面向多源关系数据的融合

丁玥^{1,2}, 王涓^{1,2}, 卢卫^{1,2*}, 荣垂田³, 杜小勇^{1,2}

1. 数据工程与知识工程教育部重点实验室 (中国人民大学), 北京 100872

2. 中国人民大学信息学院, 北京 100872

3. 天津工业大学计算机科学与技术学院, 天津 300387

* 通信作者. E-mail: lu-wei@ruc.edu.cn

收稿日期: 2019-08-16; 接受日期: 2019-09-30; 网络出版日期: 2020-04-23

北京市科技计划 (批准号: Z171100005117002) 和国家自然科学基金 (批准号: 61972403, 61872315, 61661146001) 资助项目

摘要 针对“信息孤岛”中的关系数据融合问题, 本文提出并实现了多源关系数据融合的基本框架 (multi-source relational data fusion, MSF). 框架包含 3 个主要模块: 模式匹配、实体对齐、实体融合. 模式匹配面向多源关系数据的属性对齐问题, 结合属性值的多维特征, 提出基于匈牙利 (Hungarian) 算法的属性间对齐发现机制, 实现了多源关系数据的快速模式匹配. 实体对齐连接多源关系中的元组对, 通过引入多样性取样策略和实体特征抽取方法, 提升了实体对齐的效果. 最后将对齐实体进行融合, 为数据分析提供统一的数据视图. 为了验证 MSF 的效果和效率, 实现了数据融合系统 DataPuzzle, 并在该系统上, 结合真实公开的多领域数据, 对提出的方法进行了验证. 结果表明, 所提出的方法可以高效地实现数据融合, 具有较高的查全率、查准率.

关键词 多源异构数据, 关系数据, 信息孤岛, 模式匹配, 实体对齐, 数据融合

1 引言

随着信息技术的发展, 政府、企业、医疗等众多领域的管理数据逐渐趋于信息化和网络化. 通过对海量数据进行挖掘与分析, 我们可以获取大量有价值的信息, 从而更好地释放数据价值. 然而, 无论是信息管理系统采集或管理的数据, 还是通过网络抓取到的公开数据, 都呈现出零散、异构、低质等碎片化特征. 不同部门的信息数据被分割存储, 数据之间组织形式各异, 相互独立、封闭, 无法有效地实现互连互通, 形成了数据治理的主要屏障——“信息孤岛”. 数据的管理与分析需要打破这些孤岛, 通过数据集成技术, 对碎片化数据进行有效融合.

数据集成技术主要是将自治、异构数据源中的数据进行逻辑或物理上的集中, 并向外提供统一的数据视图, 从而实现数据共享^[1,2]. 一般而言, 数据集成分为模式匹配、实体对齐、实体融合. 其中, 模

引用格式: 丁玥, 王涓, 卢卫, 等. 面向多源关系数据的融合. 中国科学: 信息科学, 2020, 50: 649–661, doi: 10.1360/SSI-2019-0172
Ding Y, Wang J, Lu W, et al. Multi-source relational data fusion (in Chinese). Sci Sin Inform, 2020, 50: 649–661, doi: 10.1360/SSI-2019-0172

式匹配主要用于查找两个或多个异构数据源之间的属性对应关系; 实体对齐是在模式匹配结果的基础上, 根据相互匹配属性中记录的特征, 将数据源中指代同一实体的记录连接起来. 但是, 由于数据质量问题的存在, 不同数据源在描述同一实体时可能存在属性值的冲突, 比如数据录入时的错误以及表达方式的不同等等. 因此, 作为数据集成的最终环节, 实体融合旨在消除不同数据源之间同一实体属性值冲突的问题, 将不同的数据信息进行综合, 从而提取出统一、丰富、高精度的数据.

目前数据集成方向的研究, 大都基于特定的场景. 比如 Pandey 等^[3] 基于语义 Web 技术, 提出了安全警示系统 Safety Check, 该系统能够对异构的警报、天气数据以及联系人信息进行集成, 可以识别和提醒在一定地理范围内可能遭受到人为或自然灾害的人. Qin 等^[4] 将数据集成技术运用于司法大数据, 介绍了“智慧法院”应用示范平台的架构以及应用效果. 这些数据集成系统的提出为不同领域的决策分析提供了数据支撑, 与上述系统不同, 本文从更为一般的应用场景出发, 分析并改进了模式匹配、实体对齐算法, 提出了一套更为通用的多源关系数据融合框架. 本文的主要贡献如下所述.

(1) 从模式匹配、实体对齐与实体融合的角度, 提出了多源关系数据融合的基本框架 (multi-source relational data fusion, MSF);

(2) 结合数据实例的多维特征, 提出基于匈牙利 (Hungarian) 算法的属性间对齐发现机制, 有效提高了模式匹配效率. 此外, 结合数据特异性, 提出结果排序算法, 用于给用户最佳匹配结果;

(3) 基于模式匹配结果, 引入多样性取样策略和实体特征抽取方法, 改进了实体对齐性能;

(4) 根据上述框架, 搭建了多源关系数据融合系统 DataPuzzle, 并在真实数据集上对系统的有效性进行了分析与测试.

2 相关工作

2.1 模式匹配

模式是对数据逻辑结构和特征的描述, 模式匹配主要用于发现并映射两个或多个异构数据源之间的属性对应关系^[5]. 在大数据背景下, 各数据源间结构和语义的异构性, 使得模式匹配的重要性不言而喻. 目前模式匹配的方法主要可以分为模式级别 (schema-level) 和实例级别 (instance-level)^[6]. 其中模式级别的方法借助于数据库中已有的模式信息来识别不同数据库中模式的对应关系, 包括语言层面、约束层面和结构层面^[7]. 语言层面利用元数据信息来发现模式之间的对应关系, 例如属性名、属性缩写、可获得的文字描述等等^[8~10]. 约束层面主要是指数据类型、取值范围、完整性约束等等^[7,10]. 结构层面是指数据库中数据表的内外模式信息以及模式属性的基数信息^[7,10]. SF (similarity flooding)^[11] 依据模式信息中的列属性以及数据类型, 进行多步迭代的定点计算得到匹配结果. 然而, 数据模式信息缺乏统一标准, 相同实体对象在不同数据库中通常具有不同的表达方式. 同时, 在特定数据场景中, 数据的模式信息难以获取, 信息量极少, 难以直接利用^[12]. 因此, 越来越多的研究开始转向实例级别的模式匹配方法, 从句法或语义层面, 对不同数据源的实例数据进行特征提取与比对分析, 从而获得模式元素之间的映射关系. SMDD (schema mapping method based on data distribution)^[13] 方法通过挖掘数据实例信息, 利用神经网络识别出具有相似分布规律的元素集合并进行相似度的计算, 最终将相似度较高的匹配作为最终结果反馈给用户. Mehdi 等^[14] 将数据分为 3 类: 数值、字符以及混合类型. 对于数值和混合类型的数据, 采用正则表达式分析数据的模式特征, 从而计算各属性元素之间的句法相似程度; 对于字符类型数据, 引入 Google 相似度距离 (Google similarity distance, GSD)^[15], 获取实例之间的语义关系. 随着自然语言处理技术的发展, 在最近的研究中, Nozaki

等^[16]将词向量引入模式匹配,利用 Word2Vec 工具^[17]计算比较实例之间的语义相似关系。

本文考虑到实际应用中孤岛数据模式信息难以完整获取的特性,采用实例级别的模式匹配方法,结合提取到的多维特征,提出基于匈牙利算法^[18]的属性间对齐发现机制,提高了模式匹配的算法性能。

2.2 实体对齐

数据融合的核心在于实体对齐 (entity alignment), 需要根据相互匹配属性的记录特征, 将数据源中指代同一实体的记录连接起来^[19]。在相关工作中, 也被称为实体解析 (entity resolution)^[20]、记录链接 (record linkage)^[21]。目前已有的方法可以划分为基于规则的方法与基于学习的方法, 可用于数据库重复数据删除 (database deduplication)^[22, 23]、知识库扩充 (knowledge base enrichment)^[24, 25] 等方面。基于规则的方法, 一般根据人为定义的规则来确定记录是否指代同一实体。最为典型的规则就是记录与记录之间的相似程度^[26, 27], 通过查找所有相似度大于阈值的记录对作为匹配实体。结合字符串相似度, Singh 等^[28]使用合取 ($\wedge$)、析取 ($\vee$)、非 ($\neg$) 等逻辑连接词将依赖规则定义成通用布尔公式 (general Boolean formulas, GBFs)。通过程序自动化地生成 GBF, 从而判断是否为同一实体。此外, Fan 等^[29]结合了数据的语义信息, 识别构建匹配依赖 (matching dependencies, MDs) 关系, 所有满足依赖规则的记录为匹配实体。基于规则的实体对齐方法, 直观简单, 但是规则制定的合理性和有效性往往强依赖于专家的先验知识, 这直接影响了算法的性能。为了解决这一问题, 有学者提出将机器学习 (machine learning, ML)^[23, 30~32]运用于这一领域。其中, Hike 框架^[32]利用人机协作的方式, 对大规模知识库进行对齐。为了减少匹配空间, 降低计算复杂度, 该框架首先通过机器学习方法对知识库进行粗略对齐, 将大规模实体对分块, 在块内建立偏序关系; 借助于推理模型将问题分发给众包平台, 并基于众包结果推断出其他问题的答案; 最后, 经过容错处理获得实体对齐的最终结果。作为 ML 的重要研究分支, 深度学习 (deep learning, DL) 也越来越多地被应用于实体对齐^[19, 33]。Mudgal 等^[33]指出, 对于长文本数据和脏数据来说, DL 方法以时间为代价, 取得了较好的对齐效果, 然而对于结构化数据来说, 基于 ML 的 Magellan 方法^[34]在效率和精度上更为优越。

本文结合数据的结构化特征, 复用并改进 Magellan 方法, 将问题转化为有监督的分类问题, 引入多样性取样策略, 提高了分类模型的泛化能力, 有效改善了对齐的效果。

3 多源关系数据融合

3.1 整体框架

结合数据融合的基本步骤, 我们构建了如图 1 所示的 MSF 多源关系数据融合框架。框架包含 4 个部分, 其中 3 个主要模块分别是模式匹配、实体对齐与实体融合。

(1) 数据选取。系统的数据可来源于网络抓取或信息管理系统, 将数据导入数据库之后, 用户可以交互性地选取需要进行模式匹配的数据表。

(2) 模式匹配。对选定数据进行多维特征的提取, 引入匈牙利算法进行属性间对齐关系的发现与计算; 结合数据特异性, 对最佳匹配结果进行排序。具体细节将在 3.2 小节进行介绍。

(3) 实体对齐。用户选定匹配属性对之后, 对两表中的数据, 计算匹配属性记录对的重叠系数 (overlap coefficient), 初步筛选出可能匹配的记录对组成候选集; 利用多样性取样策略从候选集中提取样本集, 用于用户标记; 对标记好的样本集提取特征, 以进行分类器的训练; 最终利用训练好的分类器对候选集进行进一步的筛选。具体细节将在 3.3 小节介绍。

图 1 (网络版彩图) 多源关系数据融合 MSF 框架

Figure 1 (Color online) Multi-source relational data fusion framework

(4) 实体融合. 将对齐实体进行融合, 消除不同数据源之间同一实体属性值冲突的问题, 为数据分析提供统一的数据视图. 具体细节将在 3.3 小节介绍.

3.2 模式匹配算法

模式匹配的目的在于获取不同数据源之间的属性对应关系. 直观地来讲, 在数据量小、数据模式复杂度低的情况下, 人为指定属性对应关系是一种最为简单的解决方案. 然而, 当数据表数量众多、模式复杂时, 人工判别耗时耗力且易出错. 因此, 本文采用了自动化的模式匹配方法, 可以实现多表两两之间的模式匹配关系识别. 此外, “信息孤岛”中, 各个部门之间数据的高度自治性, 导致无法准确获取数据模式信息, 并且可能存在巨大的差异 (比如在 A 部门中关于员工的家庭住址信息存储于“家庭住址”列, 而 B 部门存储于“JTZZ”列, C 部门存储于“Address”列). 在该种情况下, 仅仅根据模式信息就无法准确获知属性之间的对应关系. 基于上述分析, 本文采取了实例级别的模式匹配方法, 通过分析各列数据特征, 获取属性之间映射关系.

3.2.1 算法流程

模式匹配需要在多表输入的情况下, 识别两两数据表属性之间的映射关系. 假设当前有 n 个表输入, 则需要进行 $C(n, 2)$ 次模式匹配计算. 为了更好地分析和理解算法过程, 现对两表之间的模式匹配问题进行定义.

问题描述: 假设 A, B 两表分别具有 a, b 个属性 ($0 < a \leq b$), 称包含属性个数较少的 A 表为小表, 则需要为小表中的每个属性求解出 B 表中与之最为匹配的属性列. 在本框架中, 模式匹配分为以下 3 个步骤进行:

(1) 特征提取. 对各属性列数据进行特征提取, 以衡量各列之间的相似程度. 现有的文本特征提取方法, 广义上可以分为两大类: (1) 基于统计计算; (2) 基于语义分析^[35], 二者均可运用于模式匹配算法中的特征提取. 本文从统计学的角度, 针对不同类型的数据, 各自选取了平均长度、变异系数等具有代表性的多维统计量 (见表 1), 将实例数据统一转化为数值特征, 用于表征数据的句法特点.

表 1 特征提取
Table 1 Feature extraction

Features of string type	Features of int/float type
ItemLetterAverage	ItemAverage
LetterVariationCoefficient	ItemVariationCoefficient
LetterProbablility	ItemProbability
NormalizedLetterEntropy	—

(2) 距离矩阵生成. 得到 A 表和 B 表中每列数据的多维特征值之后, 对属性列之间两两比较, 求解出距离 d_{ij} , 构成 A, B 两表的距离矩阵 $D_{ab} = (d_{ij})_{a \times b}$, 其中

$$d_{ij} = \begin{cases} \frac{\sum_{k=1}^{\text{valuenum}} (1 - e^{f_{ik} - f_{jk}})}{\text{valuenum}} \in (0, 1), & \text{数据类型一致,} \\ 1, & \text{数据类型不一致,} \end{cases} \quad i = 1, 2, \dots, a; j = 1, 2, \dots, b. \quad (1)$$

f_{ik} 表示 A 表中第 i 列的第 k 个特征值. 当待比较的两列数据类型一致时, 需要根据公式进行 d_{ij} 的计算; 否则直接置为 1, 表示两列数据类型不一致, 无法进行匹配.

(3) 最佳匹配寻优. 基于第 2 步得到的距离矩阵 D_{ab} , 寻找出最佳匹配结果, 使得在该匹配下, 所有匹配属性之间距离之和最小. 为了形式化地描述该问题, 引入 $a \times b$ 个 0-1 变量 x_{ij} :

$$x_{ij} = \begin{cases} 1, & \text{A 表中的第 } i \text{ 列与 B 表中的第 } j \text{ 列匹配,} \\ 0, & \text{A 表中的第 } i \text{ 列与 B 表中的第 } j \text{ 列不匹配,} \end{cases} \quad i = 1, 2, \dots, a; j = 1, 2, \dots, b.$$

这样, 问题的数学模型就可以表示为

$$\begin{aligned} \min z &= \sum_{i=1}^a \sum_{j=1}^b d_{ij} x_{ij}, \\ \text{s.t.} \quad &\begin{cases} \sum_{i=1}^a x_{ij} = 1, & j = 1, 2, \dots, b, \\ \sum_{j=1}^b x_{ij} = 0 \text{ 或 } 1, & i = 1, 2, \dots, a, \\ x_{ij} = 0 \text{ 或 } 1, & i = 1, 2, \dots, a; j = 1, 2, \dots, b. \end{cases} \end{aligned} \quad (2)$$

为了求解上述模型, 得到最佳匹配, 最为直观的方法是遍历每种组合, 迭代地以回溯法寻找最佳匹配. 但是该种算法的时间复杂度为 $O(a!)$. 本文将基于距离矩阵的模式匹配寻优问题, 转化为了运筹学中经典的指派问题, 引入时间复杂度为 $O(a^3)$ 的匈牙利算法^[18], 在多项式算法内求解出最佳匹配结果. 该算法以寻找二分图中增广路径为核心, 求解所有匹配对距离之和最小的组合为最优匹配.

模式匹配算法的详细描述见算法 1.

3.2.2 结果排序策略

上述的模式匹配算法, 考虑了实例级别数据特征之间的差异, 提取数据的多维特征, 以构建距离矩阵. 采用匈牙利算法, 在保证匹配查全率的基础上, 提高了算法的效率. 类比搜索引擎中对搜索页结

Algorithm 1 Schema matching algorithm based on Hungarian**Input:** n tables to be aligned: tbs;**Output:** Schema matching results: bestmatches_list;

```

1: function multimatch(tbs)
2:   Feature_list  $\leftarrow$  feature_extraction(tbs); //Extract feature
3:   for idx1,idx2 in itertools.combinations(range(len(Feature_list)), 2)
4:     DisMatrix  $\leftarrow$  Combination(Feature_list[idx1], Feature_list[idx2]); //Generate distance matrix
5:     bestmatches  $\leftarrow$  Hungarian(DisMatrix); //Find the best matches with Hungarian
6:     bestmatches_list.append(bestmatches);
7:   end for
8:   return bestmatches_list.
9: end function

```

果排序以提升用户满意度的思想, 我们需要为两表之间模式匹配的结果设计排序算法, 使得匹配可能性大的属性对尽可能地排名靠前. 而属性对匹配与否除了与相应的属性值有关, 还与属性的特异性程度 unique 有关.

$$\text{unique} = \frac{|\text{schema_instances}|}{\text{schema_instances}} \in (0, 1], \quad (3)$$

其中 $|\text{schema_instances}|$ 指的是该属性对应的全部记录去重后的数量, 比如某个属性对应的实例为 $[1, 2, 3, 3, 2, 4, 5]$, 则 $\text{unique} = 5/7$. unique 数值表征了属性为主码的可能性, 亦为该属性对应属性值所含信息量的大小. 当 $\text{unique} = 1$ 时, 表示该属性对应的所有记录中没有重复数据, 即该属性可以唯一标识该关系模式下的一个实体, 我们希望模式匹配结果中包含此类属性的对齐关系可以优先显示. 因此综合考虑属性值的 unique 程度以及对应属性元素之间的距离, 我们定义了每对匹配属性的收益值 profit (式 (4)), 用于表示 A 表中第 i 列与 B 表中第 j 列匹配的程度. 最终两表的模式匹配结果按照该数值计算结果降序排列输出.

$$\text{profit}_i = (1 - d_{ij}) \times \max(\text{unique}_i, \text{unique}_j) \in (0, 1). \quad (4)$$

属性对的收益越大, 表示通过该对属性对两表进行实体对齐的可能性越大, 从而更好地为后续实体对齐提供模式匹配参考.

3.3 基于 Magellan 的实体对齐与融合

实体对齐旨在将不同数据源中指代同一实体的记录连接起来. 基于模式匹配结果, 不同数据源中, 需要匹配的属性对已经明确. 为了高效地实现实体对齐与实体融合, 我们将问题转化为有监督的分类问题, 在 Magellan 实体对齐方法的基础上, 引入多样性 (diversity) 取样策略, 增强了分类模型的泛化能力, 有效地提高了实体对齐与实体融合的效果.

问题描述: 假设 A, B 两表, 各有 m, n 条元组, 根据模式匹配结果, 已知两表中相匹配的属性列, 需要进行对齐操作, 以识别出两表中指代同一实体的记录. 在本框架中, 实体对齐与融合分为以下 4 个步骤进行.

(1) 分块操作. 为了减小候选集规模, 首先对数据进行分块操作, 保留匹配可能性较高的实体对进行后续操作. 在此过程中, 我们把数据源中匹配属性对应的值进行拼接, 分别构成 mixture 字符串, 通过计算字符串的重叠系数 (overlap coefficient) 来对数据集进行初步筛选, 形成候选集 C .

(2) 样本集的提取与标记. 从候选集 C 中提取一定数量的样本集, 人工判别匹配错误/正确, 并进行 0/1 标注. 为了缓解 Magellan 方法中随机取样导致的匹配效果不稳定, 本文采用多样性取样策略

进行样本集的提取. 具体地, 假设 A, B 两表中的匹配属性对为 (A_1, B_1) 与 (A_2, B_2) , 则对于候选集 C 来说, 匹配属性值的相似情况各异, 比如 $V(A_1) = V(A_2)$ 且 $V(B_1) = V(B_2)$, $V(A_1) = V(A_2)$ 且 $V(B_1) \neq V(B_2)$, $V(A_1) \neq V(A_2)$ 且 $V(B_1) = V(B_2)$, $V(A_1) \neq V(A_2)$ 且 $V(B_1) \neq V(B_2)$. 因此在进行样本集提取时, 应当考虑每一种情况, 按比例进行子样本集的抽取, 组合形成具有多样性的样本集 S .

(3) 分类器训练. 对标记好的样本集 S , 基于相似度 (如 jaccard, cosine 等) 进行实体特征抽取, 并将其划分为训练集和测试集, 用于机器学习分类器 (如 Decision Tree, Random Forest, Support Vector Machine 等) 的训练与评估. 最终将训练好的最佳分类器应用于候选集 C , 进行实体对齐正确与否的判别.

(4) 实体对齐与融合. 将分类器预测出的对齐实体进行融合, 为数据分析提供统一的数据视图. 整体实体对齐与融合的算法描述见算法 2.

Algorithm 2 Entity matching algorithm based on Magellan

Input: Two tables to be aligned: $tb1, tb2$;

Matched attribute pairs: $schema_correlations$;

Output: $fusion_results$;

```

1: function Magellan( $tb1, tb2, schema\_correlations$ );
2:    $data1, data2 \leftarrow loaddata(tb1, tb2)$ ;
3:    $C \leftarrow overlap\_coefficient\_join(data1, data2, schema\_correlations); // blocking$ 
4:    $S \leftarrow diversity\_sample(C); // sampling$ 
5:    $Label \leftarrow load\_labeled\_sample(labeled\_path); // Label candidate pairs in S with 0/1 manually, which means$ 
      $non-matched/matched respectively and load it;$ 
6:    $F \leftarrow feature\_extract(Label)$ ;
7:    $Train, Test \leftarrow split(F)$ ;
8:   for  $match$  in  $matchers$ ;  $// decision tree, random forest, support vector machine \dots$ 
9:      $model, result \leftarrow cross\_validation(match, Train); // training$ 
10:     $bestmodel, bestresult \leftarrow compare(model, result)$ ;
11:  end for
12:   $predictions \leftarrow best\_model.predict(Test); // evaluation$ 
13:   $eval\_result \leftarrow eval\_matches(predictions)$ ;
14:   $candset\_feature \leftarrow feature\_extract(C)$ ;
15:   $alignment\_results \leftarrow best\_model.predict(candset\_feature); // prediction$ 
16:   $fusion\_results \leftarrow formalize(alignment\_results); // data collation and specification$ 
17:  return  $fusion\_results$ .
18: end function

```

4 系统实现与实验验证

为了验证 MSF 数据融合框架的效果和效率, 本文实现了数据融合系统 DataPuzzle. 在该系统中, 结合真实数据, 对提出的算法进行了验证.

4.1 数据集

本文基于政务网站¹⁾ 和某企业信息查询网站中抓取到的北京市政务数据和企业数据, 对改进后的模式匹配、实体对齐与融合算法性能进行了测试与分析. 在 4.4 小节中将选取北京市水务局、北京市

1) <http://www.beijing.gov.cn>.

图 2 模式匹配时间对比 (单位: s)

Figure 2 Time comparison of schema matching

食品药品监督管理局、北京市应急管理局以及企业数据, 介绍数据融合过程及效果. 其中北京市水务局数据集包含了 2016~2018 年间北京市水务局行政许可和行政处罚信息公示; 企业法人数据集包含了北京市大中小企业的详细注册信息; 北京市食品药品监督管理局数据集包含了保健食品生产企业证件信息、保健食品企业标准备案信息、食品生产企业等; 北京市应急管理局数据, 包含了标准化达标信息等. 通过对上述规模不一、模式不一的多源数据进行融合, 可以获得维度更为丰富的数据信息.

4.2 性能分析

4.2.1 模式匹配

本文提出的基于匈牙利算法的属性间对齐发现机制, 利用属性值的多维特征进行最佳匹配属性对的寻优. 该算法在确保查全率的基础上, 提高了算法效率. 为了验证算法在时间性能上的优越性, 实验中, 基于测试数据, 分别对回溯法和匈牙利算法的执行效率进行了比对. 为了最小化实验结果的偶然性, 对于 n ($n = 2, 3, \dots, 12$) 表匹配, 我们分别进行 $C(12, n)$ 次实验, 求取时间平均值. 从图 2 中可以看到匈牙利算法执行多表匹配时, 性能明显优于回溯法. 当模式匹配输入为 12 张表时, 前者的执行耗时仅为后者的 30.1%. 此外, 随着数据表数量的增多, 匈牙利算法执行时间呈现平缓增速; 而回溯法时间增长明显. 在数据表属性规模增大、表数量增多的情况下, 后者将无法在多项式时间内计算出最佳的模式匹配结果.

在模式匹配效果方面, 二者求解得到的匹配结果是一致的. 我们将模式匹配计算结果与真实情况对比分析, 得到的实验结果如表 2 所示. 通过计算不同表数量下的查全率 (recall)、查准率 (precision), 可以发现查全率基本稳定在 77.1%, 说明原本匹配的属性对中, 绝大部分可以被准确地计算得到; 查准率基本稳定在 27.1%, 原因在于为了不丢失所有可能的匹配对, 算法在数据处理过程中, 为小表中每一个属性都设定了相对应的匹配属性, 而实际中两表间可能只有两对属性是可以匹配的.

为了弥补查准率的缺陷, 我们结合数据的特异性, 根据式 (4) 计算匹配收益值, 从而两表中最有可能、最具价值的匹配对将具备较高的收益值并且优先显示 (具体数据实例见 4.4 小节).

4.2.2 实体对齐

实体对齐部分, 我们引入多样性取样策略, 对 Magellan 实体对齐中的随机取样方法进行改进. 而后, 将具有多样性特征的样本集分为训练集和测试集, 提供给机器学习分类器训练与评估, 使得分类

表 2 模式匹配效果评估
Table 2 Evaluation of schema matching results

Evaluating indicator	2 tables	3 tables	4 tables	5 tables	6 tables	7 tables
Recall (%)	77.2	77.2	77.2	77.1	77.1	77.0
Precision (%)	27.2	27.2	27.1	27.1	27.1	27.1

Evaluating indicator	8 tables	9 tables	10 tables	11 tables	12 tables	Average
Recall (%)	76.9	76.8	76.7	76.5	76.4	77.1
Precision (%)	27.1	27.1	27.0	27.0	27.0	27.1

图 3 (网络版彩图) 实体对齐效果评估

Figure 3 (Color online) Evaluation of entity alignment results. (a) Precision; (b) recall; (c) F1-score

器可以更加全面地学习到各种对齐实体的数据特征, 从而提高分类器的泛化能力, 提升并稳定了实体对齐性能.

基于实验数据, 我们分别利用随机 (random) 和多样性 (diversity) 取样策略提取到的训练集进行分类器的训练, 并采用相同的测试集对两种分类器进行评估, 实验结果如图 3 所示.

对比发现, 二者在查准率 (precision) 上大体一致, 均为接近 100% 的状态. 随机取样下, 单一的样本特征使得分类器完美地学习了个别特征, 而忽略了全局, 进而在查全率 (recall) 方面, 远不足多样性取样策略的效果. 综合考量二者在 F1-score 方面的表现, 可以看到后者明显优于前者, 并且效果较为稳定. 实验验证了多样性取样的有效性, 泛化的分类器有利于潜在对齐实体的发现与识别.

4.3 系统实现

为了验证 MSF 框架的有效性和效率, 本文将数据融合过程可视化, 构建了数据融合系统 Data-Puzzle (图 4). 该系统实现了多表之间的两两模式匹配, 根据模式匹配结果, 用户可以参考性地对模式

图 4 (网络版彩图) 多源关系数据融合系统

Figure 4 (Color online) Multi-source relational data fusion platform

表 3 单源、多源模式匹配结果

Table 3 Schema matching results of single-source and multi-source data

Fusion type	Source table1	Source table2	Match results		Profit
Single-source	l. 水务局行政许可	r. 水务局行政处罚	l. 行政相对人名称	r. 行政相对人名称	0.7019
			l. 统一社会信用代码	r. 统一社会信用代码	0.5454
	l. 保健食品生产单位证件信息	r. 保健食品企业标准备案信息	l. 名称	r. 企业名称	0.8501
Multi-source	l. 水务局行政许可	r. 企业信息	l. 统一社会信用代码	r. creditcode	0.8067
			l. 行政相对人名称	r. name	0.7079
	l. 食品生产企业	r. 标准化达标信息	l. 企业名称	r. 企业名称	0.7169

匹配结果进行筛选, 在此基础上实现实体对齐与融合.

4.4 融合结果分析

以抓取到的北京市政务数据和企业数据为例, 对数据融合的过程和结果进行展示和分析. 首先, 对多领域数据进行自动化模式匹配. 为了多方面地对算法效果进行分析和验证, 我们选取来自北京市水务局的 (行政许可, 行政处罚), 来自北京市食品药品监督管理局的 (保健食品生产单位证件信息, 保健食品企业标准备案信息), 以及分别来自北京市水务局和网络的 (行政许可, 企业信息), 来自北京市食品药品监督管理局和北京市应急管理局的 (食品生产企业, 标准化达标信息) 四组模式匹配表对作为模式匹配算法的输入, 具体的匹配结果见表 3. 可见, 无论是对于单源关系数据, 还是多源关系数据, 模式匹配算法均可以将匹配属性准确地检测出来, 相应的匹配属性对具有较高的收益值, 可以为下一步实体对齐提供可靠参考. 上述多例模式匹配结果, 验证了本文模式匹配算法用于单源、多源数据属性对齐发现的有效性.

在此基础上, 通过人工筛选匹配的属性对, 可进行进一步的实体对齐与融合. 结合真实数据, 对融合结果进行分析, 发现指代同一实体的匹配数据中, 存在部分不完全匹配数据 (如表 4 加粗字段所示), 包括公司更名或公司简称导致的书写不一致、企业信用代码书写格式不同、录入出错等情况, 进一步验证了多样性取样策略的引入有利于潜在对齐数据的发现, 有效地解决了“信息孤岛”现象中不同数据源数据难以互连互通的问题.

表 4 多源数据融合结果 (脱敏处理)
Table 4 Fusion results of multi-source data

水务局行政许可 & 水务局行政处罚					
Single-source	L 统一社会信用代码	L 行政相对人名称	r 统一社会信用代码	r 行政相对人名称	Incomplete match reasons
	91110105M****MYD5M	北京伯佳诊所有限公司	91110105M****MYD5M	北京嘉树广渠诊所有限公司	Changed company name
	4****289-0	北京市建筑工程研究院有限责任公司	911101084****2890C	北京市建筑工程研究院有限责任公司	Inconsistent entries of credit code
	91110105****4838H	北京招商局物业管理有限公司 (招商局大厦)	911101057****4838H	北京招商局物业管理有限公司	Incorrect entries of credit code and inconsistent entries of company name
	保健食品生产单位证件信息 & 保健食品企业标准备案信息				
Multi-source	L 名称		r 企业名称		Incomplete match reasons
	北京中研万通科技有限责任公司保健食品厂		北京中研万通科技中心		Inconsistent entries of company name
	华润紫竹药业有限公司		北京紫竹药业有限公司		
	水务局行政许可 & 企业信息				
L 统一社会信用代码	L 行政相对人名称	r_creditCode	r_name	Incomplete match reasons	
	911101067****5204Q	北京八方达客运有限责任公司保修分公司	911101067****5204Q	北京公共交通控股 (集团) 有限公司第二保修分公司	Changed company name
	91110006****2434H	东方艺术大厦有限公司	911100006****2434H	东方艺术大厦有限公司	Inconsistent entries of credit code
	911101056****9797T	北京瀛洲三苏大酒楼有限公司	911101056****3797T	北京瀛洲三苏大酒楼有限公司	Incorrect entries of credit code and inconsistent entries of company name
	食品生产企业 & 标准化达标信息				
	L 企业名称		r 企业名称		Incomplete match reasons
	北京航天农业生物科技有限公司		北京首诚航天农业生物科技有限公司		Inconsistent entries of company name
	宝健中国有限公司		宝健中国日用品有限公司		

5 总结

本文针对多源关系数据的融合问题, 提出了一种通用的 MSF 框架, 框架中包含 3 个主要部分: 模式匹配、实体对齐、实体融合. 为了实现多表之间的高效模式匹配, 提出了基于匈牙利算法的属性间对齐发现机制, 并对模式匹配结果提出一种基于数据特异性的排序算法, 可以有效改善模式匹配结果的可用性. 实体对齐部分, 在 Magellan 方法中引入多样性取样策略, 提升了实体对齐的效果. 最后将对齐实体进行融合, 消除了不同数据源之间同一实体属性值冲突的问题.

基于 MSF 框架, 本文实现了多源关系数据融合系统 DataPuzzle, 结合真实数据对算法的效果和效率进行验证和分析. 实验结果表明, 本文提出的 MSF 框架具有普遍适用性, 实现的 DataPuzzle 系统可以高效地实现数据融合, 具有较高的查全率、查准率. 在之后的工作中, 我们将进一步研究如何利用自然语言处理方法, 将词汇特征表示成向量, 提取数据中的语义特征, 提高模式匹配效果, 更好地进行数据融合.

参考文献

1 Doan A H, Halevy A, Ives Z. Principles of Data Integration. Waltham: Elsevier, 2012. 19–58

2 Song X Y, Wang Y H. Data Integration and Application Integration. Beijing: China Water Power Press, 2008 [宋晓宇, 王永会. 数据集成与应用集成. 北京: 中国水利水电出版社, 2008]

3 Pandey Y, Bansal S. Safety check: a semantic web application for emergency management. In: Proceedings of the International Workshop on Semantic Big Data, New York, 2017. 1–6

4 Qin Y B, Feng L, Chen Y P, et al. “Intelligent Court” data fusion analysis and integrated application. Big Data, 2019, 3: 35–46 [秦永彬, 冯丽, 陈艳平, 等. “智慧法院”数据融合分析与集成应用. 大数据, 2019, 3: 35–46]

5 Du X Y, Chen Y G, Lu W, et al. Data wrangling: a key technique of data governance. Big Data Res, 2019, 5: 13–22 [杜小勇, 陈跃国, 卢卫, 等. 数据整理 —— 大数据治理的关键技术. 大数据, 2019, 5: 13–22]

6 Alwan A A, Nordin A, Alzeber M, et al. A survey of schema matching research using database schemas and instances. Int J Adv Computer Sci Applicat, 2017, 8: 102–111

7 Bernstein P A, Madhavan J, Rahm E. Generic schema matching, ten years later. In: Proceedings of the 37th International Conference on Very Large Data Bases, Seattle, 2011. 695–701

8 Munir S, Khan F, Riaz M A. An instance-based schema matching between opaque database schemas. In: Proceedings of the 4th International Conference on Engineering Technology and Technopreneuship, Kuala Lumpur, 2014. 177–182

9 Do H H, Erhard R. COMA: a system for flexible combination of schema matching approaches. In: Proceedings of the 28th International Conference on Very Large Data Bases, Hong Kong, 2002. 610–621

10 Zhao H, Ram S. Combining schema and instance information for integrating heterogeneous data sources. Data Knowl-

- edge Eng, 2007, 61: 281–303
- 11 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*, Washington, 2002. 117–128
- 12 Alzeber M, Alwan A A, Nordin A, et al. An empirical comparative study of instance-based schema matching. *Int J Electr Electron Syst*, 2018, 10: 1266–1277
- 13 Li Y, Liu D B, Zhang W M. A method for automatic schema matching using characteristic of data distribution. *Comput Sci*, 2005, 32: 85–87 [李由, 刘东波, 张维明. 基于数据实例分布特征的自动模式匹配方法. *计算机科学*, 2005, 32: 85–87]
- 14 Mehdi O A, Ibrahim H, Affendey L S. An approach for instance based schema matching with google similarity and regular expression. *Int Arab J Inf Technol*, 2017, 14: 755–763
- 15 Cilibrasi R L, Vitanyi P M B. The Google similarity distance. *IEEE Trans Knowl Data Eng*, 2007, 19: 370–383
- 16 Nozaki K, Hochin T, Nomiya H. Semantic schema matching for string attribute with word vectors and its evaluation. *Int J Netw Distrib Comput*, 2019, 7: 100–106
- 17 Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. In: *Proceedings of the International Conference on Learning Representations*, Scottsdale, 2013. 1–12
- 18 Kuhn H W. The Hungarian method for the assignment problem. *Naval Res Logist*, 1955, 2: 83–97
- 19 Zhao C, He Y Y. Auto-EM: end-to-end fuzzy entity-matching using pre-trained deep models and transfer learning. In: *Proceedings of the World Wide Web Conference*, San Francisco, 2019. 2413–2424
- 20 Dong X L, Rekatsinas T. Data integration and machine learning: a natural synergy. In: *Proceedings of the 2018 International Conference on Management of Data*, New York, 2018. 1645–1650
- 21 Dunn H L. Record linkage. *Am J Public Health Nations Health*, 1946, 36: 1412–1416
- 22 Christen P. Development and user experiences of an open source data cleaning, deduplication and record linkage system. *SIGKDD Explor Newsl*, 2009, 11: 39–48
- 23 Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning. In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, 2002. 269–278
- 24 Dong X, Gabrilovich E, Heitz G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 2014. 601–610
- 25 Nickel M, Murphy K, Tresp V, et al. A review of relational machine learning for knowledge graphs. *Proc IEEE*, 2016, 104: 11–33
- 26 Xiao C, Wang W, Lin X. Ed-join: an efficient algorithm for similarity joins with edit distance constraints. *Proc VLDB Endow*, 2008, 1: 933–944
- 27 Jiang Y, Li G, Feng J, et al. String similarity joins: an experimental evaluation. *Proc VLDB Endow*, 2014, 7: 625–636
- 28 Singh R, Meduri V, Elmagarmid A, et al. Generating concise entity matching rules. In: *Proceedings of the 2017 ACM International Conference on Management of Data*, New York, 2017. 1635–1638
- 29 Fan W, Jia X, Li J, et al. Reasoning about record matching rules. *Proc VLDB Endow*, 2009, 2: 407–418
- 30 Arasu A, Gotz M, Kaushik R. On active learning of record matching packages. In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, Indianapolis, 2010. 783–794
- 31 Singla P, Domingos P. Entity resolution with markov logic. In: *Proceedings of the 6th International Conference on Data Mining*, Hong Kong, 2006. 572–582
- 32 Zhuang Y, Li G, Zhong Z, et al. Hike: a hybrid human-machine method for entity alignment in large-scale knowledge bases. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Singapore, 2017. 1917–1926
- 33 Mudgal S, Li H, Rekatsinas T, et al. Deep learning for entity matching: a design space exploration. In: *Proceedings of the 2018 International Conference on Management of Data*, Houston, 2018. 19–34
- 34 Konda P, Das S, Suganthan G C P, et al. Magellan: toward building entity matching management systems. *Proc VLDB Endow*, 2016, 9: 1197–1208
- 35 Pang J A. Research on web text feature extraction method and its development. *Inf Studies Theory Appl*, 2006, 3: 338–340,367 [庞景安. Web 文本特征提取方法的研究与发展. *情报理论与实践*, 2006, 3: 338–340,367]

Multi-source relational data fusion

Yue DING^{1,2}, Juan WANG^{1,2}, Wei LU^{1,2*}, Chuitian RONG³ & Xiaoyong DU^{1,2}

1. MOE Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing 100872, China;

2. School of Information, Renmin University of China, Beijing 100872, China;

3. School of Computer Science and Technology, Tianjin Polytechnic University, Tianjin 300387, China

* Corresponding author. E-mail: lu-wei@ruc.edu.cn

Abstract Focusing on the problem of relational data fusion in the environment with “information isolated island”, this paper presents a multi-sources relational data fusion (MSF) framework. The framework consists of three components: schema matching, entity alignment, and entity fusion. Based on the Hungarian algorithm, we propose an alignment discovery mechanism for the attributes alignment among multi-sources relational data. By extracting the multi-dimensional features of attribute values, we efficiently realized schema matching of multi-sources relational data. To link the tuple pairs from multi-source data, we introduced the diversity sampling strategy and the entity feature extraction approach. These can effectively improve the performance of entity alignment. Finally, linked entities are fused to provide a unified view of data analysis. To verify the usefulness and efficiency of the proposed methods, we implemented a fusion system called Data Puzzle, which is verified with the real public multi-field data. Experimental results demonstrate that the proposed methods can fuse multi-source relational data efficiently with high recall and precision.

Keywords multi-source heterogeneous data, relational data, information isolated island, schema matching, entity alignment, data fusion

Yue DING was born in 1996. She is a master's candidate in the School of Information at Renmin University of China. Her research interests include data integration and knowledge graph.

Juan WANG was born in 1996. She is a master's candidate in the School of Information at Renmin University of China. Her research interest is in the field of natural language processing.

Wei LU was born in 1981. He is an associate professor at Renmin University of China. He received his Ph.D. degree in computer science from Renmin University of China in 2011. His research interests include query processing in the context of spatiotemporal, cloud database systems, and applications.

Xiaoyong DU was born in 1963. He is a professor at Renmin University of China. He received his Ph.D. degree from Nagoya Institute of Technology in 1997. His research focuses on intelligent information retrieval, high-performance database, and unstructured data management.