Ziyang Han, Keyan Cao

Faculty of Information & Control Engineering, Shenyang Jianzhu University, Shenyang Liaoning
Email: syjzjyh@163.com, ligui21c@sina.com

Received: Dec. 25th, 2019; accepted: Jan. 7th, 2020; published: Jan. 14th, 2020

Abstract

Web data exchange is one of the important researches on the integration of Web heterogeneous data sources. It is usually divided into two aspects: instance layer and schema layer. The research in this paper is mainly focused on the mode layer. Because a given source-to-target mode mapping usually makes the data exchange results contain a lot of redundancy, in order to generate data without redundancy as a data exchange kernel solution, this paper designs a homomorphic relationship Schema mapping design and optimization methods. This method first introduces the homomorphic relationship between the schema mappings as the basis of the schema mapping rewriting method. By decomposing the schema mappings, defining the degree of data redundancy generated by different rules, and determining the rules that need to be rewritten. Finally, the given schema mapping is rewritten into a kernel schema mapping that can directly generate a kernel solution, and it is converted into an executable SQL statement to calculate the kernel solution. This paper uses data from China Land Market Network to test the performance of the proposed method.

Keywords

Web Big Data, Data Exchange, Schema Mapping, Core Solution, Homomorphism

Web数据转换模式映射优化方法

纪宇航，李　贵，李征宇，韩子扬，曹科研

沈阳建筑大学信息与控制工程学院，辽宁 沈阳
Email: syjzjyh@163.com, ligui21c@sina.com

收稿日期：2019年12月25日；录用日期：2020年1月7日；发布日期：2020年1月14日

摘　要

Web数据转换是Web异构数据源集成的重要研究之一，通常分为实例层和模式层两方面进行。本文的研

文章引用: 纪宇航, 李贵, 李征宇, 韩子扬, 曹科研. Web 数据转换模式映射优化方法[J]. 数据挖掘, 2020, 10(1): 76-89.
DOI: 10.12677/hjdm.2020.101008

究主要针对模式层，由于给定的源到目标模式映射通常使数据转换结果包含大量冗余，为了生成不含冗余的数据作为数据转换核解，本文设计了一种基于同态关系的模式映射设计与优化方法。该方法首先引入模式映射之间的同态关系作为模式映射重写方法基础，通过对模式映射进行分解，定义不同规则生成的数据冗余的大小程度，确定需要重写的规则。最后将给定的模式映射重写为能够直接生成核解的核模式映射，并将其转换为可执行的SQL语句来计算核解。本文实验使用来自中国土地市场网的数据验证本文方法的有效性。

关键词

Web大数据，数据转换，模式映射，核解，同态关系

1. 引言

最初的数据转换问题(data exchange problem) [1]是由 Fagin 等人提出的，他们给出了数据转换的相关定义。文献[2]中说明了数据转换通常将源数据作为输入，由一组模式映射的集合(也叫元组生成依赖关系)对源数据进行选择，将其转换为满足给定的模式映射的目标数据集，在此基础上还给出了数据转换过程中通解、核解的概念以及基本求解方法，并将一些相关领域知识转换为约束条件融入到数据转换求解算法中(核解是保留映射语义的解决方案中最小的解决方案，简称核解)。由于在相关研究中核解已被确定为数据转换的最优解决方案，因此在数据集成过程中如何有效快速计算核解非常重要。现有的计算数据转换核解的方法，通常是通过 chase 方法执行原始的源到目标映射规则生成目标实例，然后应用一些实例选择方法对属于核解的目标实例进行选择。

基于上述研究，已有许多方法对计算核解进行了研究，文献[3]从模式映射的角度出发，通过对映射规则进行优化设计，将映射规则转换为可以执行的脚本来计算核解。文献[4]在给定元数据约束和数据示例的情况下，给出了一种从潜在映射空间中选择最佳映射的方法。但是这些方法通常不适用于数据源规模较大的大型映射场景，可能导致目标数据库中的数据存在大量冗余数据。文献[5]通过弱化目标数据中的约束条件来计算核解，由于该方法只能在特定条件下对有限的数据进行处理，存在一定局限性。关于模式匹配的研究中，大多数研究是关于语义相似性度量方法，文献[6]使用 WordNet 信息引入了一种新的语义相似性度量方法，来处理模式匹配问题。文献[7]同样在语义相似性度量方法的基础上提出了一种在半结构化数据和链接数据之间进行模式匹配的方法。关于数据转换的研究中，文献[8]提出了一种可伸缩实体保存数据转换(SEDEX)方法，该方法利用在模式级别和数据实例级别的信息解决使用不同方法表示模式间对应关系导致模式映射不准确的问题。为了设计准确的模式映射信息，一些方法[9] [10]以交互方式从映射设计器那里获取数据示例，用以设计源模式与目标模式之间的模式映射，并通过将范围较小的多个独立设计的方案映射关联到较大的方案映射构造复杂映射。国内关于数据转换的研究大都集中于 ETL 技术的研究[11] [12] [13]，ETL 技术通常用来描述将数据从来源端经过抽取、转换、加载至目标端的过程。研究过程中的主要问题是解决数据异构性及转换效率提升问题。

传统的模式映射设计中，通常都是通过消除数据值冗余来避免更新的低效性。研究表明，许多冗余的出现都是由于映射规则不完善导致的，因此直接通过重写规则防止在目标中生成冗余数据要比执行不完善的映射规则生成冗余数据后再尝试去删除这些冗余数据更有效。考虑如表 1，表 2 所示场景，表 1

中的源表 A，B，C，D 分别代表来自不同 web 源的源数据库表，表 2 中的目标表 T1，T2 是两个不同的目标数据库表。

Table 1. Source database summary data
表 1. 源数据库概要数据

源表 A 房小二网		源表 B 沈阳房天下网		源表 C 房谱网		源表 D 沈阳楼盘网
P NAME 项目名称	P Adr 项目地址	P NAME 项目名称	P ID 地块编号	P ID 地块编号	P Adr 项目地址	P NAME 项目名称
城南春晓	浑南新区塔南街 501 号	金石小镇	HN-17014	HN-17014	浑南区全运五路与沈营大街交汇处	金石小镇
小石城梦想小镇	浑南区沈营路与全运北路交汇处西行 500 米	小石城梦想小镇	HN0609	HN0609	浑南区沈营路与全运北路交汇处西行 500 米	首创光和城 城南春晓

Table 2. Target database summary data
表 2. 目标数据库概要数据

目标表 T1 (楼盘信息表)		目标表 T2 (地址信息表)	
P NAME 项目名称	P ID 地块编号	P Adr 项目地址	P ID 地块编号
金石小镇	N1	浑南区全运五路与沈营大街交汇处	HN-17014
首创光和城	N2	浑南区沈营路与全运北路交汇处西行 500 米	HN0609
城南春晓	N3	浑南区沈营路与全运北路交汇处西行 500 米	L1
城南春晓	L1	浑南新区塔南街 501 号	L2
小石城梦想小镇	L2		
金石小镇	HN-17014		
小石城梦想小镇	HN0609		

该映射场景初始给定的模式映射如下：

m1. $\forall Pname, Paddr : B(Pname, Paddr) \to \exists I : T1(Pname, I) \wedge T2(I, Paddr)$

m2. $\forall Pname, Pnumber : C(Pname, Pnumber) \to T1(Pname, Pnumber)$

m3. $\forall Pnumber, Paddr : D(Pnumber, Paddr) \to T2(Pnumber, Paddr)$

m4. $\forall Pname : A(Pname) \to \exists N : T1(Pname, N)$

上述映射场景中，每个映射规则的源模式都不相同，通过这些映射规则可以将源模式上的实例转换成目标模式上的实例。这些给定的映射规则基本满足映射场景的初始映射规则，可以将源模式上的数据转换到目标模式中。但依据给定的映射规则生成的目标实例包含冗余实例，即表 2 中灰色背景的部分，这些冗余会影响数据转换的准确性。

基于上述问题，本文提出如下方法：

1) 本文将给定映射规则目标模式上的查询定义为扩展，表示满足映射规则的所有查询公式的集合，它们可以作为一阶查询来捕获目标实例。将矩阵之间的同态关系概念应用到扩展中，通过扩展之间的同态关系，查找到目标实例中满足核解特征的目标实例，扩展间的同态关系称为公式同态。

考虑本文例子中的映射规则 m1，它的基扩展为 $\varepsilon_{b_1}=\exists I:T1(Pname,I)\wedge T2(I,Paddr)$，合并映射规则 m2，m3 可以得到第二个扩展 $\varepsilon''=T1(Pname,Pnumber)\wedge T2(Pnumber,Paddr)$。可以看出，存在由基扩展 ε_{b_1} 到 ε'' 的同态关系，将这种情况称为通过 m2，m3 生成的扩展覆盖了 m1 的扩展，就生成核解而言，第二个扩展比基扩展更好，因为其不含存在量词。

2) 每当为给定映射规则确定了一个比基扩展更合适的扩展时，可以根据以下策略执行映射规则来防止在目标中生成冗余元组。首先通过更合适的扩展查找目标实例，然后使用基扩展只生成那些实际向目标添加一些新内容的目标实例。通过这种方式，我们可以从扩展的角度对核解的特征进行计算，这是本文重写算法的基础。

3) 为了将上述研究转化为实际的重写方法，通过将扩展重写为源模式上的查询公式，扩展的源重写会在源数据库上声明一个条件，以生成目标中与扩展 ε 相关的目标实例。在我们的例子中，扩展的源重写如下：

$$sourceRew(T1(Pname,Pnumber))=IBLBook(t,id)$$

$$sourceRewT1(Pname,Pnumber)\wedge T2(Pnumber,Paddr)=C(Pname,Pnumber),D(Pnumber,Paddr)$$

一旦在源上重写了扩展，可以通过在原始映射规则的前提中添加否定来重写它。每当映射规则 m 有比基扩展更好的扩展 ε 后，通过添加 ε 的源重写的否定来重写他的前提。

4) 通过本文的示例研究发现，并不是所有规则都会导致目标中生成冗余实例。在对原始规则进行重写时，应去除那些不含存在量词并且源到目标对应关系明确的规则，为此我们给出初始规则选择方法，只选择那些会生成冗余元组的规则进行重写。分析给定的映射规则集，为每条规则分配一个冗余度，用来确定生成目标实例冗余程度，以便识别其中的哪条规则可能在目标中生成冗余元组。

5) 最终根据本文方法可以生成以下规则，它们比普通的映射规则更具表现力，允许在前提中添加否定，可以用来表示原始场景的核心模式映射，通过在源实例上执行这些规则能够生成核解，其中 r1、r4 为重写后的规则，r2、r3 为生成目标数据精确度较高的规则，未进行重写：

r1. $\forall x_1,x_2:B(x_1,x_2)\wedge\neg(\exists x_4,x_5:C(x_1,x_4)\wedge D(x_5,x_2)\wedge x_4=x_5)\rightarrow T1(x_1,f_{x_1})\wedge T2(f_{x_2},x_2)$

r2. $C(x_3,x_4)\rightarrow T1(x_3,x_4)$

r3. $D(x_5,x_6)\rightarrow T2(x_5,x_6)$

r4. $\forall x_7,A(x_7)\wedge\neg(\exists x_2:B(x_7,x_2))\wedge\neg(\exists x_4:C(x_7,x_4))\rightarrow T1(x_7,f_{x_7})$

2. 相关概念

定义 1. 一阶规则(FO 规则)

给定源模式 S 和目标模式 T，一阶规则是形式为 $\forall\bar{x}:\varphi(\bar{x})\rightarrow\psi(\bar{x})$ 的映射关系，其中 $\varphi(\bar{x})$ 是 S 上的一阶公式，$\psi(\bar{x})$ 是形式为 $R(t_1\cdots t_n)$ 的原子的合取式，t_i 可以是形式为 $t_i\in\{\bar{x}(x_1,x_2,\cdots,x_n)\}$ 的变量，也可以是 $\bar{x}$ 上的 Skolem 范式[14]。

定义 2. 一阶规则的执行

给定一阶规则 $\forall\bar{x}:\varphi(\bar{x})\rightarrow\psi(\bar{x})$，称 $\psi(\bar{x})$ 为从 $\varphi(\bar{x})$ 获得的 S 上的一阶顺序查询，将 $\bar{x}$ 视为自由变量。用 $Q_\varphi(I)$ 表示元组 $\bar{c}\in dom(I)^{|\bar{x}|}$，这样 $\bar{c}$ 就是在 S 上的实例 I，对于查询 Q_φ 的结果。给定 $\bar{c}\in Q_\varphi(I)$，然后用 $\psi(\bar{c})$ 表示从 ψ 获得的原子集合，用相应的 $c_i\in\bar{c}$ 替换每个变量 $x_i\in\bar{x}$，并用相应的不确定量替换每个 Skolem 项[14]。

定义 3. 核心模式映射

给定映射场景 $M=\left(S,T,\Sigma_{ST}\right)$，如果对于任意源实例 I，目标实例 $R(I)$ 是 M 在 I 上的核解，一组 FO 规则 R 被称为 M 的核心模式映射。

定义 4. 变量

给定公式 $\varphi^l\left(\bar{x},\bar{y}\right)$ 中的原子 $R^l\left(A_1:v_1,\cdots,A_k:v_k\right)$，其中变量由 $R^l.A_j:v_i$ 表示。如果 $v_i\in\bar{x}$，则 $\varphi^l\left(\bar{x},\bar{y}\right)$ 中的变量 $R^l.A_j:v_i$ 是全称量词；如果 $v_i\in\bar{y}$，它是存在量词。

在下文中，用 $occ\left(\varphi^l\left(\bar{x},\bar{y}\right)\right)$ 表示所有在 $\varphi^l\left(\bar{x},\bar{y}\right)$ 中的变量；$u-occ\left(\varphi^l\left(\bar{x},\bar{y}\right)\right)$，$e-occ\left(\varphi^l\left(\bar{x},\bar{y}\right)\right)$ 分别表示全称变量和存在变量。同样，$occ(v)$，$u-occ(v)$，$e-occ(v)$ 将表示给定变量 v 的所有(通用的，存在的)变量取值集合。

定义 5. 公式同态

给定两个合取范式：$\varphi^l\left(\bar{x},\bar{y}\right)$ 和 $\varphi'^{l}\left(\bar{x}',\bar{y}'\right)$，公式同态是一个从集合 $occ\left(\varphi^l\left(\bar{x},\bar{y}\right)\right)$ 到 $occ\left(\varphi'^{l}\left(\bar{x}',\bar{y}'\right)\right)$ 的映射 h^f 。

i) h^f 将 $\varphi^l\left(\bar{x},\bar{y}\right)$ 中的全称量词映射为 $\varphi'^{l}\left(\bar{x}',\bar{y}'\right)$ 中的全称量词；

ii) 对于每个原子 $R^l\left(A_1:v_1,\cdots,A_k:v_k\right)\in\varphi^l\left(\bar{x},\bar{y}\right)$，在集合 $\varphi'^{l}\left(\bar{x}',\bar{y}'\right)$ 上存在 $R\left(h^f\left(R^l\left(A_1:v_1,\cdots,A_k:v_k\right)\in\varphi'^{l}\left(\bar{x}',\bar{y}'\right)\right)\right)$ 与之对应；

iii) 对于存在变量 $y\in\bar{y}$，变量 $R_n^l.A_m:y$，$R_i^l.A_j:y$ 成对出现，在这种情况下，$h^f\left(R_n^l.A_m:y\right)$ 和 $h^f\left(R_i^l.A_j:y\right)$ 都是常量，或者是在 $y'\in\bar{y}'$ 中相同的存在变量。

如果一个公式同态 h^f 把 $\varphi^l\left(\bar{x},\bar{y}\right)$ 中的不同的原子映射到 $\varphi'^{l}\left(\bar{x}',\bar{y}'\right)$ 的不同的原子中，则公式同态 h^f 是单射的。如果 $\varphi'^{l}\left(\bar{x}',\bar{y}'\right)$ 中的每个原子都是依据 h^f 在 $\varphi^l\left(\bar{x},\bar{y}\right)$ 中的某个原子的图像，那么 h^f 是满射的。

对于关系 $R(A,B,C)$ 和 $T\left(A,B,C\right)$ 上的两个查询公式：

$$\varphi^l=R^1\left(x_1,x_2,Y_1\right)\wedge T^2\left(x_3,x_1,Y_1\right)\text{和}\varphi'^{l}=R^3\left(x_4',x_5',x_6'\right)\wedge T^4\left(x_9',x_7',x_8'\right)$$

在以下变量之间的映射中，存在 φ^l 到 φ'^{l} 的公式同态 h^f 。

$$h^f\left(R^1.A:x_1\right)\to R^3.A:x_4',\quad h^f\left(R^1.B:x_2\right)\to R^3.B:x_5'$$

$$h^f\left(R^1.C:Y_1\right)\to R^3.C:x_6',\quad h^f\left(T^2.A:x_3\right)\to T^4.A:x_9'$$

$$h^f\left(T^2.B:x_1\right)\to T^4.B:x_7',\quad h^f\left(T^2.C:Y_1\right)\to T^4.C:x_8'$$

可以看出，由于公式同态的影响，它们可以将左侧的相同变量与右侧不同变量相关联。在本文中，将通过 $A_j:h^f_{R^l.A_j}\left(v_i\right)$ 来引用变量 $h^f\left(R_l.A_j:v_i\right)$。$h^f_{R^l.A_j}\left(v_i\right)$ 是与 v_i 中的 $R^l.A_j$ 相关联的变量。考虑上面 h^f 的例子，φ^l 中出现两次的存在变量 x_1 被映射到 φ'^{l} 中不同的通用变量的位置，实际上，$h^f_{R^1.A}\left(x_1\right)=x_4'$，而 $h^f_{T^2.B}\left(x_1\right)=x_7'$。

定义 6. 公式同态的分类

给定两个合取公式 $\varphi^l\left(\bar{x},\bar{y}\right)$ 和 $\varphi'^{l}\left(\bar{x}',\bar{y}'\right)$，和一个从 $occ\left(\varphi^l\left(\bar{x},\bar{y}\right)\right)$ 到 $occ\left(\varphi'^{l}\left(\bar{x}',\bar{y}'\right)\right)$ 的公式同态 h^f 。

i) 如果 h^f 是满射的，则公式同态 h^f 是更紧凑的。可以是 $\left|\varphi'^{l}\left(\bar{x}',\bar{y}'\right)\right|<\left|\varphi^l\left(\bar{x},\bar{y}\right)\right|$ 或 $\left|\bar{y}'\right|<\left|\bar{y}\right|$ 的情况。即要么 $\varphi'^{l}\left(\bar{x}',\bar{y}'\right)$ 比 $\varphi^l\left(\bar{x},\bar{y}\right)$ 小，要么 $\varphi'^{l}\left(\bar{x}',\bar{y}'\right)$ 包括较少的存在变量；

ii) 如果 h^f 是单射的而不是满射的，则 h^f 被认为是更适当的，即在 $\varphi'^{l}\left(\bar{x}',\bar{y}'\right)$ 中至少有一个原子不是 $\varphi^l\left(\bar{x},\bar{y}\right)$ 的原子的图像。

讨论“公式同态”和“事实(公式的所有实例的集合)之间对的应同态”的关系是非常重要的。本文研究公式同态，以便检测作为公式事实之间可能的同态。然而，公式同态并不能保证实际同态在事实之间

产生，公式之间的同态映射可以将通用变量的值映射为其他通用变量的值。在实例化公式时，这些变量不一定接收相同的值，公式实例之间的同态可能实现也可能不实现，这都取决于通用变量所假设的值。

考虑关于 $\varphi^{l}=R^{1}\left(x_1,x_2,Y_1\right)\wedge T^{2}\left(x_3,x_1,Y_1\right)$ 和 $\varphi'^{l}=R^{3}\left(x_4',x_5',x_6'\right)\wedge T^{4}\left(x_9',x_7',x_8'\right)$ 的公式同态 h^{f}。包含两个公式的实例：$W=\left\{R\left(1,2,N_0\right),T\left(3,1,N_0\right)\right\}$ 和 $W'=\left\{R\left(1,2,4\right),T\left(3,1,4\right)\right\}$。给定变量值的赋值，可以看出，$w'$ 的事实集实际上比 w 的事实集更紧凑。如果改变赋值，通常不会这样。

现在考虑一下：$W''=\left\{R\left(1,2,N_0\right),T\left(3,1,N_0\right)\right\}$ 和 $W'''=\left\{R\left(1,4,6\right),T\left(3,5,7\right)\right\}$。$w''$ 中没有 w''' 的同态，造成这种情况的原因有两个。i) 首先，公式同态将 x_2 映射为 x_5'。通过这样做，公式同态对变量 x_2，x_5' 的值进行限制：为了实现公式实例之间的同态，两个变量必须接收相同的值；ii) 其次，h^{f} 将两次出现的 N_0 映射到 $x_6'x_8'$；这意味着为了实现同态，x_6' 的值应等于 x_8' 的值。

给定一个公式同态 h^{f}，可以在与 h^{f} 有关的通用变量之间引入几组等式：

$INTERSECT_{h^f}$ 表示 $\varphi^{l}\left(\overline{x},\overline{y}\right)$ 和 $\varphi'^{l}\left(\overline{x}',\overline{y}'\right)$ 之间全称量词的一组等价集(等式集合)，这组等式必须成立，才能实现这两个公式实例之间的同态。

$$INTERSECT_{h^f}\left(\overline{x},\overline{x}'\right)=\left\{x_i=x_j'\mid h^{f}\left(R.A:x_i\right)=R.A:x_j',x_i\in\overline{x},x_j'\in\overline{y}\right\}$$

$JOINS_{h^f}$ 表示 $\varphi'^{l}\left(\overline{x}',\overline{y}'\right)$ 的全称量词之间的等价集，其具体的值是 $\varphi^{l}\left(\overline{x},\overline{y}\right)$ 中相同存在量词的值的图像。

$$JOINS_{h^f}\left(\overline{x}\right)=\left\{x_h'=x_l'\mid x_h'=h^{f}_{R_i.A_j}\left(y_k\right),x_l'=h^{f}_{R_n.A_m}\left(y_k\right),y_k\in\overline{y}\right\}$$

直观地说，只有满足 $EQUAL_{h^f}\left(\overline{x},\overline{x}'\right)$ 的赋值才能实现公式同态。

$$EQUAL_{h^f}\left(\overline{x},\overline{x}'\right)=INTERSECT_{h^f}\left(\overline{x},\overline{x}'\right)\cup JOINS_{h^f}\left(\overline{x}'\right)$$

3. 模式映射重写

3.1. 扩展

给定一个映射场景 $M=\left(S,T,\Sigma_{ST}\right)$，由 $R=\bigcup_i\psi_i\left(\overline{x}_i,\overline{y}_i\right)$ 表示在 Σ_{st} 中的所有映射规则结论的集合，由 R_k^{pow} 表示在 R 中的所有数量≤k 的多重原子集合；每当多重集合中出现相同原子的多个副本时，本文假设它们已被正确重命名，以避免变量的冲突。给定映射场景 $M=\left(S,T,\Sigma_{ST}\right)$，本文的目标是在一组表示 M 的核心模式映射的一阶逻辑规则下重写给定的映射规则，从中生成一个 SQL 脚本，直接生成核心目标实例。为了执行重写，将依赖于核解的概念。在本节中，我们将介绍映射规则结论中基于扩展的核解概念，并在下一节中使用它来执行重写，使用扩展作为研究映射规则结论之间可能存在冗余的一种方法。

定义 7. 映射规则中的扩展

给定映射场景 M 和映射规则集 Σ_{ST}，在 Σ_{ST} 中有映射规则 $tgd.m:\phi\left(\overline{x}_2\right)\to\exists\overline{y}_2\left(\psi^{l}\left(\overline{x}_2,\overline{y}_2\right)\right)$，$m$ 的扩展集合用 $expansions_M\left(m\right)$ 表示，$expansions_M\left(m\right)$ 是一组包含存在量词的逻辑查询公式：$\varepsilon=\chi^{l}\left(\overline{x}_1,\overline{y}_1\right)\wedge\exists\overline{x}_2,\overline{y}_2:\left(\psi^{l}\left(\overline{x}_2,\overline{y}_2\right)\wedge EQUAL_{h_c^f}\left(\overline{x}_1,\overline{x}_2\right)\right)$，其中 $\chi^{l}\left(\overline{x}_1,\overline{y}_1\right)$ 是 R_k^{pow} 中带标记的原子(k 是 $\psi^{l}\left(\overline{x}_2,\overline{y}_2\right)$ 的大小)，且存在满射 h^{f}，$h_c^{f}:\psi\left(\overline{x}_2,\overline{y}_2\right)\to\chi\left(\overline{x}_1,\overline{y}_1\right)$。

在接下来的文章中，假设扩展中 $\overline{x}_1\cap\overline{x}_2=\varnothing$，$\overline{y}_1\cap\overline{y}_2=\varnothing$，即 $\overline{x}_1,\overline{x}_2,\overline{y}_1,y_2$ 是不相交的。注意扩展 ε 也可以被看作是一个含有自由变量 $\overline{x}_1,\overline{y}_1$ 的查询 $\varepsilon\left(\overline{x}_1,\overline{y}_1\right)$。接下来会证明，在通解 $J\in USol_M\left(I\right)$ 上进行这样的查询，结果恰好会返回目标实例集合 $W^{\langle I,J\rangle}$ 中的一组目标实例。

给定实例 J，并给 $\overline{x}_1,\overline{y}_1$ 赋值 a_1，如果满足以下条件，则称 $J\vert=a_1\left(\varepsilon\left(\overline{x}_1,\overline{y}_1\right)\right)$：

$$J\vert=a_1\left(\chi^{l}\left(\overline{x}_1,\overline{y}_1\right)\right)$$

a_1, a_2 是 $EQUAL_{h^f}\left(a_1\left(\overline{x}_1\right), a_2\left(\overline{x}_2\right)\right)$ 的形式。

由于映射规则结论的扩展数量随着结论之间连接的数量而增加，并且它通常是输入映射规则的大小的指数。将 $expansions_M\left(m\right)$ 称为 Σ_{ST} 中映射规则的所有扩展集。

3.1.1. 扩展与核解

根据参考文献[15]我们可以知道，在目标数据库存储的目标实例中，为了生成核解，需要选择最大的实例，具体步骤是首先选择包含属性更多的实例，再此基础上选择包含信息性更大的实例，最后将能够满足 $J_0 = \cup SelectBestBlock\left(Single\left(Full\left(W^{\langle I,J\rangle}\right)\right)\right)$ 关系的实例称之为核解。我们在本节引入核解的概念，基于扩展给出核解的相似概念，但是这个核解并不是依赖于实例的而是基于本文所说的扩展生成的。

3.1.2. 扩展的分类

在本文第 2 节中介绍的公式同态可以用来确定什么时候扩展能够产生一个比其他块包含更多属性或更具信息性的实例，由此引入了一个更紧凑的和更具信息性的扩展的并行定义。

定义 8. 更紧凑和更具信息性的扩展

给定由具体实例组成的扩展：

$$\varepsilon = \chi^l\left(\overline{x}_1, \overline{y}_1\right) \wedge \exists \overline{x}_2, \overline{y}_2 : \left(\psi^l\left(\overline{x}_2, \overline{y}_2\right) \wedge EQUAL_{h_c^f}\left(\overline{x}_1, \overline{x}_2\right)\right)$$

$$\varepsilon' = \chi'^l\left(\overline{x}'_1, \overline{y}'_1\right) \wedge \exists \overline{x}'_2, \overline{y}'_2 : \left(\psi'^l\left(\overline{x}'_2, \overline{y}'_2\right) \wedge EQUAL_{h_c^f}\left(\overline{x}'_1, \overline{x}'_2\right)\right)$$

如果存在一个更紧凑的同态 $h_c^f : \chi^l\left(\overline{x}_1, \overline{y}_1\right) \to \chi'^l\left(\overline{x}'_1, \overline{y}'_1\right)$，则称 ε' 比 ε 更紧凑，用 $\varepsilon \prec \varepsilon'$ 表示；

如果存在更恰当的同态 $h_p^f : \chi^l\left(\overline{x}_1, \overline{y}_1\right) \to \chi'^l\left(\overline{x}'_1, \overline{y}'_1\right)$，称 ε' 比 ε 更具有信息性，用 $\varepsilon < \varepsilon'$ 表示；

当根据公式同态 h^f 使得 ε' 比 ε 更紧凑(或更具信息性)时，可以这样写 $\varepsilon \prec_{h^f} \varepsilon'$ $\left(\varepsilon <_{h^f} \varepsilon'\right)$。

3.1.3. 扩展的选择

根据上述方法，给定一个扩展 ε，我们可以基于扩展 ε 生成一个新的查询，叫做 $mostComp\left(\varepsilon\right)$。主要通过向 ε 添加 ε' 的否定， ε' 比 ε 更紧凑。有如下等式：

定义 9. 更紧凑的扩展： $mostComp\left(\varepsilon\right)$

给定一个映射场景 M，以及它的一组扩展 $expansions\left(M\right)$，其中的一个扩展如下：

$$\varepsilon = \chi^l\left(\overline{x}_1, \overline{y}_1\right) \wedge \exists \overline{x}_2, \overline{y}_2 : \left(\psi^l\left(\overline{x}_2, \overline{y}_2\right) \wedge EQUAL_{h_c^f}\left(\overline{x}_1, \overline{x}_2\right)\right)$$

$mostComp\left(\varepsilon\right)$ 通过如下方式获得：

1) 初始化 $mostComp\left(\varepsilon\right) = \varepsilon$；

2) 对于 $expansions\left(M\right)$ 中的任意 $\varepsilon' = \chi'^l\left(\overline{x}'_1, \overline{y}'_1\right) \wedge \exists \overline{x}'_2, \overline{y}'_2 : \left(\psi'^l\left(\overline{x}'_2, \overline{y}'_2\right) \wedge EQUAL_{h_c^f}\left(\overline{x}'_1, \overline{x}'_2\right)\right)$ 和任意形如 $\varepsilon \prec_{h^f} \varepsilon'$ 的公式同态 h^f。即 ε' 根据 h^f 变得比 ε 更紧凑，向 $mostComp\left(\varepsilon\right)$ 添加公式 $\wedge \neg \exists \overline{x}'_1, \overline{y}'_1 : \left(\varepsilon' \wedge EQUAL_{h_c^f}\left(\overline{x}_1, \overline{x}'_1\right)\right)$。

本文将用 $mostComp\left(M\right)$ 来表示形式为 $mostComp\left(\varepsilon\right)$ 的扩展所有重写的集合，其中 $\varepsilon \in expansions\left(M\right)$。

在第一次重写之后，与通过实例生成核解的方法所述的策略相一致，我们在其他扩展中寻找有利于在目标中生成更具信息性的目标实例的扩展，并且进一步重写 $mostComp\left(\varepsilon\right)$。在这个过程中，产生成了一个在 $mostComp\left(\varepsilon\right)$ 基础上的公式，叫做 $mostInf\left(\varepsilon\right)$，如下所示：

定义 10. 更具信息性的扩展： $mostInf\left(\varepsilon\right)$

给定一个映射场景 M，以及它的一组扩展 $expansions(M)$，其中的一个扩展如下：

$$\varepsilon = \chi^{l}(\overline{x}_1, \overline{y}_1) \wedge \exists \overline{x}_2, \overline{y}_2 : \left(\psi^{l}(\overline{x}_2, \overline{y}_2) \wedge EQUAL_{h_c^f}(\overline{x}_1, \overline{x}_2)\right)$$

$mostInf(\varepsilon)$ 的具体操作如下：

1) 初始化 $mostInf(\varepsilon) = mostComp(\varepsilon)$

2) 对于 $expansions(M)$ 中的任意扩展

$\varepsilon' = \chi'^{l}(\overline{x}'_1, \overline{y}'_1) \wedge \exists \overline{x}'_2, \overline{y}'_2 : \left(\psi'^{l}(\overline{x}'_2, \overline{y}'_2) \wedge EQUAL_{h_c^f}(\overline{x}'_1, \overline{x}'_2)\right)$ 和任意形如 $\varepsilon <_{h^f} \varepsilon'$ 的公式同态 h^f，即 ε' 由 h^f 变得比 ε 更具信息性，向 $mostInf(\varepsilon)$ 添加公式

$$\wedge \neg \exists \overline{x}'_1, \overline{y}'_1 : \left(mostComp(\varepsilon') \wedge EQUAL_{h_c^f}(\overline{x}_1, \overline{x}'_1)\right)。$$

我们用 $mostInf(M)$ 表示 $mostInf(\varepsilon)$ 形式的所有重写集。

总之，为了选择最大目标实例，本文考虑映射规则 tgd 中的每个扩展 ε：

a) 首先通过添加所有的扩展 ε_i 否定，将 ε 重写成一个新形式：$mostComp(\varepsilon)$，ε_i 比 ε 更紧凑；本文期望用这些新的公式选择在与 ε 相关的目标实例中更紧凑的；

b) 然后通过添加 $mostComp(\varepsilon_j)$ 的否定，进一步将 $mostComp(\varepsilon)$ 重写为一个新的公式 $mostInf(\varepsilon)$，扩展 ε_j 比 ε 更具信息性。

与扩展类似，扩展的重写也可以被视为查询。给定扩展 ε，以及相关的查询 $\varepsilon(\overline{x}_1, \overline{y}_1)$，$mostComp(\varepsilon)$ 和 $mostInf(\varepsilon)$ 都可以看作是带有自由变量 $\overline{x}_1, \overline{y}_1$ 的查询。本文编写这些查询如下：$mostComp(\varepsilon)(\overline{x}_1, \overline{y}_1)$ 和 $mostInf(\varepsilon)(\overline{x}_1, \overline{y}_1)$。为了简化符号，本文将省略显式引用变量，仅用 $mostComp(\varepsilon)$ 和 $mostInf(\varepsilon)$ 来代表查询。

3.2. 规则选择

在前文中，通过对初始映射规则结论进行分析，给出了扩展的概念，进而我们希望通过扩展对初始模式映射规则集进行重写以让他可以直接生成核解。但是分析发现，不是所有的规则都会在目标中生成冗余实例。若对初始规则集中的所有规则都进行重写，会导致运行时间过长，效率低等问题。由于有些规则已经满足核心模式映射条件，不需要进行重写就可以直接生成满足核解的目标实例，我们希望对规则集进行选择，只对不满足核心模式映射条件的规则进行重写。

3.2.1. 规则冗余程度

为了识别不同规则生成冗余的大小程度，必须对规则源模式与目标模式之间的属性相似度进行计算。针对本文研究数据特点，由于数据来自不同的 web 数据源，在计算属性相似度时选用 Jaccard 相似系数方法来计算源和目标模式属性相似度。

定义 11. Jaccard 相似系数

给定两个集合 A、B，Jaccard 相似系数定义为 A 与 B 交集的大小与 A 与 B 并集的大小的比值，定义如下：

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

当集合 A，B 都为空时，$J(A,B)$ 定义为 1。

本文称给定规则源模式与目标模式之间的相似系数大于 0.7 时，该规则生成的目标实例冗余度较低，不需要进行重写。当相似系数低于 0.7 时，应进行规则分解，查看规则中全称量词的数量来进一步确定

该规则是否会生成冗余度较高的目标实例，是否需要重写。

3.2.2. 规则分解

给定映射场景M以及它的初始规则集，本文将根据文献[16]中研究的策略来对冗余度较高的给定规则进行分解，具体方法如下。

分析表级之间映射关系的构成，考虑源和目标模式下包含如下对应关系：

1) 1:1 映射关系

1:1 映射关系是指 web 数据源模式与目标数据模式中的表结构是一致的，即标准数据库中的某些表在源数据库中有唯一的一个数据与之对应。对于映射关系为 1:1 的映射规则，若其目标模式中不含存在量词，则不要进行重写，直接作为核模式映射输出。若其目标模式中包含存在量词，这些规则生成的数据中必定包含较多的变量，说明这些规则是冗余的，直接进入待重写的规则库。

2) 1:*N* 映射表关系

1:*N* 映射关系是指 web 数据源模式与目标数据模式的表结构并非是一致的，即源数据库中原始数据的一个表映射到目标数据库中的多个表。首先将其分解为 1:1 类型的映射，然后根据其包含存在量词的数量来确定其冗余程度的大小，判断是否需要对其进行重写。

3) *N*:*M* 映射表关系

N:*M* 映射关系是指 web 数据源模式与目标数据模式的表结构并非是一致的，即源数据库中原始数据的多个表映射到目标数据库中的多个表。当规则中的映射关系为 *M*:*N* 时，可以首先将其转换成 *M* 个 1:*N* 的映射关系；然后再将 1:*N* 的映射处理为 *N* 个 1:1 之间的映射；最后对 *M* * *N* 个 1:1 的映射关系进行判断处理即可。在其中若不包含存在量词，则作为核心模式映射输出；若其中包含存在量词，则其冗余度较大，需要按照本文方法进行重写，转换成核心模式映射来计算核解。

3.3. 重写方法

3.3.1. 源重写

扩展作为目标模式上的公式，在本节中表明可以根据源模式上的关系来重写扩展，我们将其称为扩展的源重写。虽然扩展是一个用于在目标模式中选择原子的查询，但它的源重写说明了这些原子存在的“前提条件”。通过引入标签技术[2]，在源上重写扩展的策略得以实现。事实上，扩展是从映射规则结论中得到的标记原子的合取范式。对于每一个原子，都用一个前提联系起来(前提就是相应的规则的左侧)。通过连接其所有原子的前提，能够获得扩展的源重写。注意到标签系统在这一步中起着核心作用：通过查看每个扩展的原子的标签，可以立即知道它的出自哪一个映射规则，因此也就知道了它的前提。

定义 12. 映射规则的前提

给定映射规则 $tgd.m\forall\bar{x}:\phi(\bar{x})\to\exists\bar{y}\left(\psi^{l}(\bar{x},\bar{y})\right)$ 和原子 $R^{l}(\bar{x}_i,\bar{y}_i)\in\psi^{l}(\bar{x},\bar{y})$， $R^{l}(\bar{x}_i,\bar{y}_i)\in\psi^{l}(\bar{x},\bar{y})$ 的前提 $premise\left(R^{l}(\bar{x}_i,\bar{y}_i)\right)$ 实际上是由映射规则的左侧公式 $\phi(\bar{x})$ 构成的。

给定一个查询公式 $\chi^{l}(\bar{x}_1,\bar{y}_1)$，其具体前提由如下公式表示：

$$premise\left(\chi^{l}(\bar{x}_1,\bar{y}_1)\right)=\wedge\left\{premise\left(R_i^{l}(\bar{x}_i,\bar{y}_i)\right)\mid R_i^{l}(\bar{x}_i,\bar{y}_i)\in\chi^{l}(\bar{x}_1,\bar{y}_1)\right\}$$

定义 13. 源重写

给定映射规则 $m:\phi(\bar{x}_2)\to\exists\bar{y}_2\left(\psi^{l}(\bar{x}_2,\bar{y}_2)\right)$ 和扩展集合 $expansions_M(m)$ 中的扩展 $\varepsilon=\chi^{l}(\bar{x}_1,\bar{y}_1)\wedge\exists\bar{x}_2,\bar{y}_2:\left(\psi^{l}(\bar{x}_2,\bar{y}_2)\wedge EQUAL_{h_c^f}(\bar{x}_1,\bar{x}_2)\right)$，它的源重写 $SourceRew(\varepsilon)$ 是下面的公式：

$$SourceRew(\varepsilon) = premise\left(\chi^{l}(\bar{x}_1, \bar{y}_1)\right) \wedge \exists \bar{x}_2 : \left(\phi(\bar{x}_2) \wedge EQUAL_{h_c^f}(\bar{x}_1, \bar{x}_2)\right)$$

注意到，虽然 tgd 的扩展 ε 可以看作是一个查询 $\varepsilon(\bar{x}_1, \bar{y}_1)$，$\varepsilon(\bar{x}_1, \bar{y}_1)$ 包括映射规则 m 的全称量词和存在量词，它的源重写 $SourceRew(\varepsilon)$，也是一个查询 $SourceRew(\varepsilon)(\bar{x}_1)$，但其中所有自由变量都是映射规则 m 的全称量词。

3.3.2. 源重写的否定

给定扩展 ε，它的源重写 $SourceRew(\varepsilon)$ 说明了产生其所有实例的前提条件。本文只想选择关于信息量最大的，最具代表性的实例。因此，需要生成新的表达式，分别为与扩展相关的最紧凑和更具信息性的见证块集提供前提条件。

为了进行上述操作，给定扩展 ε，引入公式 $SourceRew(mostComp(\varepsilon))$ 和 $SourceRew(mostInf(\varepsilon))$，类似于 $mostComp(\varepsilon)$ 和 $mostInf(\varepsilon)$，但是在源上重写了。为了生成 $SourceRew(mostComp(\varepsilon))$，每当扩展 ε' 比 ε 更紧凑时，在 ε 的源重写中加入 $SourceRew(\varepsilon')$ 的否定。

定义 14. 重写源上的 *mostComp*()

给定场景 M，以及 M 上的一组扩展 $expansions(M)$，他们的重写是 $mostComp(\varepsilon)$，对于扩展集 $expansions(M)$ 中的每个 ε，其公式 $SourceRew(mostComp(\varepsilon))$ 生成的方式如下：

i) 首先初始化

$$SourceRew(mostComp(\varepsilon)) = SourceRew(\varepsilon)$$

ii) 然后对于扩展集 $expansions(M)$ 中的任何扩展 ε'，若 ε' 是比 ε 更紧凑的，称 h_c^f 为从 $\chi^{l}(\bar{x}_1, \bar{y}_1)$ 到 $\chi'^{l}(\bar{x}'_1, \bar{y}'_1)$ 的更紧凑的同态；向 $SourceRew(mostComp(\varepsilon))$ 添加一个公式：

$$\wedge \neg \exists \bar{x}' : \left(SourceRew(\varepsilon') \wedge EQUAL_{h_c^f}(\bar{x}_1, \bar{x}'_1)\right)$$

定义 15. 重写源上的 *mostInf*()

给定场景 M，以及 M 上的一组扩展 $expansions(M)$，对于扩展集 $expansions(M)$ 中的每个 ε，其公式 $SourceRew(mostInf(\varepsilon))$ 生成的方式如下：

i) 初始化

$$SourceRew(mostInf(\varepsilon)) = SourceRew(mostComp(\varepsilon))$$

ii) 然后对于扩展集 $expansions(M)$ 中的任何扩展 ε'，若 ε' 是比 ε 更具信息性的，称 h_c^f 为从 $\chi^{l}(\bar{x}_1, \bar{y}_1)$ 到 $\chi'^{l}(\bar{x}'_1, \bar{y}'_1)$ 的更恰当的同态；向 $SourceRew(mostInf(\varepsilon))$ 添加一个公式：

$$\wedge \neg \exists \bar{x}'_1 : \left(SourceRew(mostComp(\varepsilon')) \wedge EQUAL_{h_c^f}(\bar{x}_1, \bar{x}'_1)\right)$$

3.3.3. 重写算法

现在准备介绍本文的最终重写算法。主要介绍一个映射场景 M 中扩展规则的概念，由于希望规则被规范化。因此，使用公式 $SourceRew(mostInf(\varepsilon))$ 为每个规范化的扩展 ε 都建立一个规则，其中将只有全称量词出现作为前提条件。在 ε 不是规范化的情况下，生成一系列规范化规则，由 $normalize(\varepsilon)$ 表示，$normalize(\varepsilon)$ 中每个规范化分量对应一个规则。

定义 16. 扩展规则

对于扩展集 $expansions(M)$ 中的每个扩展 ε，都生成一组扩展规则 $expansionRule(\varepsilon)$，其形式如下：

$$\forall \bar{x}_1 : SourceRew(mostInf(\varepsilon))(\bar{x}_1) \rightarrow \chi_i(\bar{x}_1, \bar{y}_1)$$

其中 $\chi_i\left(\overline{x}_1,\overline{y}_1\right)$ 是 $normalize(\varepsilon)$ 的规范化分量。映射场景 M 的扩展规则集如下：

$$\Sigma_M^e=\left\{expansionRule(\varepsilon)\mid\varepsilon\in expansions(M)\right\}$$

正如前面几节所讨论的，一个简单的优化是由仅为原子集为核解的扩展生成的扩展规则组成。

为了处理非标准化块，我们在规则结论中寻找适当的同态：我们将公式同态的概念扩展到 Skolem 公式，将 Skolem 项考虑为存在量词；然后进一步的进行重写，如下所示：

定义 17. 最终重写 *finalRew*()

对于每条属于 Σ_M^e 的规则 r，规则 $finalRew(r)$ 通过如下方式获得：

将在规则选择过程中冗余度较高的规则 r 称为分解规则。

若 r 不是分解规则， $finalRew(r)=r$ 。

若 r 是形式为 $r.\phi(x)\to\psi(x)$ 的分解规则：

i) 首先初始化 $finalRew(r)=r$ ；

ii) 对于 Σ_M^e 中的任意规则 $r'.\phi'(x')\to\psi'(x')$，根据同态 h_i^f 说明 $\psi'(x')$ 是比 $\psi(x)$ 更具信息性的，向 $finalRew(r)$ 增加一个前提公式 $\wedge\neg\exists(\overline{x}'):\left(\phi'(x')\wedge EQUAL_{h_i^f}(\overline{x},\overline{x}')\right)$。

最后构成了一组新的一阶逻辑规则，具体表示如下：

$$\Sigma_M^c=\left\{finalRew(r)\mid r\in\Sigma_M^e\right\}$$

4. 实验结果

4.1. 数据集

本文采用的数据集是从中国土地网、链家网等房产信息网站爬取的 500 多万条数据，经过垃圾清理、无意义信息删除两个预处理过程，数据信息属性如表 3 所示。

Table 3. Real estate information data sheet attribute description
表 3. 房地产信息数据表属性说明

序号	列名	数据类型	说明
1	PId	Int	编号(主键)
2	Pname	Varchar	楼盘名称
3	Paddr	Varchar	楼盘地址

为了评估本文提出的模式映射优化方法的可行性和有效性，我们从数据集中筛选出部分高质量房产信息数据构建实验数据集，现实世界中的真实数据很难全面的展现所有问题，因此，采用上述数据集构造了人工数据集，表 4 显示了构建数据集的关键特征。针对个本文构建数据集模式及目标数据模式即 S 和 T 构建模式映射，选取其中 12 个作为基本映射集，这些模式映射可以满足源和目标模式的基本要求。

Table 4. Property description of the Weibo data table
表 4. 实验用评测数据集

序号	Set1	Set2	Set3
数据集大小(TB)	6.32	4.17	1.98
平均实例大小(KB)	132.6	71.0	202.3
数据冗余程度(%)	26.14	33.71	23.26

4.2. 实验设置

本文将按照以下两种策略来衡量本文算法的有效性以及与同类工作相比的优越性。首先比较不同数据集下的信息准确率，信息准确率是指映射规则能描述完整的源模式及目标模式信息，包括文档结构信息以及语义约束信息等。映射规则中的每一个分量都不可再分，任意两个属性不能完全相同。映射规则的完整性代表着规则的完善程度，也标志着根据其生成的目标实例存在少量冗余信息。在本文节中，信息准确率是根据本文方法重写规则后生成的满足核解特征的实例数量占总目标数据库中存储实例数量的比例，如下列公式所示：

$$\text{Precision} = \frac{\text{满足核解特征的实例数}}{\text{目标数据库中存储实例总数}} \times 100\%$$

接下来本文将根据优化后的模式映射生成的数据中冗余缩小程度来评估本文方法的可行性，冗余缩小程度即结果生成的数据包含冗余数量占总数据的百分比与之前数据集中重复率大小的比值，如下列公式所示：

$$\text{Reduction of repetition} = \frac{\text{最终生成数据冗余度}}{\text{初始冗余程度}} \times 100\%$$

4.3. 实验结果

图 1 比较了本文方法 SMO 与其他两种方法 TKM [17]，SRMIS [18]的比较结果。如图所示，在三个数据集的测试中，本文方法都能保持较高性能，信息准确率始终高于 80%，TKM 方法性能表现次之，这主要是由于其固有的缺点，即它的基于语义逻辑的映射选择策略，这可能导致不同应用场景的模式映射结果存在巨大差异。SRMIS 的信息准确率仅达到 70%左右，在本文数据集中的表现结果最差。

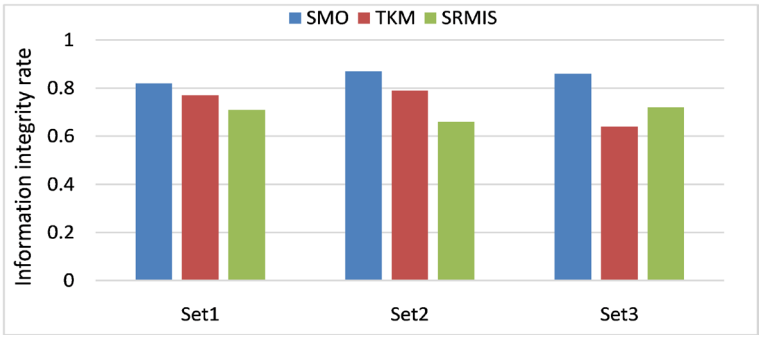

Figure 1. Information integrity comparison
图 1. 信息完整性对比

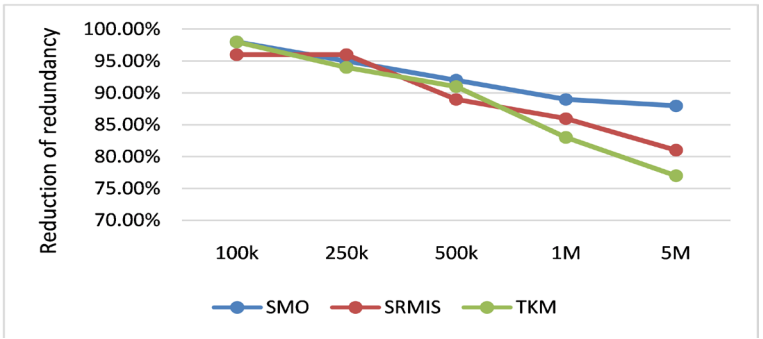

Figure 2. Reduction of redundancy
图 2. 冗余缩小程度

图 2 显示了随着数据规模的增大，本文方法与其他两种方法的冗余程度。如图所示，当数据数量为100 k时，当源实例数据量小于500 k时，三种方法都具有较好的处理性能。当源实例数据量增加到1 M时，SRMIS，TKM两种算法对冗余的缩小程度明显下降。当源实例数据量增加到5 M时，本文中提出SMO方法可以将冗余缩小程度维持在85%以上，其他两种方法在处理数据量较大的数据集时，表现不佳，但很明显本文方法在降低大型数据集冗余程度时性能更好，效率更高。

5. 结论

本文首先分析了传统的数据转换核解计算方法的不足之处，继而对上述现有的核解计算方法提出改进，并针对模式层数据转换问题提出了模式层数据转换映射重写方法，进而对本文方法的查询效果做了理论分析，在真实数据集中验证了本文方法的有效性，证明本文方法在很大程度上减少了冗余数据数量，降低了转换成本。本文方法需要在已有研究的基础上，进一步研究包含连接情况的映射处理的问题，分析映射关系之间的依赖性和数据的分布特性，优化转换效率，以进一步降低转换成本，提高算法的转换效率。

参考文献

[1] Fagin, R., Kolaitis, P.G., *et al*. (2003) Data Exchange: Semantics and Query Answering. In: *Database Theory—ICDT* 2003, Springer, Berlin, Heidelber, 207-224. https://doi.org/10.1007/3-540-36285-1_14

[2] Fagin, R., Kolaitis, P.G. and Popa, L. (2005) Data Exchange: Getting to the Core. *ACM Transactions on Database Systems*, **30**, 174-210. https://doi.org/10.1145/1061318.1061323

[3] Pichler, R. and Savenkov, V. (2010) Towards Practical Feasibility of Core Computation in Data Exchange. *Theoretical Computer Science*, **411**, 935-957. https://doi.org/10.1016/j.tcs.2009.09.035

[4] Kimmig, A., Memory, A., Miller, R.J. and Getoor, L. (2017) A Collective, Probabilistic Approach to Schema Mapping. 2017 *IEEE* 33*rd International Conference on Data Engineering* (*ICDE*), San Diego, CA, 19-22 April 2017, 921-932. https://doi.org/10.1109/ICDE.2017.140

[5] Gottlob, G. and Nash, A. (2006) Data Exchange: Computing Cores in Polynomial Time. *ACM Sigmod-Sigact-Sigart Symposium on Principles of Database Systems*, June 2006, 40-49. https://doi.org/10.1145/1142351.1142358

[6] Yousfi, A., Elyazidi, M.H. and Zellou, A. (2018) Assessing the Performance of a New Semantic Similarity Measure Designed for Schema Matching for Mediation Systems. In: *International Conference on Computational Collective Intelligence*, Springer, Cham, 64-74. https://doi.org/10.1007/978-3-319-98443-8_7

[7] Kettouch, M., Luca, C. and Hobbs, M. (2017) Schema Matching for Semi-structured and Linked Data. 2017 *IEEE* 11*th International Conference on Semantic Computing* (*ICSC*), San Diego, CA, 30 January-1 February 2017, 270-271. https://doi.org/10.1109/ICSC.2017.104

[8] Sekhavat, Y.A. and Parsons, J. (2017) SEDEX: Scalable Entity Preserving Data Exchange. 2017 *IEEE* 33*rd International Conference on Data Engineering* (*ICDE*), San Diego, CA, 19-22 April 2017, 65-66. https://doi.org/10.1109/ICDE.2017.39

[9] Alexe, B., ten Cate, B., Kolaitis, P.G. and Tan, W. (2011) EIRENE: Interactive Design and Refinement of Schema Mappings via Data Examples. *Proceedings of the VLDB Endowment*, **4**, 1414-1417. https://doi.org/10.1145/2043652.2043656

[10] Alexe, B., Hernndez, M., Popa, L. and Tan, W.C. (2012) MapMerge: Correlating Independent Schema Mappings. *The VLDB Journal*, **21**, 191-211. https://doi.org/10.1007/s00778-012-0264-z

[11] 解筱, 张克, 任伯群, 等. ETL 技术在商业银行数据整合中的研究与应用[J]. 信息技术与信息化, 2019(7): 45-47.

[12] 丁强龙, 王津, 张学杰. 基于子模式的关系数据到图数据 ETL 方法研究[J]. 计算机工程与应用, 2017, 53(12): 76-84.

[13] 李磊. ETL 任务集群调度方法[J]. 计算机技术与发展, 2018, 28(11): 41-44.

[14] Baker, C.A. (1995) Extended Skolem Sequences. *Journal of Combinatorial Designs*, **3**, 363-379. https://doi.org/10.1002/jcd.3180030507

[15] Ravichandra, S. and Somayajulu, D.V.L.N. (2015) Core Schema Mappings: Computing Core Solution with Target

Dependencies in Data Exchange.

[16] 吕劲松, 王忠. 金融审计中的数据分析[J]. 审计研究, 2014(5): 28-33.

[17] Fan, H., Deng, K. and Liu, J. (2016) An Approach of XML Schema Matching Using Top-K Mapping. 2016 3*rd International Conference on Information Science & Control Engineering* (*ICISCE*), Beijing, 8-10 July 2016, 174-178. https://doi.org/10.1109/ICISCE.2016.47

[18] Hsu, I.C., Yang, L.J., Huang, D.C., *et al*. (2014) Integrating Semantic Web Technologies with XML Schema Using Role-Mapping Annotations. *The Electronic Library*, **32**, 147-169. https://doi.org/10.1108/EL-07-2012-0096