模式匹配中的结构差异识别及消解

杜小坤[1] 李国徽[2] 李艳红[1]

(中南民族大学计算机科学学院 武汉 430074)[1] (华中科技大学计算机学院 武汉 430074)[2]

摘 要 模式匹配是数据空间、语义Web等热点研究领域的一个关键问题。已有的研究成果以元素为操作对象,通过元素的自身信息、结构信息和数据信息等来获取元素语义并选取语义相近的元素作为匹配元素,取得了较好的效果。但不同模式在元素自身信息、结构信息上的巨大差异严重阻碍了语义的获取。分析了模式结构差异产生的原因,总结了几种模式差异常见的形式,并给出了相应的检测和消解算法来消除差异。实验表明,对模式进行差异消解后再匹配能显著提高匹配结果的准确率。

关键词 模式匹配,结构信息,结构差异

中图法分类号 TP301.131 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.040

Structural Difference Recognition and Dispelling in Schema Matching

DU Xiao-kun[1] LI Guo-hui[2] LI Yan-hong[1]

(College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China)[1]

(Department of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)[2]

Abstract Schema matching is a primary problem in the hot research field such as data space, semantic Web and so on. The existing method extracts the element's own information, structural information and data information, and then chooses the pair of elements having most similar semantic as matching elements. But the difference between elements in element's own information and structural information hinders the extraction of semantic. Through analyzing the reason of structure information difference, this paper summarized some kinds of structure information difference and proposed the corresponding detecting and dispelling algorithm. Extensive simulation experiments were conducted and the results show that the accuracy of matching result is increased by the dispelling of structure information difference.

Keywords Schema matching, Structural information, Structural difference

1 绪论

模式匹配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,其广泛应用于数据空间、数据集成、语义Web等热点研究领域。模式匹配经过十多年的研究,已经取得了丰富的成果[1-5,8]。目前的一些研究成果普遍以模式元素为对象,利用模式中的各种信息(元素自身信息、模式结构信息、数据实例信息等)挖掘元素的语义并计算语义相似度,最后选取语义相似度较高的元素对作为匹配元素。对同一现实世界对象,不同的设计习惯、设计目的会导致模式中元素自身信息及元素间关联的结构信息产生较大的差异,严重影响了元素语义的获取及相似度的计算。对于元素自身信息的差异,目前已有较多的研究成果[4,6,7,9],它们采用了近义词检索、缩/简写还原、数据类型匹配等措施处理元素名称、数据类型等自身信息上的差异,使元素自身信息具有相对统一的描述形式,以提高语义相似度计算结果的准确性。但对于模式结构信息的差异,目前还没有有效的处理手段。本文分析了模式结构差异产生的原因,并给出了相应的识别和消解算法,通过消除结构差异来优化结构信息的使用,进而提高匹配结果的准确率。实验表明,在匹配前进行结构差异识别及消解处理能显著提高匹配结果的准确率。

本文的主要创新点如下:

①提出了通过消除结构差异来优化结构信息,进而提高匹配准确率的新思路。由于不同模式元素自身信息的差异,对获取的语义信息进行一致性处理能够提高匹配准确率,所以通过结构信息进行一致性处理同样也能提高匹配准确率。

②总结了几种结构差异的常见形式,并给出了相应的识别及消解算法。

本文第2节介绍相关的研究成果;结构差异识别及消解算法在第3节给出;第4节对本文提出的方法与已有方法进行实验对比;最后是结论与展望。

2 相关工作

模式匹配的主要目标是获取高准确率的元素匹配关系,

到稿日期:2014-04-02 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),中央高校基本科研业务费(CZQ14015)资助。

杜小坤(1980—),男,博士,讲师,主要研究方向为数据集成、模式映射,E-mail:hustdxkun@163.com;李国徽(1973—),男,博士,教授,主要研究方向为现代数据库;李艳红(1973—),女,博士,副教授,主要研究方向为路网查询。

目前已有的研究成果基本都从如下两个方面着手：①获取新的辅助信息；②优化辅助信息的使用方式。但由于模式自身表达能力的缺陷，模式中的各种辅助信息并不能够完全准确地描述对应数据的语义，因此自动模式映射算法并不能保证获取正确的匹配结果，高效、方便的人工干预方式也是目前的一个研究重点。下面分别对其中的典型算法进行介绍。

2.1 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[4,8]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[4,9,10-13]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)获取元素匹配关系。其中 SF 方法[12]首先建立模式结构图并计算节点间的相似度，然后根据图中相关节点相似度相互影响的原理不断地调整节点相似度直至收敛到一个稳定的值，最后根据调整后的值选取元素匹配。该方法虽然有效利用了模式结构信息，但其将模式中所有信息都以节点的形式在图中表示，极大地增加了图结构的复杂性，从而使得算法的时间复杂度很高，并且某些无效信息也参与相似度调整，削弱了结构信息的作用。PFD_based 方法[13]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，但由于其仍然以元素为基础描述结构关联，因此存在信息量大、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，研究人员还发现了一些其它类型的信息可用以辅助匹配。申德荣等人提出的 SKM 模型[14]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[15]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[16]提出一种利用本体辅助获取模式间复杂元素匹配关系的方法。

2.2 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 等人[17]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 等人[18]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。

2.3 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 等人[19]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 等人[1]提出借助用户对现实世界的敏感直觉来辅助映射的思路，即针对匹配过程中的不确定因素生成一些简单问题，通过 CrowdSourcing 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。黄少滨等人[20]提出的 PVMM 模型将专家的手工干预过程放到匹配算法前，通过专家提供的少量准确匹配关系来大幅提高全局匹配准确率。

3 结构差异识别及消解

关系数据库的规范化理论为模式结构设计提供了理论指导。规范化理论一般要求建立的模式满足 3NF，对 3NF 较为通俗的理解是：对每个对象分别建立关系，将同一对象的各个属性放到同一关系中。对于同一组现实世界对象，若统一按照 3NF 的要求建立模式，则建立模式的结构基本相同(称为原始结构)。但 3NF 并不是一个强制要求，设计人员根据 3NF 要求建立模式后，还需根据具体的应用需求进行相应的结构优化。不同的优化目的和优化方法会对模式结构产生不同的影响，从而导致描述相同信息的不同模式在结构信息上存在较大差异，进而降低了结构信息对匹配的辅助作用。若能准确识别模式中的相关优化操作，并将其还原为模式的原始结构(即将待匹配模式的结构还原到一个较为一致的状态)，则能显著提高结构信息的辅助作用，提高匹配准确率。

文献[21,22]介绍了常见的模式结构优化的方法，这些结构优化方法采用的策略是空间换时间，即通过增加一定量的数据冗余来提高查询的速度。具体采用的优化措施主要有：①增加冗余列；②增加派生列；③合并表；④重复表；⑤分割表。通过分析这些优化措施对模式结构产生的影响，我们将其引起的模式结构变化主要分为如下 3 种形式：①属性冗余；②纵向合并；③横向分割。下面分别给出每种结构变化相应的识别及还原算法。

3.1 部分函数依赖

元素间的函数依赖关系是结构信息的一种表现形式，模式结构的改变会对其产生相应的影响(隐藏一部分依赖关系)，若能挖掘出隐藏的结构信息，就能还原出原始结构信息，通过分析原始结构信息，就可识别出模式结构的变化。

元素间的依赖关系是元素对应数据间的固有关系，模式结构的改变虽然改变了元素的组织形式，但元素数据的依赖关系不会随之改变。Berzal F 等人[23]将元素对应数据间的依赖关系称为元素间的部分函数依赖关系，并对此进行了深入的研究。我们首先挖掘元素间隐含的部分函数依赖关系，然后以此为依据分析其中出现的模式结构变化。

定义 1 对关系 r 中任意两个属性集 X,Y，我们称满足如下条件的元组集合 $r_e \subset r$ 为关系 r 的函数依赖例外集：

(1)$(r-r_e)$中所有元组满足 $X \rightarrow Y$。

(2)$\forall t \in r_e$，$(r-r_e) \cup \{t\}$中的元组不都满足 $X \rightarrow Y$。

(3)不存在 $r_e' \subset r$ 满足条件(1)和(2)并且 $\#(r_e') < \#(r_e)$($\#(r)$表示关系 r 中的元组数)。

同时将 r_e 中元组数称为关系 r 上部分函数依赖例外数，记为 $\exp_e(X \rightarrow Y) = |r_e|$；将 $|r-r_e|$ 与总元组数的比值称为部分函数依赖度，记为 $pfd_e(X \rightarrow Y) = |r-r_e|/|r|$。

根据上述定义，以模式中一定量的数据实例为基础计算出任意两个元素之间的部分函数依赖度，然后选取部分函数依赖度值较高的依赖关系作为分析结构变化的依据。该过程需要考虑数据集的规模，数据集的规模对部分函数依赖度的计算有如下影响：当数据集的规模较小时，计算的部分函数依

赖度可能会产生较大的随机误差；当数据集的规模较大时，计算会占用大量的时间。选取一个合适的数据集规模是一个较为重要的问题，我们在对不同规模的数据集进行测试后选取数据集规模为 5000 个元组。对图 2 中的模式 T，根据定义计算元素间的部分函数依赖度，如表 1 所列。

表 1　图 2 中模式 T 的元素间部分函数依赖度

	ID	Name	Singer	Issue Date	Birthday	Nationality	Company
ID		1	1	1	1	1	1
Name	1		1	1	1	1	1
Singer	0.14	0.14		0.14	1	1	1
IssueDate	0.64	0.64	0.64		0.64	0.64	0.64
Birthday	0.13	0.13	0.98	0.14		0.01	0.01
Nationality	0.01	0.01	0.02	0.01	0.01		0.74
Company	0.01	0.01	0.02	0.01	0.01	0.11	

下面以元素间的部分函数依赖关系为基础分析其中的结构变化。

3.2　属性冗余识别与消解

模式设计人员为了提高某些常用查询的执行速度，会将某些属性重复存储以避免费时的连接操作。例如图 1 中，为了提高目标模式 T 中根据专辑名称查询歌手名称的速度，设计人员添加了属性 MusicAlbum. SName，该属性是 SingerInfo. Name 属性的重复。在模式匹配时，这两个元素应匹配源模式 S 中的同一元素，但已有的模式匹配方法却未考虑该情况，为其分别选取匹配元素，这显然会造成错误匹配。

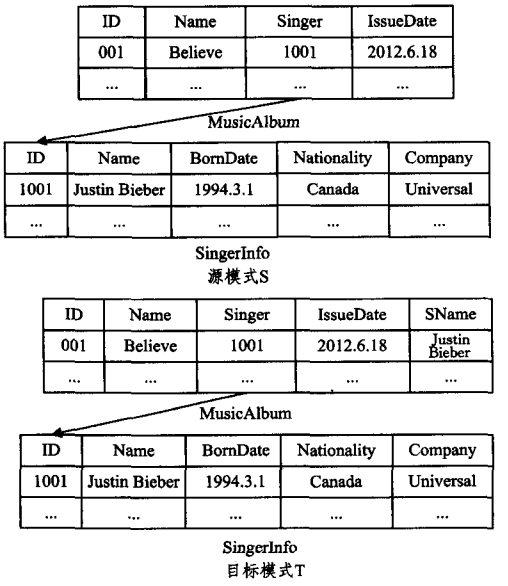

图 1　存在属性冗余的模式 S、T

定义 2　模式中存在函数依赖关系 $A\rightarrow B, A\rightarrow B'$，若对 $DOM(A)$ 中任意值 x，属性 B 和 B' 中对应值 y 和 y' 满足下列情况之一：

①若 $y=$null，则 $y'=$null；

②若 $y!=$null，则 $y'=y$ 或 $y'=$null。

则称依赖关系 $A\rightarrow B$ 与 $A\rightarrow B'$ 相同，且 B' 是 B 的冗余属性，B 是 B' 的源属性，源属性 B 可能存在多个冗余属性，我们把 B 的所有冗余属性的集合记为 RedundantSet(B)。

根据上述冗余属性的定义，我们选取模式中所有部分函数依赖度为 1 的依赖关系作为依赖关系集进行冗余属性的判断。对所有具有相同左部的函数依赖关系的右部进行检测，可发现所有的冗余属性。冗余属性的判断过程需要考虑如下两个问题：数据集规模和算法执行效率。对于数据集规模，我们选取与计算部分函数依赖度相同大小的数据集。由于冗余属性与源属性对应数据间的严格等值关系，仅需判断个别数据即可排除大量的属性。因此在判断冗余属性时，我们首先仅做少量的数据连接，当发现属性间的值满足冗余属性的条件时才进行所有数据的连接判断，提高了冗余属性的识别效率。

由于冗余属性是源属性的重复，可根据源属性推导出冗余属性的值，因此对冗余属性的消解采取简单的删除策略，即在模式匹配时不考虑冗余属性，而最终匹配结果中冗余属性与其对应的源属性具有相同的匹配元素。

3.3　纵向合并识别与消解

模式设计人员经常采用的另一种结构优化手段是将多个相互关联的实体合并到同一关系中，这样能够显著提高实体间相互查询的响应速度。如图 2 所示，模式 T 将专辑和歌手信息存储到同一个关系中，这虽然增加了添加和修改操作的难度，但提高了专辑和歌手信息互相查询的速度。我们把这种形式的模式结构变化称为纵向合并。

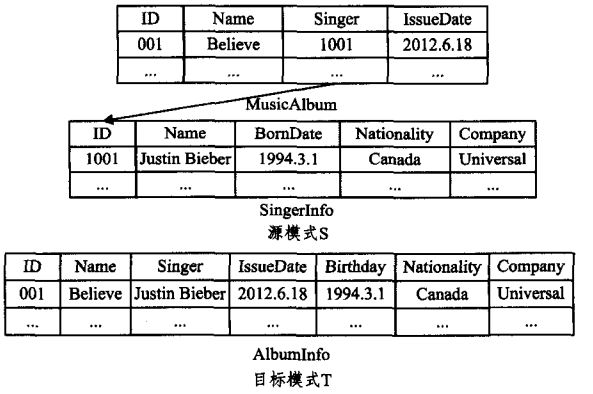

图 2　存在结构差异的模式 S 和 T

在模式匹配时，目标模式 T 中关系 AlbumInfo 中的所有属性具有基本相同的结构信息，若其中存在自身信息相似的元素，则难以有效识别，例如：目标模式 T 中属性 Name 与源模式 S 中的两个名称都为 Name 的属性间的匹配关系难以确定。若能有效地将目标模式 T 中的关系根据描述实体的不同进行分解，则能有效利用结构信息解决属性名称为 Name 的元素间的匹配问题。

纵向合并优化将多个关系合并，我们通过分析其中元素的函数依赖关系可对其进行还原。对元素依赖关系的选取，可从 3.1 节中计算出的部分函数依赖关系中选取依赖度值为 1 的依赖关系作为最终依赖关系。例如根据表 1 选取图 2 中模式 T 的函数依赖关系集为 $F_T=\{$ID→(Name, Singer, IssueDate, Birthday, Nationality, Company), Name→(ID, Singer, IssueDate, Birthday, Nationality, Company), Singer→(Birthday, Nationality, Company)$\}$。得到函数依赖关系后，以依赖关系为基础，将模式按照描述实体的不同分解为不同的分块。具体的分解流程如算法 1 所示：

算法 1　Schema_Decompose(R)

输入：关系 R(U,F)，F 为关系 R 中元素满足的函数依赖关系集

输出：关系 R 的分解 ρ

{

1. 函数依赖集 F 极小化处理(去除冗余依赖，获取最小覆盖)；
2. 将 F 中未出现的属性单独构成一个关系，并从 U 中删除；
3. 如有 F 存在函数依赖 X→A∈F 且 XA=U，则算法终止；否则进行第 4 步。
4. 对 F 按相同左部原则分组，每组的全部属性为一个属性集 U_i，如生成的某一属性集 U_i 被其它属性集 U_j 包含，则去掉 U_i；否则令 F_i 为 F 在 U_i 上的投影，每对〈U_i，F_i〉构成一个分解后的关系模式。
5. 在第 4 步得到的模式分解结果 $\rho=\{R_1(U_1,F_1),R_2(U_2,F_2),\cdots,R_n(U_n,F_n)\}$ 中判断是否存在某个 U_i 使得关系 R 的主码 $X\in U_i$，若存在则返回 ρ 为分解结果；否则返回 $\rho'=\rho\cup R*\langle X,FX\rangle$。

}

对于图 2 中的模式 T，第 1 步对 T 中的函数依赖集 F_T 进行极小化处理，获取 F_T 的最小函数依赖集为 F_{TM}={ID→Name，ID→Singer，ID→IssueDate，Name→ID，Singer→BornDate，Singer→Nationality，Singer→Company}，由于 F_{TM} 不满足第 2、3 步提出的条件，因此直接进行第 4 步，将关系划分为：$\rho=\{R_1$(ID，Name，Singer，IssueDate)，R_2(Singer，BornDate，Nationality，Company)$\}$。由于上述划分中，主键 ID 已经包含在 R_1 中，因此 ρ 为模式分解的结果。该结果与图 2 中模式 S 的结构基本一致，消除了模式间由于纵向合并优化引入的结构差异。

3.4 横向分割识别与消解

对于模式中某个包含较多元组的关系，设计人员可采用横向分割的方式进行优化。例如图 3 中，设计人员为提高对歌手信息的查询速度，将歌手根据其所属的唱片公司的不同进行分割并分别存储到不同的关系中，我们把这样的一组关系称为相似关系。

ID	Name	BornDate	Nationality	Company
1001	Justin Bieber	1994.3.1	Canada	Universal
...	...	...	...	...

SingerInfo

源模式S

ID	Name	BornDate	Nationality
1001	Justin Bieber	1994.3.1	Canada
...	...	...	...

SingerInfo_Universal

ID	Name	BornDate	Nationality
2001	Craig David	1981.5.5	UK
...	...	...	...

SingerInfo_Warner

……

目标模式T

图 3 存在横向合并与分割差异的模式 S、T

若直接对源、目标模式进行匹配操作，常规匹配算法会为源模式中的属性寻找唯一的匹配属性，这会产生错误的匹配结果。例如在已有匹配方法的最终结果中，源模式中属性 SingerInfo. Name 仅与目标模式中名称为 Name 的多个属性之一匹配，但事实上前者与目标模式中一组相似关系中的多个名称为 Name 的属性相匹配。若能在匹配前正确识别目标模式中的相似关系并采取措施，则能有效避免匹配结果中的类似问题。算法 2 为相似关系的识别算法。

算法 2

横向分割识别(R)

{

1. 采用传统模式匹配方法，计算 R 中元素间自身信息的相似度；
2. 为 R 中每个元素选取候选匹配元素(设定阈值为 α)；
3. 根据元素相似性获取关系的元素相似度并选取元素相似的关系；
4. 根据结构相似规则从元素相似的关系中选取相似关系。

}

对算法第 1 步，采用 COMA++方法计算元素自身信息相似度；然后在第 2 步中为每个元素 e 选取相似度大于阈值 α 的候选匹配组成集合 $CAND(e)$。

定义 3 对模式中任意两个关系 $R_1(e_1,e_2,\cdots,e_n)$ 和 $R_2(f_1,f_2,\cdots,f_m)$，若满足如下关系：

$$\frac{|\{e_i|(CAND(e_i)\cap R_2)\neq\emptyset\}|+|\{f_i|(CAND(f_i)\cap R_1)\neq\emptyset\}|}{m+n}\geqslant\beta$$

则称 R_1 和 R_2 元素相似(β 为给定阈值)。

定义 4 对于任意元素相似的关系对 R_1 和 R_2，若 R_1 中存在外键 e_i，其对应关系 R' 的主键，则 R_2 中一定也存在对应 R' 主键(或 R' 的相似关系)的外键 f_j，且 $f_j\in CAND(e_i)$，则称 R_1 和 R_2 为相似关系。

算法 2 的第 3 步和第 4 步可分别根据定义 3 和定义 4 进行计算，并最终得到模式中的相似关系。由于相似关系具有传递性，可将具有相似性的关系划分为一组。例如可将图 3 中模式 T 中的多个相似关系(SingerInfo_Universal，SingerInfo_Warner，…)划分到同一组。获取模式中的相似关系后，可依据相似关系间元素的对应关系进行关系合并。例如将上述相似关系组进行合并后得到关系 SingerInfo{ID，Name，BornDate，Nationality，Type}。

前面给出了常见的结构优化识别算法，本文第 4 节将对识别结果的准确率进行实验验证。在进行模式匹配操作前，首先对模式进行结构优化识别及消解，不同的识别及消解算法按照一定的顺序执行，例如：首先进行属性冗余的识别及消解，然后进行纵向合并的识别与消解，再进行横向分割的识别与消解。识别和消解算法使模式结构信息处于一个较为一致的状态，从而提高匹配准确率。另外，当在一些特殊模式中还存在其它类型的结构差异时，操作人员可有针对性地设计相应的识别及消解算法并添加到执行序列的合适位置，所以该策略具备较强的可扩展性。

4 实验结果分析

测试用例的选取分为两个步骤，首先确定模式，然后获得数据实例。选取的源、目标模式分别为两家销售同类产品的公司的进销存管理数据库，称为 DB1 和 DB2(两个数据库中都有 20 多个关系和 200 多个属性)。模式中的数据实例取自两个公司的实际数据，若关系中有超过 5000 条记录，则取前 5000 个记录，否则取所有数据。数据库管理系统为MySql5. 5，使用 ODBC 连接数据库获取各种信息。主键硬件采用 Intel Core i3 双核 2. 27G 处理器，4G 内存；操作系统为 Windows 7。

4.1 结构差异识别准确率实验

为验证第3节中各种识别算法的识别准确率，首先通过人工方式对DB1和DB2中的各种结构优化进行识别，然后利用第3节中的识别算法对结构优化进行自动识别，实验结果如表2所列。

表2 DB1和DB2中的结构差异的情况统计

	DB1			DB2		
	人工识别	正确识别	错误识别	人工识别	正确识别	错误识别
属性冗余	10	10	2	13	13	1
纵向合并	4	3	1	3	3	1
横向分割	2	2	0	1	1	0

通过表2中的测试数据可知，3.2节提出的冗余属性识别算法可识别出DB1和DB2中所有的冗余属性，但也存在少量的错误识别情况；3.3节对纵向合并的识别及分解算法分别识别出DB1中4个纵向合并中的3个以及DB2中的全部3个纵向合并，但都有1个错误识别；3.4节中的横向分割识别算法能够完全正确地识别两个模式中所有的3个横向分割操作，且没有错误识别。通过上述数据可知，本文第3节提出的几种典型结构优化识别算法具有较高的准确率，能够正确识别出模式中绝大多数结构优化。

4.2 提高匹配准确率实验

为了验证结构差异识别及消解操作能够有效提高匹配结果的准确率，我们以Cupid算法为基础，将直接运行该算法得到的匹配结果与结构差异识别及消解处理后再运行该算法得到的匹配结果进行对比，匹配结果通过如下3个参数来评价。

(1)查准率(Precision)：匹配结果中正确匹配结果占所有匹配结果的比率。

$$Precision=T/P=T/(T+F)$$

(2)查全率(Recall)：匹配结果中正确匹配结果占实际匹配结果的比率。

$$Recall=T/R$$

(3)全面性(Overall)：通过使用匹配算法所节省的工作量占总的匹配工作量的比率。

$$Overall=Precision*(2-\frac{1}{Recall})=\frac{T-F}{R}$$

其中，T为匹配算法返回的正确匹配结果；P为匹配算法返回的所有匹配结果；F为匹配算法返回的错误匹配结果；R为所有正确的匹配结果。实验结果如图4所示。

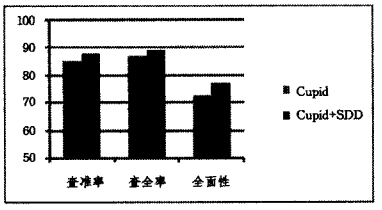

图4 进行结构差异消解与不进行结构差异消解的Cupid算法结果对比

通过图4可知，由于结构差异识别和消解操作为Cupid算法提供了更准确的结构信息(同属于一个信息元比同属于一个关系更能准确地描述元素间的关联形式)，因此匹配结果在各个指标上均有所提高。其中查准率从85%提高到88%，查全率从87%提高到89%，全面性则从72%提高到78%。综上，在匹配前进行结构差异识别和消解操作能够提高匹配结果的准确度。

结束语 不同设计人员针对同一信息创建模式时，不同的设计习惯、设计目的会导致最终的模式存在较大差异，主要表现为元素自身信息和模式结构信息的不同，这给获取正确的元素匹配关系带来了极大的障碍。对于元素自身信息差异的处理，目前已有较多的研究成果，但对于模式结构差异却没有有效的处理手段。本文针对模式结构差异，提出一种在匹配前进行识别并消解的策略，通过对结构差异进行有效的识别和消解获取准确的结构信息，进而提高匹配的准确率。在后续研究中，我们将进一步简化结构信息的描述形式，提高用户干预的工作效率，减小用户的实际工作量。

参考文献

[1] 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

[2] Nguyen Q V H, Weidlich M, Nguyen Thanh T, et al. Pay-as-you-go Reconciliation in Schema Matching Networks[OL]. http://infoscience.epf/.ch/record/189892

[3] Lee Y, Sayyadian M, Doan A H, et al. eTuner: tuning schema matching software using synthetic scenarios[J]. The VLDB Journal, 2007, 16(1): 97-122

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

[5] De Carvalho M S G, Laender A H F, Gonçalves M A, et al. An evolutionary approach to complex schema matching[J]. Information Systems, 2013, 38(3): 302-316

[6] Madhavan J, Bernstein P A, Rahm E. Generic schema matching with cupid[OL]. http://db.cs.washington.edu/papers/CupidTechReport.pdf

[7] Do Hong-hai, Rahm E. COMA-a system for flexible combination of schema matching approaches[C]// Proc. of VLDB. 2002: 610-621

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

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

[10] De Carvalho M S G, Laender A H F, Gonçalves M A, et al. An evolutionary approach to complex schema matching[J]. Information Systems, 2013, 38(3): 302-316

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

[12] Melnik S, Garcia-Molina H, Rahm E. Similarity flooding: A versatile graphmatching algorithm and its application to schema matching[C]// Proceedings of the 18th International Conference on Data Engineering. 2002: 117-128

[13] 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

[14] 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

[15] 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. 2011: 121-132

[16] PinkelC. InteractivePayasYouGoRelational-to-OntologyMapping [C]// The Semantic Web-ISWC. 2013:456-464

[17] 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

[18] 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

[19] 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

[20] 黄少滨，刘国峰，万庆生，等. 一种基于部分已验证匹配关系的模式匹配模型[J]. 自动化学报，2013，39(10)：1642-1652

[21] 董慧，刘厚嘉. 文献数据库优化设计的探讨[J]. 情报学报，1999，18(1)：43-49

[22] 崔跃生，张勇，曾春，等. 数据库物理结构优化技术[J]. 软件学报，2013，24(4)：761-780

[23] Berzal F, Cubero J C, Cuenca F, et al. Relational decomposition through partial functional dependencies[J]. Data & Knowledge Engineering, 2002, 43(2): 207-234

(上接第 184 页)

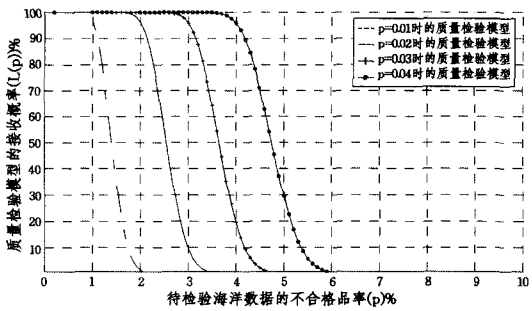

图 5 模糊检验模型和概率检验模型的 OC 曲线比较($n=N*30\%$)

由表 2—表 4 和图 3—图 5 可以看出：1)基于模糊不合格品率可推导出两端点模糊抽样检验模型，即上、下限模糊抽样检验模型。以抽样比为 10%的观测站 1 为例，其上限模糊抽样检验模型为 $S(7560,756,12)$，接收数为 12；下限模糊抽样检验模型为 $S(7560,756,40)$，接收数为 40。即因该海洋数据具有不确定的不合格品率，其质量检验模型的接收数可在 12 至 40 之间选取。2)基于不确定不合格品率的模糊抽样检验模型是具有明确不合格品率质量参数的抽样检验模型的扩充，其可涵盖模糊不合格品率的所有变化情况。即上、下限模糊抽样检验模型的接收数区间涵盖了其不合格品率为确定参数(0.02 或 0.03)时的概率抽样检验模型。3)不同模糊不合格品率的模糊抽样检验模型的辨别率亦不同，即上限模糊抽样检验模型具有最强的辨别力，而下限模糊抽样检验模型的辨别力最弱；用户在不确定不合格品率的情况下，可根据精度要求选择适当的质量抽样检验模型。

结束语 不同于传统工业产品的生产形式，海洋数据的采集方式多式多样，包括实地测量、遥感、摄影测量、数据化、文档报表等，因此海洋数据质量特性具有不确定性。传统的抽样检验方式不能满足海洋数据的质量检验要求。针对海洋数据不确定的质量特性，本文在抽样检验模型的制定中引入了梯形模糊数，扩充了传统概率优化抽样检验模型的制定方法，完善了海洋数据的抽样检验理论体系。

参 考 文 献

[1] 张耀中. 质量抽样检验标准实施指南[M]. 深圳：海天出版社，2004：3-16

[2] 于善奇. 抽样检验与质量控制[M]. 北京：北京大学出版社，1991：15-49

[3] Dodge H F, Roming H G. Single sampling and double sampling inspection tables[J]. The Bell System Technical Journal, 1941, 20(1): 1-61

[4] Dodge, H F. A sampling inspection plan for continuous production[J]. The Annals of Mathematical Statistics, 1943, 14(3): 264-279

[5] Dodge H F, Roming H G. Sampling Inspection Table, Single and Double Sampling[M]. New York: John Wiley & Sons, 1959: 118-220

[6] Jun C H, Balamurali S, Kalyanasundaram M, et al. Evalution and design of two level continuous sampling plans[J]. Tamkang-Journal of Science and Engineering, 2006, 9(4): 409-417

[7] Duarte B P M, Saraiva P M. An optimization-based approach for designing attribute acceptance sampling plans[J]. International Journal of Quality & Reliability Management, 2008, 25(8): 824-841

[8] Eleftherion M, Farmakis N. Continuous sampling plan under quadratically varying acceptance cost[C]// The XIII International Conference "Applied Stochastic Models and Data Analysis". Vilnius, Lithuania, 2009: 289-293

[9] Wang Jing-feng, R Hai-ning, Cao Zhi-dong, et al. Sampling surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning[J]. International Journal of Geographical Information Science, 2010, 24(4): 523-543

[10] Aslam M, Balamurali S, Jun C H, et al. Optimal designing of a skip lot sampling plan by two point method[J]. Pakistan Journal of Statistics, 2010, 26(4): 585-592

[11] Ma M, Friedman M, Kandel, et al. A new fuzzy arithmetic[J]. Fuzzy Sets and Systems, 1999, 108: 83-90

[12] Wetherill G B. Sampling Inspection and Quality Control[M]. Chapman and Hall, London, 1977: 233-267

[13] Govindaraju K, Balainurali S. Chain sampling plan for variables inspection[J]. Journal of Applied Statistics, 1998, 25(1): 103-109

[14] 刘大杰，刘春. GIS 数字产品质量抽样检验方案探讨[J]. 武汉测绘科技大学学报，2000，24(4)：348-361