一种基于部分已验证匹配关系的模式匹配模型

黄少滨 [1] 刘国峰 [1] 万庆生 [1] 程 媛 [1] 申林山 [1]

摘 要 模式匹配是模式集成、语义 WEB 及电子商务等领域的重点及难点问题. 为了有效利用专家知识提高匹配质量, 提出了一种基于部分已验证匹配关系的模式匹配模型. 在该模型中, 首先, 人工验证待匹配模式元素间的少量对应关系, 进而推理出当前任务下部分已知的匹配关系及单独匹配器的缺省权重; 然后, 基于上述已收集到的先验知识对多种匹配器所生成的相似度矩阵进行合并及调整, 并在全局范围内进行优化; 最后, 对优化矩阵的选择性进行评估, 从而为不同匹配任务推荐最合理的候选匹配生成方案. 实验结果表明, 部分已验证匹配关系的使用有助于模式匹配质量的提高.

关键词 模式匹配, 模式集成, 相似度矩阵, 候选匹配

引用格式 黄少滨, 刘国峰, 万庆生, 程媛, 申林山. 一种基于部分已验证匹配关系的模式匹配模型. 自动化学报, 2013, **39**(10): 1642−1652

DOI 10.3724/SP.J.1004.2013.01642

A Schema Matching Model Based on Partial Verified Matching Relations

HUANG Shao-Bin[1] LIU Guo-Feng[1] WAN Qing-Sheng[1] CHENG Yuan[1] SHEN Lin-Shan[1]

Abstract Schema matching is an important and difficult problem in many database application domains, such as data integration, semantic web and data warehousing and so on. In order to use experts′ knowledge to improve the matching quality effectively, a schema matching model is proposed based on partial verified matching relations. In this model, first, a small amount of correspondences between schemas elements are verified by manual, and the partial known matching relations and default weights of different matchers are reasoned on the current task; second, the similarity matrices of multiple matchers are combined based on the collected priori knowledge, and optimized under global scope; finally, the selectivity of the optimization matrix is evaluated, and the most reasonable candidate matching generation plans for different matching tasks are generated. Experimental results show that the use of partial verified matching helps to improve the quality of schema matching.

Key words Schema matching, schema integration, similarity matrix, candidate matching

Citation Huang Shao-Bin, Liu Guo-Feng, Wan Qing-Sheng, Cheng Yuan, Shen Lin-Shan. A schema matching model based on partial verified matching relations. *Acta Automatica Sinica*, 2013, **39**(10): 1642−1652

模式匹配在众多应用领域当中发挥着至关重要的作用, 如数据集成中模式元素关联关系的标识, 语义 WEB 中不同本体概念间的映射, 电子商务中不同 XML 信息间的交换等. 模式匹配是一个在模式元素间生成对应关系的操作, 它以两个待匹配模式作为输入, 以语义相近或相似的元素映射关系作为输出[1]. 目前的模式匹配操作大都是由系统开发人员或数据库管理员 (Database administrator, DBA) 手工完成, 匹配过程不仅费时、费力, 而且容易出错, 因此, 急需一种自动化的模式匹配方法来代替或辅助人工完成匹配任务.

近年来, 一些自动化的模式匹配方法被提出[2−11], 这些方法利用元素名称[2−3, 5, 10]、数据实例特征[4, 6−9, 11] 和模式结构[2−4, 9−11] 等信息进行匹配. 其中具代表性的模式匹配方法有: Microsoft 的 CUPID 方法[2], 该方法利用名称匹配器并结合模式内部元素间的关联关系对匹配进行推理; Stanford 大学的 SF (Similarity flooding) 方法[3] 同样利用名称匹配器计算元素间的初始相似度, 并基于两个相似节点的相邻节点也相似的假设对初始相似度进行求精; Leipzig 大学的 COMA (Combination of schema matching approaches) 方法[5] 使用不同策略对多种类型匹配器的结果进行合并, 以提高匹配结果的准确性及匹配操作的灵活性; Northwestern 大学的 SemInt 方法[6] 和 Washington 大学的 LSD

收稿日期 2012-10-23 录用日期 2013-01-18
Manuscript received October 23, 2012; accepted January 18, 2013
国家自然科学基金 (60873038, 60903080, 71272216), 国家科技支撑计划项目 (2009BAH42B02, 2012BAH08B02), 中央高校基本科研业务专项资金项目 (HEUCF100603, HEUCFZ1212) 资助
Supported by National Natural Science Foundation of China (60873038, 60903080, 71272216), National Key Project of Scientific and Technical Supporting Programs (2009BAH42B02, 2012BAH08B02), and Fundamental Research Funds for the Central Universities (HEUCF100603, HEUCFZ1212)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Ling
1. 哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001
1. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001

方法[7] 基于机器学习技术对目标模式中的元素特征进行刻画, 并利用机器学习的泛化能力建立新元素到目标元素的映射; Duplicates 方法[8] 利用重复检测算法发现待匹配模式间的重复数据, 并根据重复数据中相同数据对应的元素相同的原理寻找对应的属性.

上述方法在匹配执行阶段主要采用自动化的方式, 通常没有专家对匹配过程进行指导, 即便存在, 也仅限于匹配前模式信息的提取及辅助信息的准备. 实践表明, 该类方法并不能较好地适应模式匹配任务, 往往存在匹配结果准确性不高、匹配方法适用性差等缺陷[12]. 为此, 部分研究人员试图在匹配中引入专家知识, 以改善模式匹配的质量. 目前, 专家知识的引入主要有以下两种方式: 1) 提供一个图形用户界面, 操作人员可以在其上对匹配关系进行手动编辑或检查、纠正已得到的对应关系, 如 COMA++[13]; 2) 对自动匹配过程中出现的不确定性进行标识, 并交由专家确认, 从而实现对匹配过程的有效干预[14].

总体来看, 上述两种专家知识的引入仍主要集中在匹配算法运行过程当中或之后, 即专家知识的引入点由匹配算法决定. 实际上, 在数据迁移或集成工作中, 相关的领域专家在匹配之前并非一无所知, 而是通常掌握本领域一定量的匹配知识, 即部分元素匹配或不匹配关系, 在面对新的匹配任务时, 仅需在此基础上对未知的匹配关系进行推理, 最终完成全局匹配. 可见, 已有匹配方法并没能对上述信息进行有效运用, 因而需要反复运行较为复杂的匹配运算或在匹配过程中引入较多的专家干预. 如果一个匹配方法在匹配之前即能获得部分专家知识, 并以此为基础指导匹配方法的运行, 进而减少匹配中的不确定性, 那么该方法不仅有利于减少专家的干预量, 而且其匹配质量极有可能高于没有专家知识指导的模式匹配方法.

为此, 本文提出一种利用专家知识来辅助匹配的模式匹配模型. 区别于以往专家知识的引入方式, 本文试图在匹配的初始阶段即输入少量专家信号, 将部分匹配关系的验证工作转移到全局匹配之前, 进而从中提取用于辅助全局匹配的先验知识, 使匹配系统在具有某些知识的前提下执行匹配操作. 与已有匹配方法相比, 该操作并没有增加专家的工作量, 因为任何匹配方法所得结果的正确性都需要专家进行确认. 具体地, 该模型首先基于元素的名称信息构建待匹配模式元素间的初始相似度矩阵, 同时引入少量专家信号对矩阵中的部分匹配关系进行修正, 并以此为基础推断并扩展已知的匹配关系及对匹配器的准确性进行评估; 然后, 利用已收集到的知识指导不同匹配器结果的合并、调整及优化. 人工数据集和真实数据集中的实验结果表明, 在全局匹配之前, 少量专家知识的引入可有效提高模式匹配的准确性.

本文其余部分安排如下: 第 1 节介绍模型的整体框架; 第 2 节介绍基于专家验证的先验知识收集过程; 第 3 节介绍基于先验知识的模式匹配方法; 第 4 节进行对比实验分析; 第 5 节总结全文.

1　模型的整体框架

模式匹配操作具有内在不确定性, 实践证明, 完全自动化的模式匹配方法并不适用[12]. 为了有效利用专家知识提高模式匹配质量, 本文提出一种基于部分已验证匹配关系的模式匹配模型 (Partial verified matching relations based matching model, PVMM). 该模型的主要思想是: 匹配之前, 尽可能地利用已获取到的专家知识, 对未知的匹配关系进行推理, 以发现更多有价值的信息, 在最大程度上辅助后续匹配操作. 该模型由先验知识收集模块 (Priori knowledge collecting module, PKCM) 和基于先验知识匹配模块 (Priori knowledge based matching module, PKMM) 两部分组成. 其中, PKCM 模块用于完成匹配前先验知识的收集工作, 如图 1 所示.

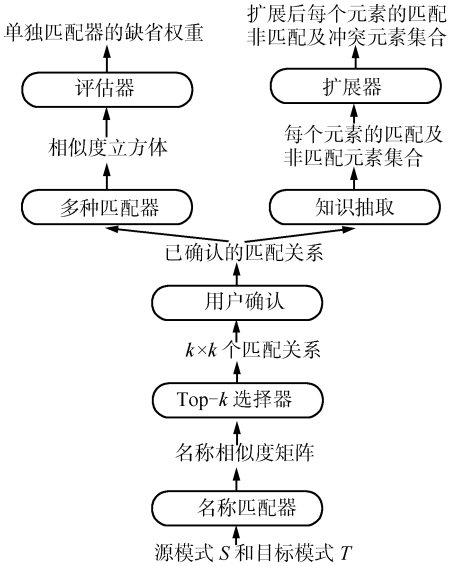


图 1　PKCM 的整体框架

Fig. 1　Overall structure of PKCM

PKCM 模块具体执行步骤如下: 1) 利用名称匹配器计算模式 S 和 T 间的初始相似度矩阵 IM_{ST}; 2) 利用 Top-k 选择器生成 IM_{ST} 中少量元素的对应关系列表 $\mathrm{Top}k(S,T)$; 3) 专家对 $\mathrm{Top}k(S,T)$ 中的元素进行确认, 指出 $\mathrm{Top}k(S,T)$ 中存在的匹配与不匹配关系; 4) 基于专家已验证的信息, 挖掘 $\mathrm{Top}k(S,T)$ 中的隐含知识, 为模式 S 和 T 中每个元

素 e 生成其匹配元素集合 Match(e)、非匹配元素集合 Mismatch(e) 和冲突元素集合 Conflict(e); 5) 利用多种匹配器对 S, T 中已确定的对应关系进行匹配, 分析每种匹配器结果的准确程度, 估计其在当前任务中的缺省权重 w.

PKMM 模块利用已收集到的先验知识对模式 S 和 T 进行重新匹配, 如图 2 所示. 该模块具体执行步骤如下: 1) 利用 PKCM 中已评估过的多种匹配器对 S 和 T 进行匹配, 并构建一个 $l \times m \times n$ 的相似度立方体 Cube$_{ST}$, 其中, l 表示匹配器的数量, m, n 分别表示模式 S 和 T 所包含的元素数量; 2) 相似度合并器以 Cube$_{ST}$、当前任务中单独匹配器的缺省权重 w、每个元素的 Match、Mismatch 和 Conflict 集合作为输入, 将 Cube$_{ST}$ 合并为相似度矩阵 M_{ST}; 3) 结构化推理器利用元素间相似性传递的性质对 M_{ST} 进行优化, 在全局范围内对 M_{ST} 中的相似度值进行调整; 4) 选择性评估器对优化后的 M_{ST} 的特征进行分析, 并给出每个元素的候选匹配集合.

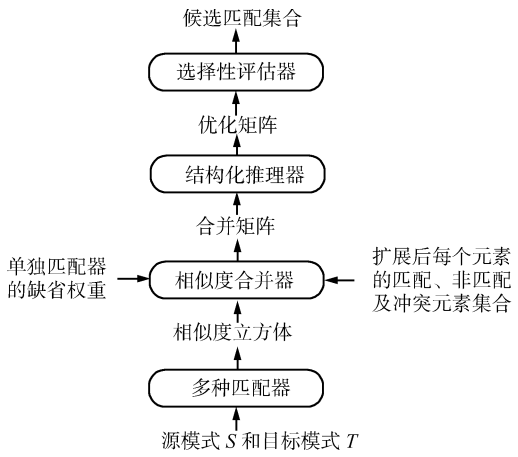


图 2 PKMM 的整体框架

Fig. 2 Overall structure of PKMM

2 基于专家验证的先验知识收集过程

为了在匹配初始阶段即有效引入专家知识, 进而减少匹配过程中可能出现的不确定性和专家的干预量. 本文首先选取待匹配模式间部分元素的对应关系用于专家确认, 然后, 对已确认的对应关系进行扩展, 以收集用于辅助全局匹配的先验知识.

2.1 部分对应关系的选取

对于待匹配模式间部分元素对应关系的选择, 一种方法是采用随机采样的方式, 即随机选取两个模式间的若干元素对, 但该方法的缺点在于可能导致所选取的对应关系都不匹配, 与存在匹配关系相比, 所有对应关系都不匹配时, 仅能提供少量信息. 因此, 为了获取更多有价值的信息, 本文通过计算待匹配模式间的 k 个最优匹配关系, 进而确定用于专家验证的对应关系集合.

对于待匹配模式 S 和 T, 首先, 利用名称匹配器构建两个模式元素间的初始匹配关系. 一般来说, 良好的命名方式可充分标识元素的语义, 但实际应用中元素的名称往往由特殊字符、常见词 (介词、连词) 构成. 因此, 匹配前需要对名称做规范化处理, 包括缩写词扩展、去除特殊字符、常见词等, 从而将其转化为词条向量的形式.

对于词条向量 (w_{11}, w_{12} ,$\cdots$, w_{1n}), (w_{21}, w_{22}, $\cdots$, w_{2m}), 任意词条 (w_{1i}, w_{2j}) 间的相似度 (Token similarity, Tsim(w_{1i}, w_{2j})) 可基于 3-gram 定义为

$$\frac{|2 \times \sum\limits_{t} \log p(t)|}{|\sum\limits_{t_1} \log p(t_1)| + |\sum\limits_{t_2} \log p(t_2)|} \tag{1}$$

其中, $t \in$ 3-gram(w_{1i}) $\cap$ 3-gram(w_{2j}), $t_1 \in$ 3-gram(w_{1i}), $t_2 \in$ 3-gram(w_{2j}); 3-gram(w_{1i}), 3-gram(w_{2j}) 分别表示 w_{1i}, w_{2j} 划分后长度为 3 的子串集合; $p(t)$, $p(t_1)$, $p(t_2)$ 分别表示子串 t, t_1, t_2 在词条 w_{1i} 和 w_{2j} 中出现的概率.

考虑到不同词条在整个模式中的重要程度不同, 本文采用信息检索领域的 TF/IDF (Term frequency-inverse document frequency) 算法对每个词条的权重进行度量. 令单个词条向量构成文档 d, 所有词条向量构成文档集 D, 则词条 i 在文档 d 中的权重 W_i 可定义为

$$W_i = \mathrm{tf}_{id} \times \mathrm{idf}_{iD} = \frac{\mathrm{fre}_{id}}{\max_q(\mathrm{fre}_{qd})} \times \log\left(\frac{|D|}{|D_i|}\right) \tag{2}$$

其中, tf_{id} 衡量词条 i 在文档 d 中的重要性, 而 idf_{iD} 则是对词条 i 在整个文档集 D 中的重要性进行度量, fre_{id} 表示词条 i 在文档 d 中出现的频率, $\max_q(\mathrm{fre}_{qd})$ 表示文档 d 中出现频率最高的词条 q 的频率. 由于元素名称总是由较少的词条构成, 且每个词条出现的频率均为 1, 因此, 在计算权重时可忽略对 tf_{id} 的考虑, 并将权重计算公式定义为[15]

$$W_i = \mathrm{idf}_{iD} = \frac{\log\left(\dfrac{|D|}{|D_i|}\right)}{\log\left(\dfrac{|D|}{1}\right)} \tag{3}$$

基于单词条间的相似度及词条在模式中的重要

性, 元素 s, t 的名称相似度 $\mathrm{Nsim}(s,t)$ 可定义为

$$\mathrm{Nsim}(s,t)=\frac{1}{2}\left[\frac{\sum_{i=1}^{n}W_{1i}W_{2c_i}\mathrm{Tsim}(w_{1i},w_{2c_i})}{\sum_{i=1}^{n}W_{1i}W_{2c_i}}+\frac{\sum_{j=1}^{m}W_{1r_j}W_{2j}\mathrm{Tsim}(w_{1r_j},w_{2j})}{\sum_{j=1}^{m}W_{1r_j}W_{2j}}\right] \tag{4}$$

其中, $c_i=\arg\max_{1\le j\le m}\mathrm{Tsim}(w_{1i},w_{2j})$ 表示当 i 固定时, 提取使得 $\mathrm{Tsim}(w_{1i},w_{2j})$ 值最大的 j; 同理, $r_j=\arg\max_{1\le i\le n}\mathrm{Tsim}(w_{1i},w_{2j})$ 表示当 j 固定时, 提取使得 $\mathrm{Tsim}(w_{1i},w_{2j})$ 值最大的 i.

假设模式 S 含有 m 个元素, 模式 T 含有 n 个元素, 上述名称匹配器将返回一个大小为 $m\times n$ 的名称相似度矩阵 IM_{ST}, 我们可基于 IM_{ST} 生成 k 最优匹配关系. 对于 IM_{ST} 中 k 个最优匹配关系的选择通常可采用如下策略[4]: MaxN、Threshold、MaxDelta. 其中, MaxN 将选取相似度值最高的 N 个关系; Threshold 将选取相似度大于阈值 α 的关系; MaxDelta 将选取与相似度最大值差值小于 d 的关系. 上述策略虽然能够保证选取当前 k 个最优匹配关系, 但可能导致所选取的对应关系均包含同一源模式元素或目标模式元素, 从而降低信息含量. 因此, 本文将 MaxN 与稳定婚姻法[3] 策略相结合, 即选取的候选匹配应满足以下三个条件:

1) 集合中所有元素的相似度之和最大;

2) 不存在以下元素对 (x_1,y_1) 和 (x_2,y_2), 其中, (x_1,y_2) 的名称相似度大于 (x_1,y_1) 的名称相似度, 同时 (x_2,y_1) 的名称相似度也大于 (x_1,y_1) 的名称相似度;

3) 相似度值最高的 k 个元素对.

相对于其他选择策略, 该方法能够保证每个元素都具有唯一匹配, 同时返回相似度值最高的 k 个元素对. 对于 k 的取值, 显然较大的 k 值可以保留更多的候选匹配, 提供更多有价值的信息, 但随着 k 值的增大, 所需的专家确认工作量也会随之增加. 当 k 达到最大值, 即等于模式 S 或 T 中所含元素的最小值时, 此时等价于由专家完成整个匹配任务. 因此, k 的取值可由专家根据待匹配模式所包含元素数量的多少进行调整, k 的缺省值为 3.

2.2　基于专家验证的部分已知匹配关系提取

k 个最优匹配关系确定后, 我们将得到一个包含 $k\times k$ 个元素的列表 $\mathrm{Top}k(S,T)$, 该列表不仅包含 k 个最优匹配, 还将包含最优匹配所涉及元素间的对应关系. 对于 $\mathrm{Top}k(S,T)$ 中的元素, 专家可首先对 k 个最优匹配进行确认, 若推荐的 k 个最优匹配全部正确时, 则对其进行标识, 并无需再对其它对应关系进行确认, 因为此时 k 个元素的候选匹配已全部找到; 若专家认为某一推荐的最优匹配不正确时, 此时, 还需对 $\mathrm{Top}k(S,T)$ 中与该匹配相关联的剩余元素进行确认, 如果在剩余元素中也未发现匹配关系, 则表明该匹配所涉及元素的候选匹配不在该列表内. 由于 $\mathrm{Top}k(S,T)$ 中的元素数量较少, 因此专家只需花费少量的时间即可完成确认工作.

图 3 给出了模式 S 和 T 之间的部分已确认匹配关系示例. 假设基于名称相似度矩阵 IM_{ST}, 模型返回以下 Top$k$ 列表: $\{(a_2,b_1),(a_3,b_3),(a_4,b_4),(a_2,b_3),(a_3,b_3),(a_4,b_3),(a_2,b_4),(a_3,b_4),(a_4,b_4)\}$, 即图中灰色区域. 其中, (a_2,b_1), (a_3,b_3), (a_4,b_4) 为 k 个最优匹配, 由稳定婚姻法和 Max3 选择得到, 即图中黑色矩形区域, 剩余元素为与 k 个最优匹配所含元素相关联的匹配关系. "×" 号区域表示专家确认元素 (a_2,b_1), (a_4,b_4) 不匹配, 且在与上述元素相关联的元素 (a_2,b_4), (a_4,b_1) 中也没有发现匹配关系; "√" 号区域表示专家确认元素 (a_3,b_3) 相匹配; 灰色空白区域表明由于元素 a_3 和 b_3 的匹配元素已找到, 因此, $\mathrm{Top}k(S,T)$ 中与其相关联的元素 (a_3,b_1), (a_3,b_4), (a_2,b_3), (a_4,b_3) 不可能匹配; "?" 号区域表示目前该元素的匹配关系未知.

T \ S	a_1	a_2	a_3	a_4	a_5	a_6
b_1	?	×		×	?	?
b_2	?	?	?	?	?	?
b_3	?		√		?	?
b_4	?	×		×	?	?
b_5	?	?	?	?	?	?
b_6	?	?	?	?	?	?

图 3　模式 S 和 T 的部分已确认匹配关系示例

Fig. 3　Example of partial verified matching relation diagram between schema S and T

在匹配基数为 1:1 的情况下, 基于已验证的匹配关系, 我们可得到现有条件下模式 S 和 T 中每个元素的匹配元素集合 Match 及非匹配元素集合 Mismatch. 由于 $\mathrm{Top}k(S,T)$ 中既可能存在匹配关系又可能所有元素都不匹配, 因此, 需要对上述情况进行分别讨论.

1) 若 $\mathrm{Top}k(S,T)$ 中存在匹配关系, 假设元素 (s,t) 匹配时, 令 $\mathrm{sim}(s,t)=1$, 则有 $\mathrm{Match}(s)=\{t\}$, $\mathrm{Mismatch}(s)=\{t_i\mid t_i\in T\wedge t_i\neq t\}$; $\mathrm{Match}(t)=\{s\}$, $\mathrm{Mismatch}(t)=\{s_i|s_i\in S\wedge s_i\neq s\}$. 对于 $\mathrm{Top}k(S,T)$ 中除 s, t 以外的模式元素 s_i 和 t_j, 则有 $\mathrm{Match}(s_i)=\{\mathrm{unknown}\}$, $\mathrm{Mismatch}(s_i)=\{t_j|(s_i,t_j)\in\mathrm{Top}k(S,T)\}$;

$\mathrm{Match}(t_j) = \{\mathrm{unknown}\}$, $\mathrm{Mismatch}(t_j) = \{s_i|(s_i,t_j) \in \mathrm{Top}k(S,T)\}$. 对于 $\mathrm{Top}k(S,T)$ 以外的模式元素 s 和 t, 则有 $\mathrm{Match}(s) = \{\mathrm{unknown}\}$, $\mathrm{Mismatch}(s) = \{t_j|(s_i,t_j) \in \mathrm{Top}k(S,T) \wedge \mathrm{sim}(s_i,t_j) = 1\}$, $\mathrm{Match}(t) = \{\mathrm{unknown}\}$, $\mathrm{Mismatch}(t) = \{s_i|(s_i,t_j) \in \mathrm{Top}k(S,T) \wedge \mathrm{sim}(s_i,t_j) = 1\}$.

2) 若 $\mathrm{Top}k(S,T)$ 中所有对应关系都不匹配时, 即对于 $\mathrm{Top}k(S,T)$ 中的任一模式元素 s 和 t, 令 $\mathrm{sim}(s,t) = 0$, 则有 $\mathrm{Match}(s) = \{\mathrm{unknown}\}$; $\mathrm{Mismatch}(s) = \{t_i|(s,t_i) \in \mathrm{Top}k(S,T)\}$; $\mathrm{Match}(t) = \{\mathrm{unknown}\}$, $\mathrm{Mismatch}(t) = \{s_i|(s_i,t) \in \mathrm{Top}k(S,T)\}$. 对于 $\mathrm{Top}k(S,T)$ 以外的模式元素 s 和 t, 则有 $\mathrm{Match}(s) = \{\mathrm{unknown}\}$, $\mathrm{Mismatch}(s) = \{\mathrm{unknown}\}$, $\mathrm{Match}(t) = \{\mathrm{unknown}\}$, $\mathrm{Mismatch}(t) = \{\mathrm{unknown}\}$.

由图 3 可以看出, 在最坏的情况下, 当 $\mathrm{Top}k(S,T)$ 中所有元素都不匹配时, 该集合提供的信息量较少 (仅包含集合内部元素间的匹配关系); 当少量匹配对存在时, 便可提供较大的信息量 (除集合内部元素间的匹配关系外, 还包括其他元素的部分匹配关系), 如 (a_1,b_3), (a_5,b_3), (a_6,b_3).

2.3 部分已知匹配关系的扩展

基于图 3 中已提取到的匹配关系, 可对 "?" 区域做进一步推理, 从而挖掘出更多的先验知识. 观察图 3 中的 "×" 号部分, 例如元素 (a_4,b_1) 和 (a_4,b_4). 由于已确认其不匹配, 因此, 说明元素 b_1 和 b_4 在某些方面一定与元素 a_4 存在冲突 (如数据类型不同等). 如果我们能够在模式 T 的剩余元素 $\{b_2,b_5,b_6\}$ 中找到这样一个元素 e, 它与 a_4 具有相同的冲突, 则该元素与 a_4 可能也不相似.

文献 [16] 对属性间的语义冲突进行了阐述, 并将其归结为命名 (Name) 冲突、默认值 (Default value) 冲突和约束 (Constraint) 冲突. 其中, 约束冲突又可划分为数据类型 (Datatype) 冲突、精度 (Precision) 冲突等. 显然, 若属性之间存在上述语义冲突, 数据转换将失败, 因此, 基于上述语义冲突, 我们对 b_1、b_4 及 a_4 所具有的特征进行描述, 并对属性间的语义冲突进行标识.

对于不同模式中的元素, 通常可采用名称作为特征来进行区分, 但在同一模式中, 由于元素名称不可能出现重名, 因此, 该特征并不适合作为区分 b_1、b_4 的特征. 本文对文献 [6] 中提到特征进行了筛选, 并采用表 1 给出的 4 类特征对元素进行描述, 这些特征均可由数据库系统表或数据实例得到.

表 1 元素的特征

Table 1 Features of the elements

序号	特征类型	特征描述
1	数据类型	用于区分元素的数据类型 如 String、Numeric、Date
2	取值范围	用于区分元素的取值范围 如长度、精度等
3	约束	用于区分元素的约束条件 如值是否唯一、允许为空等
4	数据内容	用于区分元素所包含的数据 特征, 如平均值、标准差等

由于所选取的特征隶属于不同的域, 为了便于计算, 可采用以下方式将其值域映射到 [0, 1] 范围内[6]:

1) 二元值: 特征位取值为 True/False (如是否存在主键), 则调整对应值分别为 1 和 0, 若该特征允许出现空值, 则对应值为 0.5.

2) 类别值: 特征位取值由多种类别构成 (如数据类型), 则调整其所属的类别值为 1, 其余类别值为 0.

3) 范围值: 特征位取值为离散值和连续值, 则采用函数 $f(x) = 1/(1+k^{-x})$ 将其映射到 [0, 1] 范围内, 通过调整 k 值的大小以实现对不同范围值的缩放.

经上述处理后, 每个模式元素将抽象成一个特征向量的形式. 本文采用以下方式对元素特征向量间的冲突进行标识:

1) 对于已知不匹配的元素, 若特征位取值相同, 则将该特征位的冲突标记为 0, 否则标记为 1.

2) 对于匹配关系未知的元素, 若特征位取值为二元值时, 可采用 1) 中的方法进行标识. 若特征位取值为范围值时, 此时需与已有不匹配关系在该特征位的取值进行对比, 若特征位的取值大于已有不匹配关系中该特征位的最小值, 则表明该特征位可能是造成元素不匹配的一个因素, 冲突标记为 1, 否则标记为 0.

3) 对于任意两个特征向量 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 间的冲突差异, 采用式 (5) 进行估计:

$$\mathrm{dis}(\boldsymbol{v}_1,\boldsymbol{v}_2) = 1 - \frac{\sum\limits_{k=1}^{n} v_{1k}v_{2k}}{\sqrt{\left(\sum\limits_{k=1}^{n} v_{1k}^2\right)\left(\sum\limits_{k=1}^{n} v_{2k}^2\right)}} \tag{5}$$

本文采用算法 1 对部分元素的已知匹配关系进行扩展, 给出了模式 X 中每个元素 x 的冲突元素集合 $\mathrm{Conflict}_X(x)$. 算法以模式 S、T 的所有元素集合 X 和 Y, X 和 Y 对应的特征向量集合 V_X、V_Y 及 X 中每个元素的非匹配元素集合所构成的集合

Mismatch_X 作为输入. 对于 X 中的元素 x, 步骤 2 将逐一标识 x 与 $\text{Mismatch}(x)$ 中每个元素 e 在特征向量上的差异 d, 并将 d 添加到差异向量集合 D 中. 对于 Y 中的元素 y, 若其不在 $\text{Mismatch}(x)$ 中, 步骤 4～6 计算 y 与 x 在特征向量上的差异 d'. 并判断 d' 是否存在于 D 中, 若 D 中存在 d' 或存在包含 d' 的差异向量, 则计算 y 与 x 的差异值 v, 并将 (y, v) 添加到 x 的冲突元素集合 $\text{Conflict}_X(x)$ 中.

算法 1. ExpandRelation ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{V_X}$, $\boldsymbol{V_Y}$, $\textbf{Mismatch}_{\boldsymbol{X}}$)

输入. 模式 S 和 T 中的所有元素集合 X 和 Y, X 和 Y 对应的特征向量集合 $\boldsymbol{V}_X$, $\boldsymbol{V}_Y$, X 中每个元素的非匹配元素集合所构成的集合 Mismatch_X.

输出. X 中每个元素的冲突元素所构成的集合 Conflict_X.

步骤 1. 从 X 中任取元素 x, 若 $\text{Mismatch}(x) \neq \{\text{unkown}\}$, 从 X 中去除 x, 转入步骤 2; 否则, 转入步骤 8.

步骤 2. 从 $\text{Mismatch}(x)$ 中逐一选取元素 e, 采用方法 1) 计算特征向量 $\boldsymbol{V}_X(x)$ 与 $\boldsymbol{V}_Y(e)$ 间的差异向量 $\boldsymbol{d}$, 并添加 $(x, \boldsymbol{V}_X(x), e, \boldsymbol{V}_Y(e), \boldsymbol{d})$ 到差异集合 D 中.

步骤 3. 从 Y 中任取元素 y, 若 $y \notin \text{Mismatch}(x)$, 从 Y 中去除 y, 转入步骤 4; 否则, 转入步骤 7.

步骤 4. 采用方法 2) 计算 y 与 x 之间的特征向量差异 $\boldsymbol{d}'$.

步骤 5. 若 $\exists\, (x, \boldsymbol{V}_X(x), e, \boldsymbol{V}_Y(e), \boldsymbol{d}) \in D$ 且差异向量 $\boldsymbol{d} = \boldsymbol{d}'$ 或 $\boldsymbol{d}$ 包含 $\boldsymbol{d}'$, 转入步骤 6; 否则, 转入步骤 7.

步骤 6. 采用方法 3) 计算 $V_X(x)$ 和 $V_Y(y)$ 的差异值 v, 并添加 (y, v) 到 x 的冲突元素集合 $\text{Conflict}_X(x)$.

步骤 7. 若 $Y \neq \emptyset$, 转入步骤 3.

步骤 8. 添加 unkown 到 x 的冲突元素集合 $\text{Conflict}_X(x)$, 若 $X \neq \emptyset$, 清空 D 中的元素, 转入步骤 1.

步骤 9. 返回 Conflict_X.

当 X 中的所有元素扩展完成后, 需要对 Y 中的元素执行相同操作, 从而标识出 Y 中每个元素 y 的冲突元素集合 $\text{Conflict}_Y(y)$. 上述收集到的确定及扩展知识将用于相似度矩阵的合并与优化阶段, 即第 3.1 节, 以提高合并后所得相似度值的准确性.

2.4 单独匹配器的准确性评估

目前, 利用多种匹配器进行组合匹配被视为一个有效方式, 其中, 单独匹配器的选择及结果的合并通常由专家根据经验进行设定, 但由于所面临的匹配问题不同, 可能导致专家对每种匹配器的性能并不能做出较好的判断. 为此, 本文基于对应关系已知的元素, 给出一种用于合并单独匹配器结果的简单策略. 在针对不同匹配任务时, 该方法通过对单独匹配器的准确性进行评估, 从而为专家对单独匹配器的选择和加权提供参考依据.

对于对应关系已知的元素, 其相似度值可划分为 0 和 1 两类, 其中, 1 表示匹配关系, 0 表示不匹配关系. 可采用多种匹配器对上述元素进行匹配. 令匹配器 l 对元素 i 的相似度估计值为 sim_i, 而已知该元素的相似度值为 sim'_i, 则匹配器 l 在计算元素 i 的相似度时, 所产生的计算误差 Error 可定义为

$$\text{Error}_i = |\text{sim}_i - \text{sim}'_i| \tag{6}$$

同理, 对于对应关系已知的所有元素, 可定义匹配器 l 的平均误差 AvgError_l 为

$$\text{AvgError}_l = \frac{1}{n}\sum_{i=1}^{n}\text{Error}_i \tag{7}$$

式 (7) 表明, 匹配器 l 对所有对应关系已知的元素估计值接近已知值时, 匹配器 l 的平均误差将达到最小值. 基于式 (7), 匹配器 l 的准确性可定义为

$$\text{Accuracy}_l = 1 - \text{AvgError}_l \tag{8}$$

可以看出上述方法的准确性与对应关系已知的元素数量密切相关. 当 $\text{Top}k(S, T)$ 中存在匹配关系时, 由第 2.2 节可知, 对应关系已知的元素数量将增加, 此时单独匹配器的准确性评估要明显优于 $\text{Top}k(S, T)$ 中不存在匹配关系的情况.

3 基于先验知识的模式匹配方法

上述先验知识收集完成后, 模型将对模式 S 和 T 执行全局匹配. 为了综合利用模式信息, 可考虑组合多种类型匹配器, 如名称类匹配器、约束类匹配器、结构类匹配器等. 具体匹配器的选择可首先采用第 2.4 节给出的方法对其进行评估, 然后, 选取准确性较高的匹配器. 对于由不同匹配器返回的相似度矩阵所构成的相似度立方体 Cube_{ST}, 本节将利用已有的先验知识对其进行合并及优化, 并生成每个元素的候选匹配集合.

3.1 相似度矩阵的合并与优化

假设采用 l 种匹配器对模式 S 和 T 进行匹配, 匹配器 i 生成的相似度矩阵为 M_i $(1 \leq i \leq l)$. 基于 2.4 节所获取的每种匹配器准确性估计值, 对上述矩阵进行加权合并, 生成矩阵 M_{ST}, 如式 (9):

$$M_{ST} = w_1 M_1 + w_2 M_2 + \cdots + w_l M_l \tag{9}$$

其中, w_i $(1 \le i \le 3)$ 表示为不同匹配器所赋予的权重.

由于模式 S 和 T 的部分元素匹配关系已经确定, 因此, 应对 M_{ST} 中上述元素的对应值进行修正. 本文基于第 2.2 节和第 2.3 节所取得的每个元素的 Match、Mismatch、Conflict 集合对 M_{ST} 中的相似度值进行调整.

首先, 对 M_{ST} 进行归一化处理, 并对模式 S 和 T 中的任意元素 s 逐一执行以下操作步骤:

步骤 1. 若 Match(s) ≠{unkown}, 则对于任意元素 $e \in$ Match(s), 调整 $M_{ST}(s,e)$ 的值为 1, 与 s 同行或同列的其他矩阵值调整为 0; 若 Match(s) = {unkown}, 执行步骤 2;

步骤 2. 若 Mismatch$(s) \neq$ {unkown}, 则对于任意元素 $e \in$ Mismatch(s), 调整 $M_{ST}(s,e)$ 的值为 0, 与 s 同行或同列的其他矩阵值保持不变, 执行步骤 3; 若 Mismatch(s) = {unkown}, 执行步骤 3;

步骤 3. 若 Conflict$(s) \neq$ {unkown}, 则对于任意元素 $(e,v) \in$ Conflict(s), 考虑到特征冲突对相似度的影响程度, 本文对 v 的值域进行缩小, 将其映射到 [0, 0.2] 区间内, 并调整 $M_{ST}(s,e)$ 值为 $M_{ST}(s,e) - (1/(1+k^{-v}) - 0.5)$, 其中, $k = 2.3$.

上述操作完成后, M_{ST} 中的部分元素值将得到调整. 为了更直观地说明本文后续思想, 采用二部图的形式对矩阵 M_{ST} 进行表示. 图 4 给出了 M_{ST} 所对应的二部图, 图 4 中灰色结点表示模式 S 所包含的元素, 白色结点表示模式 T 所包含的元素, 元素之间边的权值与矩阵中的值相对应. 为了清晰起见, 图中省略了部分边的权值和元素.

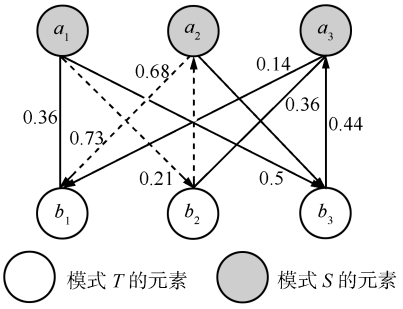


图 4 模式 S 和 T 间部分元素关系图

Fig. 4 Part of relation diagram between schema S and T

以元素 a_1 与 b_1 为例, 其相似度值除自身信息决定外, 还应考虑与其相关联元素 $\{(a_1,b_2),(b_2,a_2),(a_2,b_1)\}$ 所反映出 (a_1,b_1) 的相似性 (图 4 中虚线部分), 即当 $a_1 \leftrightarrow b_2$, $b_2 \leftrightarrow a_2$, $a_2 \leftrightarrow b_1$ 分别相似时, 由传递性可知, $a_1 \leftrightarrow b_1$ 在某种程度上也相似.

基于上述性质, 对于元素 (a_i,b_j), 本文将与其相关联的元素 $\{(a_i,b_m),(a_n,b_m),(a_n,b_j)\}$ 至少所反映出来的 (a_i,b_j) 的相似程度定义为

$$\mathrm{sim}(a_i,b_m)\mathrm{sim}(a_n,b_m)\mathrm{sim}(a_n,b_j) \tag{10}$$

基于式 (10), 元素 (a_i,b_j) 调整后的相似度 $\mathrm{sim}'(a_i,b_j)$ 可定义为

$$\begin{aligned}\mathrm{sim}'(a_i,b_j) = {} & \alpha\mathrm{sim}(a_i,b_j)+\\ & (1-\alpha)\sum \mathrm{sim}(a_i,b_m)\mathrm{sim}(a_n,b_m)\mathrm{sim}(a_n,b_j)\end{aligned} \tag{11}$$

其中, $\mathrm{sim}(a_i,b_j)$ 表示初始相似度, $\mathrm{sim}'(a_i,b_j)$ 表示元素 (a_i,b_j) 由元素间传递关系调整之后的相似度值, 它由初始相似度 $\mathrm{sim}(a_i,b_j)$ 和 $\sum \mathrm{sim}(a_i,b_m)\mathrm{sim}(a_n,b_m)\mathrm{sim}(a_n,b_j)$ 加权得到, α 表示所赋予的权重, 取值 0.7.

由式 (11) 可以看出, 当 $\mathrm{sim}(a_i,b_j) \neq 0$ 且 $(a_i,b_m)(a_n,b_m)(a_n,b_j) \neq 0$ 时, a_i, b_j 的相似度值将受到元素 (a_i,b_m)、(a_n,b_m)、(a_n,b_j) 的影响. 当 $\mathrm{sim}(a_i,b_j) \neq 0$, 且 $(a_i,b_m)(a_n,b_m)(a_n,b_j) = 0$ 时, 即其中某一元素或某几个元素被确定为非匹配元素时, 该组合传递给 (a_i,b_j) 的相似度值为 0, 即 (a_i,b_j) 的相似度值不受该组合的影响; 当 $\mathrm{sim}(a_i,b_j) = 1$ 或 0 时, 由于已确定其为匹配或非匹配元素, 因此, 其他元素对其相似度的影响不需要计算.

考虑到所传递相似性的大小所起到的促进程度不同, 式 (11) 受到以下条件的限制, 即当 $\sum \mathrm{sim}(a_i,b_m)\mathrm{sim}(a_n,b_m)\mathrm{sim}(a_n,b_j) \ge \beta$ 时, 表明该元素集合对 a_i, b_j 的相似度起促进作用, 该值为正数; 当 $\sum \mathrm{sim}(a_i,b_m)\mathrm{sim}(a_n,b_m)\mathrm{sim}(a_n,b_j) < \beta$ 时, 表明该元素集合对 a_i, b_j 的相似度起到抑制作用, 该值为负数. 其中, β 取值 0.5.

重复上述调整过程, 经过有限次迭代后, 当每对元素前后两次迭代相似度值小于 ω 或达到固定步数 N 时, 则相似度调整过程结束.

3.2 选择方案评估及候选匹配生成

目前, 已有匹配方法在候选匹配生成阶段通常采用一种固定策略, 这显然不能适应所有匹配任务. 例如, 在某些情况下为每个元素生成多个候选匹配可能要比单一候选匹配更可取, 或者恰好相反. 为此, 本文基于文献 [17] 中的矩阵特征判别思想, 对 M_{ST} 的选择性进行评估, 进而确定当前最佳的候选匹配选择策略, 并采用相应的算法进行选择.

当 M_{ST} 中存在已确认的匹配元素时, 这些元素可从矩阵中被剔除, 从而降低计算复杂性. 令 M_{ST} 的剩余部分构成一个维数削减的子矩阵 M'_{ST}, a_ib_j 和 a_ib_k 分别表示矩阵 M'_{ST} 中, 第 i 行的最大值和次大值, 则矩阵 M'_{ST} 的行平均差值 Avgdiffer(M'_{ST})

可定义如下:

$$\frac{\sum_{i=1}^{n}(a_ib_j - a_ib_k)}{|\{(a_ib_j - a_ib_k)|(a_ib_j - a_ib_k) > 0\}|} \tag{12}$$

1) 当 Avgdiffer(M'_{ST}) 的值较高时, 结果暗示矩阵中每一行的最优值和次优值之间差距较大, 每个元素的最优匹配相对确定, 此时选择 Top-1 方案可获得较好的匹配结果. 如稳定婚姻法、Max1 等. 上述两种方法的选择可根据同一目标元素在所有源模式元素候选匹配集合中出现的数量进行区分.

2) 当 Avgdiffer(M'_{ST}) 的值较低时, 结果更多的暗示了最优匹配的不确定性, 此时, 应采用 Top-k 方案.

对于 Top-k 方案, 一般可采用 MaxN 策略进行选择, 即为源模式 S 中的元素 s 在目标模式 T 中选取与其相似度值最高的 N 个元素作为其候选匹配. 但该方法的缺陷在于: 当某一目标模式元素与多个源模式元素都具有较高的相似度值时, 该元素将出现在多个源模式元素的候选匹配集合当中. 显然, 该目标元素不可能与多数源模式元素都相似, 这种错误的选择将会造成其他正确的目标元素无法被选入候选匹配集合, 从而导致候选匹配质量不高[15]. 因此, 区别于以往的 MaxN 策略, 上述情况下的 Top-k 选择不仅需要对 MaxN 中的 N 值进行说明, 还需要对 N 中每个元素允许出现的总体次数 num 进行约束, 上述条件下的候选匹配生成问题可视为限制条件下的多任务–多人员分配问题, 该问题可采用最小费用最大流算法进行求解.

为了能够使用最小费用最大流算法生成候选匹配, 图 4 中的二部图还需要进一步扩展, 如图 5 所示.

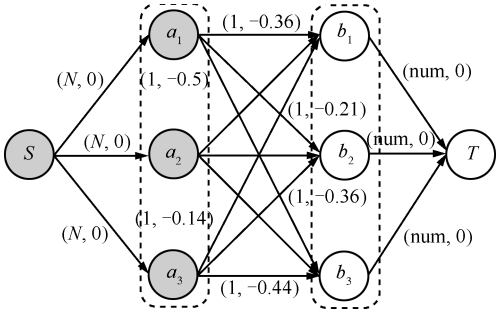


图 5　扩展后的模式 S 和 T 间部分元素关系图

Fig. 5　Part of expanded relation diagram between schema S and T

首先, 为二部图添加源结点 S 和目标结点 T, 其中 S 连接模式 S 的所有结点 a_i, T 连接模式 T 的所有结点 b_j; 其次, 为添加结点后的二部图添加流动方向, 流动方向由源结点 S 到目标结点 T; 最后, 为每条边添加 "流量" 和 "费用" 信息, 其中, 结点 S 到的所有结点 a_i 的流量和结点 T 到的所有结点 b_j 的流量分别为 N 和 num, 费用均为 0. 而结点 a_i 与 b_j 间的流量为 1, 费用则是矩阵 M'_{ST} 中的对应单元值 $M'_{ST}(a_i, b_j)$ 取负值. 算法的最终目标是找出从 S 到 T 并满足限制条件 N 和 num 的候选匹配元素集合.

4　实验分析

4.1　评价指标的选择

为了验证方法的有效性, 本节将对本文所提出的方法进行实验验证, 并采用模式匹配领域中的 Precision、Recall 和 Overall 三个评价指标来衡量匹配方法的质量. 令匹配方法返回的所有匹配结果为 P, 其中正确匹配结果和错误匹配结果分别为 T 和 F, 所有正确的匹配结果为 R, 则三个指标可定义为

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, 即:

$$\text{Precision} = \frac{T}{P} \tag{13}$$

2) 查全率 (Recall): 匹配结果中正确匹配结果占所有正确匹配的比率, 即:

$$\text{Recall} = \frac{T}{R} \tag{14}$$

3) 全面性 (Overall): 利用匹配算法所节省的工作量占总工作量的比率, 即:

$$\text{Overall} = \text{Recall} \times \left(2 - \frac{1}{\text{Precision}}\right) = \frac{T-F}{R} \tag{15}$$

其中, Precision 用于说明匹配方法的可信度, Recall 说明匹配方法覆盖正确匹配的范围, Overall 则是对上述两个指标的权衡.

4.2　人工数据集下的实验结果及分析

本节用于实验验证的测试数据集采用 http://metaquerier.cs.uiuc.edu/repository/上提供的数据模式, 这些模式大都来自于实际应用当中, 且被认为是模式匹配领域的公共测试数据集. 本文采用 ICQ Query Interfaces 数据集中提供的样例模式, 该数据集中的数据来自 5 个不同领域, 我们选取其中 4 个进行实验验证, 包括: Airfare、Auto、Book 和 Job. 每个领域所涉及的模式数量及平均元素个数如表 2 所示, 上述模式将通过 XML 格式重写的方式导入到数据库中. 由于模式缺少主/外键定义, 且部分模式数据实例较少, 因

此, 采用人工的方式对其进行补充, 并基于 Wordnet 本体对元素名称进行扩展. 4 个领域内部匹配结果的平均值将作为匹配方法在该领域的评价指标值.

表 2 人工数据集中的模式特征

Table 2 Features of schema in artificial data set

模式	Airfare	Auto	Book	Job
源模式数量	4	6	3	4
目标模式数量	1	1	1	1
平均元素个数	16	7	9	11

1) 基于名称和结构化信息的匹配性能

本实验的目的是为了观察在仅利用名称和结构化信息进行匹配时, PVMM 所能达到的性能. 匹配过程由三部分构成, 首先, 利用第 2.1 节中的名称相似度计算方法, 生成待匹配模式间的初始相似度矩阵, 随后利用第 3.1 节中的结构化推理算法对初始相似度矩阵进行调整, 最后, 采用稳定婚姻法策略生成每个元素的候选匹配集合.

实验结果如图 6 所示. 由图 6 可以看出, 在未使用部分已验证匹配关系的情况下, 只基于名称和结构化信息的匹配方法同样具有较好的表现, 此时的匹配思想类似于 SF 方法, 即通过名称匹配的方式计算元素间的语义相似度, 并由元素间相似性传递的性质对语义相似度在全局范围内进行调整.

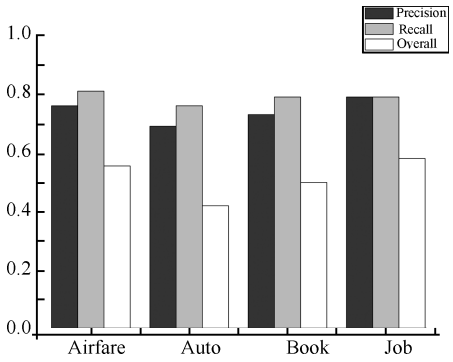


图 6 只基于名称和结构化信息的匹配性能

Fig. 6 Matching performance based on name and structured information

2) PVMM 与不同匹配方法间的性能对比

本节将 PVMM 与具有代表性的 SF、COMA 和 CUPID 方法进行对比分析. 上述三种方法代表了两种不同的匹配策略, 其中, COMA 方法利用多种类型匹配器单独执行匹配任务, 并使用不同组合策略对其结果进行合并; 而 CUPID 和 SF 方法则利用名称匹配器计算元素间的初始相似度, 并基于模式内部元素间的关联关系对其进行调整. 与上述方法相比, 本文在基于先验知识匹配阶段更倾向于二者策略的结合, 即首先采用多种匹配器进行匹配, 并对匹配结果进行合并及调整, 然后, 利用结构化调整算法对合并后的结果进行优化.

实验中, PVMM 将添加部分已验证匹配关系, 并采用组合匹配策略. 由于名称和数据类型对大多数模式来说比较普遍, 因此本文将第 2.1 节中的名称匹配器和 COMA 方法中的 EditDistance + DataType 匹配器进行组合. 其中, EditDistance + DataType 在已有匹配方法中较为常见, 该匹配器可以同时对名称和数据类型进行考虑. 然而, 在实际应用中, 匹配器的选择可根据当前所能获取到的信息进行确定. 上述匹配器的缺省权重将基于第 2.4 节计算得到, 实验结果如图 7～9 所示.

由图 7 可以看出, 在查准率方面 PVMM 要明显优于其他三种方法. 由查准率的定义可知, 其代表匹配结果中正确匹配结果占所有匹配结果的比率, 由于部分已验证匹配关系的引入, 正确匹配关系的数量得到保证, 剩余元素的相似度值也基于此先验知识得到了有效调整, 减少了发生错误匹配的可能性, 使匹配结果中正确匹配结果比例增加. 同时, CUPID 和 SF 在针对不同匹配任务时也显示出较高的准确性, 原因在于上述两种方法有效地利用了模式中的结构信息, 减少名称匹配器所得匹配结果中的不确定性. 而 COMA 虽然未使用结构信息, 但通过组合多种匹配器的方式, 使模式中蕴含的不同种类信息得到有效利用, 因此, 同样获得较好的准确性.

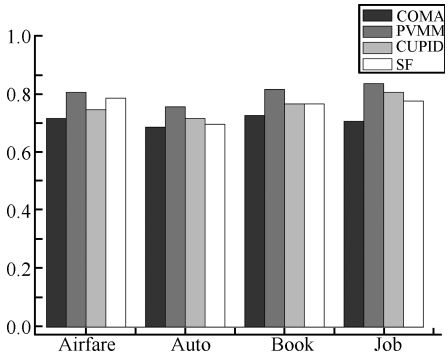


图 7 4 种匹配方法的查准率比较

Fig. 7 Precision comparison of four matching methods

图 8 给出了 4 种不同匹配方法在查全率方面的对比. 可以看出, PVMM 在查全率方面同样优于其他三种方法. 部分已验证匹配关系的运用, 有效提高了正确匹配关系被发现的可能性, 进而提高匹配结果的查全率. 与 PVMM、CUPID 和 SF 相比, COMA 在查全率方面表现一般, 主要是因为 COMA 的候选匹配选择策略采用 Threshold 方法, 虽然该方法能够保证相似度值较高的匹配对被返回, 但却无法保证每个元素都具有唯一匹配, 从而遗漏某些相似度值较低的正确匹配关系, 在 Auto 和 Book 领域该现象尤为明显.

理想情况下, 当所有返回的结果都为正确匹配

时, 即 $I = P = R$, $F = \emptyset$, 方法达到最优, 此时, Precision = Recall = Overall = 1. 由于查准率和查全率都具有一定片面性, 例如, 某些方法具有较高的查准率, 但其查全率却可能很低, 反之亦然. 因此, 使用单一的查准率和查全率并不能反映匹配方法的好坏, 仍需要对不同匹配方法的全面性进行对比分析. 图 9 给出了 4 种不同匹配方法的全面性对比, 由全面性的定义可知, 它反映了匹配方法所节省的工作量, 因此, 具有较高的全面性指标值, 则可以说明该匹配方法具有较高的可信度. 由图 9 可以看出, 本文所提出的方法在每种匹配任务中的全面性仍高于其他三种模式匹配方法. 此外, CUPID 和 SF 在全面性方面也要优于 COMA.

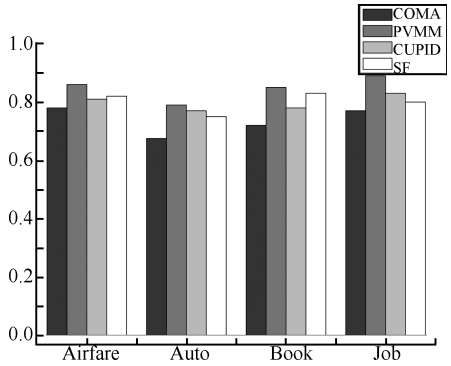


图 8　4 种匹配方法的查全率比较

Fig. 8　Recall comparison of four matching methods

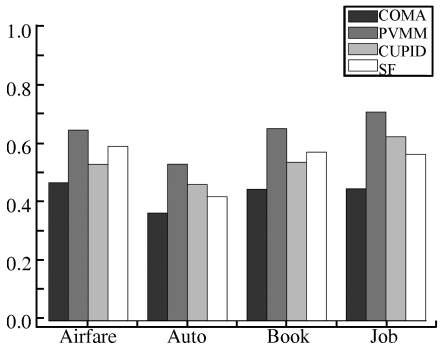


图 9　4 种匹配方法的全面性比较

Fig. 9　Overall comparison of four matching methods

4.3　真实数据集下的实验结果及分析

由第 4.1 节中的实验结果可以看出, 本文所提出的方法在性能上要明显优于其他三种方法. 为了进一步验证该方法的有效性, 本文选取实际项目中某两个地区社会保障信息系统底层数据库中的关系模式作为实验数据集, 该项目的目标是实现多地域社保信息系统底层数据库的集成, 从而为数据分析系统提供支持. 现将上述两个数据库记为 DB1 和 DB2, 其关系和属性的基本情况如表 3 所示.

表 3　DB1 和 DB2 中关系及属性

Table 3　Relations and attributes of DB1 and DB2

数据库	关系数量	属性数量
DB1	12	132
DB2	10	108

实验中, 同样采用 COMA、CUPID 和 SF 三种匹配方法与 PVMM 进行对比, 实验结果如图 10 所示. 可以看出, 在查准率、查全率方面, PVMM 要明显高于 COMA、CUPID 和 SF 方法, 因此可知, 4 种方法中, PVMM 方法所返回的结果集中正确匹配结果最多, 且正确匹配结果占返回结果比例最高. 在全面性方面, PVMM 同样要高于其他三种方法. 而与 COMA 相比, CUPID 和 SF 通过利用关系间的主/外键及关系和属性间的包含关系辅助名称匹配, 同样获得较好的匹配效果. 因此, 可以得出如下结论: 当模式结构信息可用时, 选择结构匹配器同样有助于提高模式匹配质量.

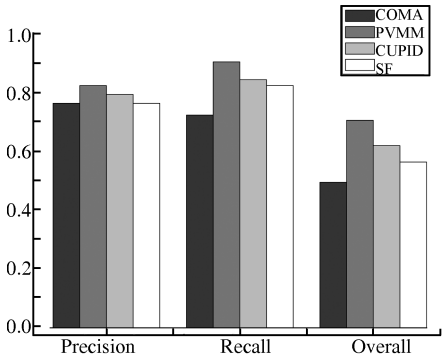


图 10　4 种匹配方法的性能比较

Fig. 10　Performance comparison of four matching methods

5　总结

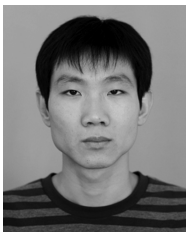
本文提出了一种基于部分已验证匹配关系的模式匹配模型. 它通过专家验证对部分对应关系的验证来获取已知的部分匹配关系及评估不同匹配器的缺省权重, 并利用获取到的先验知识来辅助匹配, 对未知的匹配关系进行推理, 从而完成全局匹配元素间相似度值的合并及优化. 实验表明本文所提出的方法在查准率、查全率和全面性方面都具有较好的表现. 相对于其他匹配方法来说, 本文将部分匹配结果的验证工作放于全局匹配之前, 这种方法不仅没有增加操作人员的工作量, 而且有助于后续匹配质量的提高.

References

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

2 Madhavan J, Bernstein P A, Rahm E. Generic schema matching with cupid. In: Proceedings of the 27th International Conference on Very Large Data Bases. San Francisco, USA: Morgan Kaufman Publishers, 2001. 49−58
3 Melnik S, Garcia-Molina H, Rahm E. Similarity flooding: a versatile graph matching algorithm and its application to schema matching. In: Proceedings of the 18th International Conference on Data Engineering. San Jose, California: IEEE, 2002. 117−128
4 Kang J, Naughton J F. Schema matching using interattribute dependencies. *IEEE Transactions on Knowledge and Data Engineering*, 2008, **20**(10): 1393−1407
5 Do Hong-Hai, Rahm E. COMA — A system for flexible combination of schema matching approaches. In: Proceedings of the 28th International Conference on Very large Data Bases. Hong Kong, China: VLDB, 2002. 610−621
6 Li W S, Clifton C. SEMINT: a tool for identifying attribute correspondences in heterogeneous databases using neural networks. *Data and Knowledge Engineering*, 2000, **33**(1): 49−84
7 Doan A, Domingos P, Halevy A Y. Reconciling schemas of disparate data sources: a machine-learning approach. In: Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data. Santa Barbara, USA: ACM, 2002. 509−520
8 Bilke A, Naumann F. Schema matching using duplicates. In: Proceedings of the 21st International Conference on Data Engineering. Tokyo, Japan: IEEE, 2005. 69−80
9 Zhang M H, Hadjieleftheriou M, Ooi B C, Procopiuc C M, Srivastava D. Automatic discovery of attributes in relational databases. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. New York, USA: ACM, 2011. 109−120
10 Shen De-Rong, Yu En-Yun, Zhang Xu, Kou Yue, Nie Tie-Zheng, Yu Ge. SKM: a schema matching model based on schema structure and known matching knowledge. *Journal of Software*, 2009, **20**(2): 327−338
(申德荣, 余恩运, 张旭, 寇月, 聂铁铮, 于戈. SKM: 一种基于模式结构和已有匹配知识的模式匹配模型. 软件学报, 2009, **20**(2): 327−338)
11 Li Guo-Hui, Du Xiao-Kun, Du Jian-Qiang. A structure matching method based on partial functional dependencies. *Chinese Journal of Computers*, 2010, **33**(2): 240−250
(李国辉, 杜小坤, 杜建强. 基于部分函数依赖的结构匹配方法. 计算机学报, 2010, **33**(2): 240−250)
12 Bellahsene Z, Bonifati A, Rahm E. *Schema Matching and Mapping*. Berlin: Springer-Verlag, 2011. 29−52
13 Aumueller D, Do H H, Massmann S, Rahm E. Schema and ontology matching with COMA++. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. New York, USA: ACM, 2005. 906−908
14 Wang G L, Goguen J, Nam Y K, Lin K. Critical points for interactive schema matching. In: Proceedings of the 6th Asia-Pacific Web Conference. Hangzhou, China: Springer-Verlag, 2004. 654−664
15 Peukert E, Rahm E. Restricting the overlap of top-n sets in schema matching. In: Proceedings of the 1st Workshop on New Trends in Similarity Search. New York, USA: ACM, 2011. 20−25
16 Kim W, Seo J. Classifying schematic and data heterogeneity in multidatabase systems. *IEEE Computer*, 1991, **24**(12): 12−18
17 Peukert E, Eberius J, Rahm E. Rule-based construction of matching processes. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2011. 2421−2424

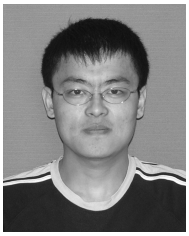
黄少滨 哈尔滨工程大学计算机科学与技术学院教授. 主要研究方向为分布式计算与仿真, 模型检测, 数据集成.
E-mail: huangshaobin@hrbeu.edu.cn
(**HUANG Shao-Bin** Professor at the College of Computer Science and Technology, Harbin Engineering University. His research interest covers distributed computing and simulation, model checking, and data integration.)



刘国峰 哈尔滨工程大学计算机科学与技术学院博士研究生. 主要研究方向为数据集成, 模式匹配, 数据挖掘. 本文通信作者.
E-mail: liuguofeng22@163.com
(**LIU Guo-Feng** Ph. D. candidate at the College of Computer Science and Technology, Harbin Engineering University. His research interest covers data integration, schema matching, and data mining. Corresponding author of this paper.)

万庆生 哈尔滨工程大学计算机科学与技术学院博士研究生. 主要研究方向为专家系统, 问答系统.
E-mail: wanqingsheng@hrbeu.edu.cn
(**WAN Qing-Sheng** Ph. D. candidate at the College of Computer Science and Technology, Harbin Engineering University. His research interest covers expert system and answering system.)

程 媛 哈尔滨工程大学计算机科学与技术学院博士研究生. 主要研究方向为数据挖掘, 机器学习.
E-mail: changuang7@sina.com
(**CHENG Yuan** Ph. D. candidate at the College of Computer Science and Technology, Harbin Engineering University. Her research interest covers data mining and machine learning.)



申林山 哈尔滨工程大学计算机科学与技术学院博士研究生. 主要研究方向为分布式计算, 自主计算.
E-mail: shenlinshan@hrbeu.edu.cn
(**SHEN Lin-Shan** Ph. D. candidate at the College of Computer Science and Technology, Harbin Engineering University. His research interest covers distributed computing and autonomic computing.)