综述评介

关于实体解析基本方法的研究和述评*

高广尚
(桂林理工大学商学院　桂林 541004)

摘要: 【**目的**】探讨实体解析理论中经典的实体解析方法及逻辑思路。【**文献范围**】在 Google Scholar 和 CNKI 中分别以检索词“Entity Resolution”、“Collective Analysis”、“Crowdsourced”、“Active Learning”、“Privacy-Preserving”和“实体解析”进行文献检索, 再结合主题筛选, 精读并使用追溯法获得实体解析研究的代表性文献共 86 篇。【**方法**】针对每种实体解析方法, 归纳分析该方法的基本思想, 并通过图示直观地呈现其中的解析过程; 重点分析梳理方法实现过程中, 现有研究所采用的关键策略、算法或技术等。【**结果**】实体解析是数据质量管理的基本操作, 也是发现数据价值的关键步骤。【**局限**】未深入分析各实体解析方法的评价指标和应用情况。【**结论**】尽管现有实体解析方法能在一定程度上满足大部分应用的需求, 但在大数据环境下其仍然面临着数据混杂性、隐私保护和分布式环境等方面的挑战。

关键词: 实体解析　协同分析　众包　主动学习　隐私保护
分类号: TP393
DOI: 10.11925/infotech.2096-3467.2018.1388

1　引　言

大数据环境下, 促进各种不同环境中所收集数据的集成共享, 并减少数据分析成本, 成为众多大数据应用的一项核心要求。然而, 有关数据质量、计算复杂度和数据演化等问题让大数据应用面临诸多挑战。在这一背景下, 作为大数据应用中关键技术之一的实体解析(Entity Resolution, ER), 也相应地引起国内外人工智能、深度学习和信息融合等领域学者高度关注[1-4]。实体解析指识别出数据集(或数据库)中那些描述同一现实世界实体的数据对象, 以此达到数据清洗和集成的目的[5]。这里的数据对象是结构化的, 又称为数据记录或近似重复记录(Approximate Duplicate Records), 通常包括多个属性, 例如姓名、年龄和地址等。

自 Newcombe 等[6]于 1959 年首次提出应用计算机自动进行数据识别, 并提出实体解析概念以来, 实体解析研究已取得一系列重要发展和创新, 主要包括以下方面: 相似度计算、通用扩展、协同分析、众包模式、主动学习、实时应用和隐私保护等。然而, 随着信息技术和互联网技术的迅猛发展, 各种数据呈现出爆炸式增长态势, 其所具有的大规模性、类型多样性和关联复杂性等给现有的实体解析研究带来巨大挑战, 例如需根据实际情况重新设计算法等, 进而影响大数据应用前景。鉴于此, 本文对现有实体解析的基本方法及逻辑思路进行梳理, 以期为大数据环境下的实体解析研究提供进一步的理论指导和实践参考。

2　基于概率决策的实体解析方法

基于概率决策的实体解析方法(Probabilistic Entity Resolution)的基本思想是: 通过计算对应属性值之间

通讯作者: 高广尚, ORCID: 0000-0003-4140-1735, E-mail: 25969393@qq.com。

*本文系国家自然科学基金项目“面向数据演化的增量实体解析方法研究”(项目编号: 71761008)和广西高校人文社会科学重点研究基地基金项目“面向企业数据治理的数据质量改善研究”(项目编号: 16YB010)的研究成果之一。

的相似度，并结合属性级阈值和记录级阈值，决定两条记录是否匹配[7]。解析过程如图 1 所示，其中的“比较”模块涉及记录之间两两成对比较，这里暂不讨论“索引分块”和“匹配决策”两个模块。

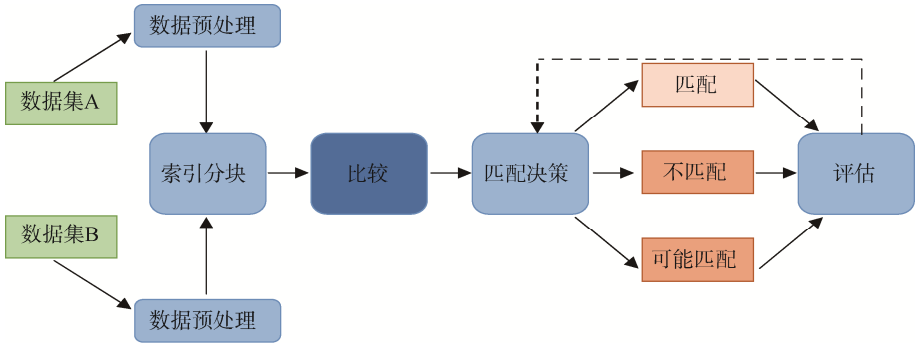

图 1　基于概率决策的实体解析过程[8]

该方法的典型代表是 Fellegi 等[8]提出的 Fellegi-Sunter 解析模型，该模型将实体解析问题视为传统分类问题，即在属性间条件独立的假设下，汇总所有对应属性值之间的相似性得分(相似度)，并计算当前记录对的条件概率比值(R)，然后将 R 与上下两个阈值(记录级)进行比较以判定当前记录对的匹配状态(匹配、不匹配或可能匹配)。很显然，字符串相似度计算是实体解析中的核心基础模块。为从字符串相似度计算角度实现基于概率决策的实体解析，现有研究主要从三个方面开展工作：基于词项的相似度、基于编辑距离的相似度和基于混合的相似度。

2.1　基于词项的相似度

这类相似度算法将属性值看作词项(token)的集合，主要包括 Jaccard 相似度、TF-IDF 相似度和 q-grams 相似度。Jaccard 相似度是一种常见的判定相似程度的指标，作用在两个集合上，其值为集合交与集合并的比值[9]。Jaccard 相似度多用于检测无拼写错误字符串的相似度，且对字符串内各词项顺序的变化不敏感。TF-IDF 相似度将属性值作为文档(包含词项)处理，并用向量表示，最后通过余弦算法计算出两个向量之间的相似度，以表示属性值之间的相似度[10]。TF-IDF 相似度背后隐含的直觉是如果两个字符串有相同的可区分项，则它们是相似的。q-grams 相似度首先将各字符串切割成长度为 q 的 grams，然后再基于 Jaccard 相似度算法或余弦算法对它们进行相似度计算[11]。q-grams 相似度计算更适用于比较存在拼写错误的字符串。

2.2　基于编辑距离的相似度

这类相似度算法认为编辑距离反映人们可能会犯的各种编辑错误，比如插入一个额外的字符，或者两个字符互换，主要包括 Levenshtein 距离算法和 Jaro 距离算法。Levenshtein 距离算法由俄国科学家 Levenshtein 提出[12]。两个字符串 str_1 和 str_2 的编辑距离是将字符串 str_1 转换成字符串 str_2 所使用的最少编辑操作次数，因此编辑距离越小，两个字符串越相似。Jaro 距离算法是一种主要用于比较姓名的字符串比较算法[13]，然而，Winkler 等[14]认为前缀匹配比姓名匹配更加重要，于是他们通过前缀匹配对 Jaro 距离算法进行修正。

2.3　基于混合的相似度

基于混合的相似度结合前述两类相似度算法的优点，主要包括泛 Jaccard 相似度(Generalized Jaccard)、泛 TF-IDF 相似度(Soft TF/IDF Similarity)和 Monge-Elkan 相似度。泛 Jaccard 相似度是 Jaccard 相似度的一种自然泛化，即弱化词项间的匹配要求，只要求彼此相似即可，而不必完全一致[15]。泛 Jaccard 相似度考虑词项可能相似这一情形，是因为现实中单词有可能会拼错。类似于泛 Jaccard 相似度方法，泛 TF-IDF 相似度是在 TF-IDF 相似度的基础上，结合应用辅助相似度函数(例如 Jaro-Winkler)计算词项的相似度得分，并用它作为最后字符串相似度计算的权重[16-17]。Monge-Elkan 相似度通过对子字符串设置辅助相似度方法 s′对字符串相似度有更多控制[18]。例如，s′可以判定如果 A_i 是 B_i 的前缀(例如 Comput 和 Computer)，则两者匹配，这样它们的相似度得分是 1。

3　面向通用扩展的实体解析方法

面向通用扩展的实体解析方法(Generic Entity Resolution)的思想是：仅从可扩展层面设计算法，以大大减少调用通用“黑盒”函数的次数，并且不考虑用

作匹配、合并记录的“黑盒”函数的具体实现[19-20]。这一解析过程如图 2 所示，其中，“匹配”函数用于判定两条记录是否匹配(只有若干对应属性值之间的相似度满足条件时才返回记录对匹配，否则返回不匹配)，“合并”函数用于组合并统计匹配记录对的对应属性值(可能不同)以产生新的合并后的记录(如 r_{12})。

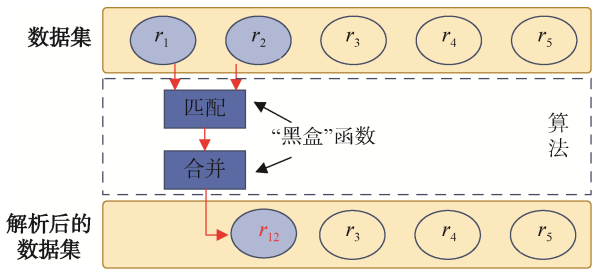

图 2　面向通用扩展的实体解析过程

这种方法的典型代表是斯坦福信息实验室开发的 SERF 解析模型[21]，该模型定义了用于“匹配”、“合并”函数的以能产生有效解析策略的 4 种性质：幂等性(Idempotence)、交换性(Commutativity)、结合性(Associativity)和代表性(Representativity)。为进一步说明如何在不同情况下利用上述性质实现较优的解析，针对三种不同情况分别设计三个算法：G-Swoosh 算法(记录层面，避免不必要的比较)、R-Swoosh 算法(记录层面，进一步避免 G-Swoosh 算法中不必要的比较)和 F-Swoosh 算法(属性层面，避免重复的属性比较)。对于每个算法，实验表明不仅能计算出正确的解析结果，而且就执行的比较次数而言也是“最优的”。为从算法设计角度开展面向通用扩展的实体解析研究，现有研究主要从三个方面开展工作：基于并行算法、基于缓冲算法和基于分布式算法。

3.1　基于并行算法

Kawai 等[22]认为实体解析过程通常是计算密集型的，因此在多个处理器之间分配解析负载将是非常重要的。为此，提出并行算法 P-Swoosh，它使用通用的匹配和合并函数，同时考虑在处理器之间实现负载平衡。实验结果显示，P-Swoosh 算法具有三方面优点：即使在并行环境中，算法也可以避免许多冗余比较；算法可以降低 R-Swoosh 算法中合并记录时的计算成本；无论是否具有领域知识，算法在 2-15 个处理器上具有较好的线性可扩展性。类似地，Kim 等[23]认为合并的记录可能会与数据集中其他记录匹配，因此要求解析方案必须是可迭代的，这会让解析过程复杂化。鉴于此，提出基于数据“特征(Characteristics)”的并行算法，例如 a 包含 b、a 等同于 b，以及 a 和 b 重叠等特征，从而使得多处理器之间的冗余计算和开销最小化。

3.2　基于缓冲算法

Kawai 等[24]认为尽管解析过程的成本很高，但对于有限的内存或非常大的数据集来说，解析过程中磁盘 I/O 读写的成本同样可能是关键的，因此在给定“匹配”、“合并”函数的情况下，在内存有限的单处理器中缓冲(buffer)一个大的数据集是必要的。鉴于此，提出基于延迟磁盘更新(Lazy Disk Update)和位置感知匹配调度(Locality-Aware Match Scheduling)的缓冲算法 Bufoosh，利用初始记录排序定位内存中的匹配候选对象。实验结果显示，算法可以实现非常高的命中率并能显著降低磁盘 I/O 读写次数。

3.3　基于分布式算法

为能在多个处理器上并行运行解析过程，Benjelloun 等[25]对 R-Swoosh 算法进行扩展，并提出分布式算法 D-Swoosh，以在多个处理器之间分配解析工作负载。D-Swoosh 算法使用多种范围函数(Scope Functions)和负责函数(Responsible Functions)，前者用于将输入记录分发给处理器，后者用于避免冗余的比较。这些函数中的一些用于针对没有语义知识的场景，另一些则利用 ER 应用程序中经常出现的知识。

4　基于协同分析的实体解析方法

基于协同分析的解析方法(Collective Entity Resolution)的思想是：以迭代的方式对共同出现的多种类型或同种类型的数据对象协同而非独立地进行解析[26]。与传统上使用记录属性的成对相似度进行解析的方法不同，这种方法通过考虑在早期迭代中已经发现的实体关系作为证据递增地解析实体。以通过共同作者这一关系来解析出其他作者为例，这一解析过程如图 3 所示。其中，矩形表示一篇学术论文(p_i 表示论文名称)，椭圆表示仅包括单一属性(作者)的数据对象，其中“W. Wang”等人名表示作者。

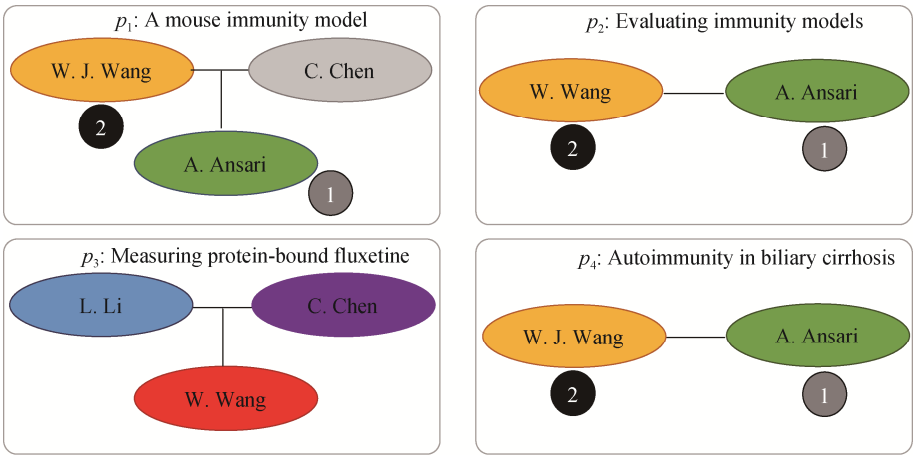

图 3　基于协同分析的实体解析过程[26]

这种方法的典型代表是 Bhattacharya 等[26]提出的协同解析学术论文中共同作者是否为同一实体的方法。为进一步理解这种协同解析过程，本文对这一过程的理论分析如下：

(1) 解析过程从最有信心的数据对象开始，例如两个名为“A. Ansari”的数据对象更可能表示同一实体(如标记 1 指示的绿色椭圆)，因为“A. Ansari”是一个不常见的姓名，这为解析其他数据对象提供额外的证据。

(2) 在名为“A. Ansari”的数据对象被解析后，那么分别来自论文 p_1, p_2 和 p_4 的名为“Wang”的数据对象就有一个共同的合作者，即“A. Ansari”，为解析它们表示同一实体提供了证据(如标记 2 指示的橙色椭圆)。相比之下，来自论文 p_3 名为“W. Wang”的数据对象因有不同的合作者，从而说明它表示另外的实体。以此类推，可以解析出其他合作者。为从协同分析角度开展实体解析研究，现有研究主要从基于聚类算法和基于关联算法两方面开展工作。

4.1　基于聚类算法

基于聚类算法的思想是利用共现数据对象的聚簇相似度进行迭代聚类，从而解析出多个相互关联的数据对象。Malin 等[27]利用层次聚类算法(Hierarchical Clustering)对不明确的“人名”数据对象进行聚类，聚簇之间的相似度计算使用余弦相似度函数，其中，最相似的聚簇将会合并成一个新的聚簇(表示它们共同描述同一实体)，此过程将会继续，直到满足预先指定的停止条件，或所有的“人名”数据对象被合并在一个共同的聚簇中。类似地，高学东等[28]针对包含两种类型对象的实体解析问题，提出基于联合聚类思想的两阶段协同实体解析通用框架，以同时对其中的决定对象和辅助对象分别进行聚类。

4.2　基于关联算法

基于关联算法的思想是针对复杂信息空间中多类型的数据对象相互关联，并且每个数据对象仅具有少量属性的特点，利用数据对象间丰富的关联关系帮助进行实体解析，并将一些数据对象的实体解析结果传递到其关联的数据对象，同时通过解析过程中的信息增益解决部分数据对象属性信息不足的问题[29]。Dong 等[29]提出一个面向复杂信息空间的协同实体解析方法(Joint Entity Resolution in Complex Information Space, JER-CIS)，其基本思想是利用数据对象间丰富的关联关系来帮助进行实体解析，并迭代地处理整个过程。具体来说，首先利用数据对象的多种上下文信息进行数据对象匹配；之后，将一些数据对象的实体解析结果传递到与其关联的数据对象，实现匹配传播。当两个数据对象匹配后，它们的属性值将分别组成属性值集合，以产生数据对象信息增益，从而解决部分数据对象属性信息不足的问题，进而提升后续数据对象的匹配准确性。总之，JER-CIS 方法通过挖掘上下文信息、匹配传播和数据对象信息增益得到更精确的实体解析结果。类似地，孙琛琛等[30]提出一种基于图形的迭代协同解析方法(Graph-Based Iterative Joint Entity Resolution Approach, GBi-JER)，该方法与领域无关，且适合于任何关联的数据。GBi-JER 方法充分发掘逐渐收敛的对象图，迭代地对多类型关联数据对象进行协同解析，并利用不同数据对象间的关系促进彼此的匹配。

此外, Kalashnikov[31]提出一种领域无关的 R_{EL} DC 方法, 将传统的基于特征的相似度技术与分析关系的技术相结合, 以协同地进行实体解析。R_{EL} DC 方法将数据集视为对应的数据对象关系图, 然后利用图论技术分析图中节点之间存在的路径, 即分析数据对象之间的关系链(Chains of Relationships)。类似地, Naumann 等[10]在数据对象关系图的基础上, 分别从基于连接分量的划分和基于中心节点的划分两个角度, 设计相应的关联算法找出图中关联的节点。

5　基于众包模式的实体解析方法

基于众包模式的实体解析方法(Crowdsourced Entity Resolution)的思想是: 通过整合计算机和互联网上未知的大众(工作者), 完成计算机难以单独完成的人类智能任务(Human Intelligence Tasks, HIT), 从而有效提高实体解析准确率[32-33]。该方法包括三个典型步骤:

(1) 由机器对所有数据进行预处理, 并将待验证的最有影响力的候选匹配对提交到众包平台上;

(2) 由大众通过众包平台验证最可能的匹配对;

(3) 验证过程分析平台返回的结果, 确定最终匹配结果。

众包解析过程如图 4 所示, 生成策略是关键步骤, 关系到如何以最小代价获得最佳收益, 即通过算法选出最有影响力的候选匹配对放在众包平台上, 当其数量达到预算的时候则停止执行算法。

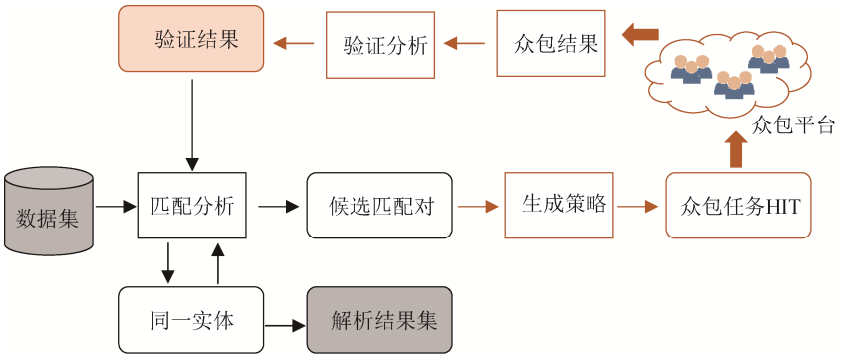

图 4　基于众包模式的实体解析过程

为从众包角度开展实体解析研究, 现有研究主要从三个方面开展工作: 结合聚类或迭代步骤、结合 HIT 生成步骤和结合返回结果验证步骤。

5.1　结合聚类或迭代步骤

除了通过图 4 中的成对匹配(Pairwise Matching)以及分块(Blocking)方式进行实体解析外, 还可以通过聚类(Clustering)方式进行[5], 因此可将众包融入实体解析的聚类过程或迭代过程中。具体来说, 将众包融入聚类过程的相似度计算中, 依据相似度决定数据对象是否属于同一个聚类, 能够有效改善数据对象间的相似度计算精度, 提高聚类准确性。

5.2　结合 HIT 生成步骤

为在众包过程中生成批处理的 HIT, 并平衡花费、质量和时间等问题, Wang 等[34]提出两种典型的众包任务生成策略: 基于成对的批处理方法和基于聚簇的批处理方法。前者将候选匹配对简单地按照 HIT 的最大量划分, 即每个 HIT 都由候选匹配对组成。例如, 在一个 HIT 上需要大众判定记录 1 和 2、5 和 7 是否描述同一实体。后者在给出一聚簇记录时, 要求大众判断哪些描述同一实体。例如, 在 HIT 上需要大众判定记录 1、2、5 和 7 中的哪些描述同一实体。事实上, 划分记录聚簇的工作已被证明是图上的简化问题, 并且是 NP 难的, 因此可以采用优化方法解决[10]。最后, 实验证明划分聚簇的方法比成对的方法更适合众包。

5.3　结合返回结果验证步骤

在确定返回结果中的匹配对方面, 现有研究主要采用 4 种方法: 基于投票的方法、黄金标准数据法、期望最大化的评估方法和结合传递性的方法。

(1) 基于投票的方法将一个任务分配给多个工作者独立回答, 然后将答案通过投票方式进行整合, 最后将大多数的意见作为最终的正确结果[35]。这种方法假定每个工作者的准确率一致。

(2) 黄金标准数据法通过设计一些具有标准答案的问题作为测试题目, 在任务开始前或者在任务进行

过程中由工作者回答，最后根据答题结果识别欺诈者，同时对工作者的准确率进行评估，进而依据贝叶斯模型或概率模型获得任务的最终结果[36-37]。这种方法假定每个工作者的答题准确率是固定的。

(3) 期望最大化的评估方法通过对任务结果和工作者的准确率不断进行迭代估计，直至收敛得到任务结果[36]。这种方法能够实现对任务结果的精确评估，但当任务或工作者较多的时候，算法运行效率较低。

(4) 结合传递性的方法通过数据对象匹配的传递关系减少所需 HIT 的数量。例如，如果 a 和 b 描述同一实体，b 和 c 不是同一实体，显然 a 和 c 不是同一实体。这种方法通过考虑传递关系提高实体解析效率。

6 基于主动学习的实体解析方法

基于主动学习的实体解析方法(Entity Resolution Using Active Learning)的思想是：通过主动学习器(Active Learner)从未标注数据对象集中有策略性地选出某些数据对象以让专家进行标注，标注后的数据对象有助于以最快的速度加强分类器[38-39]。主动学习本质上是一种机器学习方法。这一解析过程如图 5 所示，主动学习器试图迭代地从未标注数据对象集中选择信息最丰富的数据对象，而不是错误或无关的数据对象，以让专家对该数据对象进行标注，这对学习有效的分类器最有用。

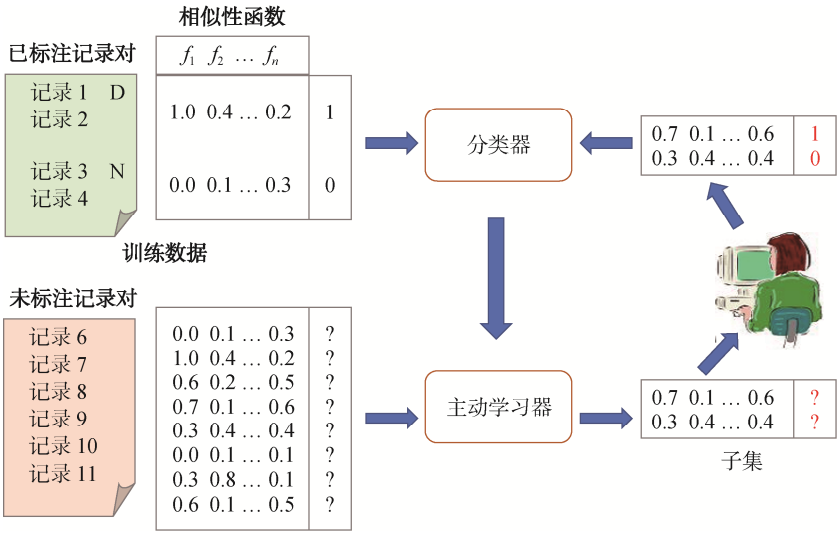

图 5 基于主动学习的实体解析过程

这种方法的典型代表是 Sarawagi 等[40]提出的基于主动学习的交互式实体解析系统 ALIAS，在专家标注的基础上分别实现基于决策树(Decision Trees)、基于朴素贝叶斯(Naive Bayes)以及基于支持向量机(Support Vector Machines, SVM)的分类方法来训练分类器，并对这些方法的性能进行比较。由于 ALIAS 系统通过主动学习自动构建匹配函数，因而能较好地解决匹配函数难以确定、词汇共现度难以利用等问题。类似地，Tejada 等[41]提出 Active Atlas 解析系统，它使用由三个决策树分类器组成的委员会来学习区分匹配和非匹配的最佳规则。从主动学习角度开展实体解析研究，主要包括从基于采样策略和基于其他策略两方面的工作。

6.1 基于采样策略

Bellare 等[42]提出基于采样策略的主动学习方法，主动选择未标注样本中信息量较大的对象交给专家进行标注，然后将标注后的样本添加到训练集中，从而让分类器在标注代价较小的情况下获得较高的分类性能。Arasu 等[43]提出利用特定问题特征的主动学习方法，其主要思想是利用额外的灵活性选择对学习任务最有用的示例，从而消除用户挑选合适示例或良好过滤器的负担。由于这种方法采用决策树和 SVM 分类模型，并让用户可以指定最终分类模型以达到所需精度，因此它能克服原有主动学习方法中存在的局限性：算法不提供原则性强的接口以供用户使用其以控制分类器的质量；算法不能扩展到较大的输入。

6.2 基于其他策略

Qian 等[44]提出一种主动学习系统，该系统可以大规模地学习多个规则，每个规则都具有对重复空间的显著覆盖，从而在高精确度的基础上实现高召回率。

具体来说，其中的算法可以通过最大化召回率来学习规则，同时满足给定的一组标注示例的高精度约束。Fisher 等[45]应用主动学习技术产生训练数据，以用于基于马尔可夫逻辑网络(Markov Logic Networks, MLNs)的实体解析模型，同时学习马尔可夫逻辑网络公式中必要的权重系数。此外，作者提出一种允许领域专家向马尔可夫逻辑网络添加新规则的方法，以捕获现有模型未正确分类的记录对。此外，Fu 等[46]从实例选择角度研究已有的主动学习研究成果并将其分为两类：一类是仅基于独立同分布(Independent and Identically Distributed, IID)实例的不确定性；另一类是基于实例相关性。在上述分类基础上，作者总结该领域的主要方法以及其技术优势和劣势，进行简单的运行时性能比较，并讨论新兴的主动学习应用程序和其中的实例选择挑战。

7　面向实时应用的实体解析方法

面向实时应用的实体解析方法(Real-Time Entity Resolution)的思想是：通过索引技术从数据集中选取一个记录块(记录子集)，用该记录块解析到来的查询记录流(Stream of Query Records)，这一过程持续时间(响应时间)为亚秒级[47-48]。这一解析过程如图 6 所示，其中，比较次数由|D|次减少为|B1|、|B2|或|B3|次。这种解析方法主要考虑两个因素：如何根据到来的查询记录(Query Record)快速确定记录块；如何快速确定块中的记录是否与查询记录匹配。显然，在大数据环境下的解析过程中，前者比后者更难处理。值得一提的是，这种解析方法在功能上与文本和网络搜索引擎有相同之处：实时、近似匹配和结果排名[49-51]。

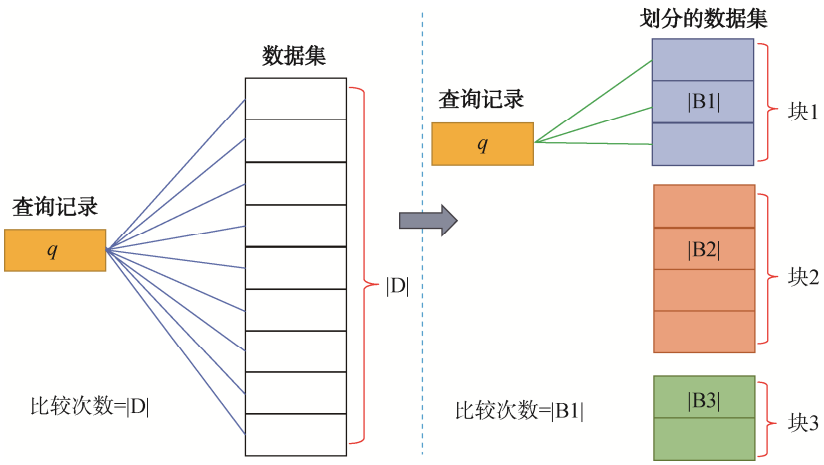

图 6　面向实时应用的实体解析过程[47]

面向实时应用的实体解析研究工作主要包括三个方面：基于索引过滤、基于 Meta-Blocking 过滤和基于其他策略。

7.1　基于索引过滤

索引(Indexing)技术的思想是通过定义某种形式的块键(Blocking Key)，以让尽可能多的真实匹配对包括在候选记录对集合中[52]。Christen 等[53]提出一种适用于实时解析的相似感知倒排索引技术(Similarity-Aware Inverted Index)，主要思想是预先计算同一块中属性值之间的相似度，并存储在主存储器中，以便在后续匹配查询记录的过程中使用。由于在匹配过程中避免了相似度计算，因而该技术能显著减少匹配时所需的时间。但该技术的缺点是只能处理静态数据集，因为索引一旦被创建，新的记录和属性值就不能被加入其中。鉴于此，Ramadan 等[54]提出一种更具灵活性的动态相似感知倒排索引技术(Dynamic Similarity-Aware Inverted Indexing, DySimII)，每当处理新的查询记录时相应的属性值就能加入到索引中。具体来说，采用三种索引结构克服最初相似感知倒排索引技术的局限性：块索引(Block Index, BI)，用来存储唯一的属性值及其关联的块键；相似性索引(Similarity Index, SI)，用来存储相同块中属性值之间预计算的相似度；记录索引(Record Index, RI)，用来存储所有唯一属性值及它们关联的记录标识符。类似地，Ramadan[47]在以前研究的基础上提出三种动态索引技术：基于分块的 DySimII 索引技术，每当新查询

记录到达时索引便会更新；基于树(Tree-Based)的DySNI 动态索引技术；基于多树(Multi-Tree Based)的F-DySNI 索引技术，它在索引数据结构中使用多个不同的树，其中每棵树都有唯一的排序键，该索引技术的目标是减少属性值开始处错误和变种的影响。此外，Ramadan 等[55]提出一种可自动选择最优块键的无监督学习方法，构建可用于实时解析的索引。

7.2 基于 Meta-Blocking 过滤

索引过滤技术在进行分块时会产生重叠的块，从而导致比较冗余，而 Meta-Blocking 技术则可清除重叠的块以避免不必要的比较，从而进一步提高解析效率。Papadakis 等[56]提出利用 Meta-Blocking 技术直接优化重叠的分块，将信息封装在块到实体关系(Block-to-Entity Relationships)中并构建块图(Blocking Graph)；将问题转化为度量图中边的权重和图修剪问题。这种做法独立于底层的索引技术，与模式无关，并具有通用性。Meta-Blocking 技术并不取代现有索引技术，而是对其进行补充。为进一步提高 Meta-Blocking 技术的效率，Efthymiou 等[57]从并行计算角度提出基于 MapReduce 的 Meta-Blocking 变种技术，其中包含两种替代策略：具有更高可扩展性的基本策略(显式地构建块图)和高级策略(隐式地构建块图，减少数据交换的开销)。

7.3 基于其他策略

Dragut 等[58]针对在线环境中的实体解析提出一种基于迭代缓存(Iterative Caching)的方法，该方法的思想是解析并缓存一组频繁请求的记录以用作将来参考，其中记录是通过从不同 Web 数据库中采样(Sampling)获得。在这种方法中，响应查询的新到达记录将与缓存中的记录一起被解析，解析后的结果呈现给用户并适当地追加到缓存中。这种方法允许对当前查询进行“快速”响应，并为后续查询提供“改进”的数据质量。类似地，Gruenheid 等[59]提出一种可增量、有效更新解析结果的端到端框架，可以在数据更新到达时逐步有效地更新解析结果，其中的算法不仅允许将数据更新中的记录与现有聚簇合并，还允许利用其中的新证据修正以前的错误解析。Whang 等[60]提出一种即付即用解析方法(Pay-as-you-go Entity Resolution)，这种方法建议使用提示(Hints)提供关于记录的信息，这些记录可能描述同一现实世界实体。提示可以用各种格式表示(例如，根据匹配的可能性对记录进行分组)，实体解析可以使用这些信息作为首先比较哪些记录的指南。该方法的目的是通过匹配的可能性排序候选记录对，从而在仅占用总运行时间的一小部分时间内找到大多数匹配记录，进而改进解析过程。

8 隐私保护下的实体解析方法

隐私保护下的实体解析方法(Privacy-Preserving Entity Resolution)的思想是：确定两个或多个组织之间的数据集中是否包含描述同一现实世界实体的记录，同时又不会向对方或其他任何组织泄露除匹配记录以外的任何信息[61-62]。解析过程如图 7 所示，其中灰色框部分表示步骤中的任何敏感数据都要以某种方式进行编码，以免任何一方了解任何其他方的数据，这是与图 1 中明显不同的地方。数据预处理步骤可以由数据集所有者单独进行，因此这一步骤在隐私保护范围之外。

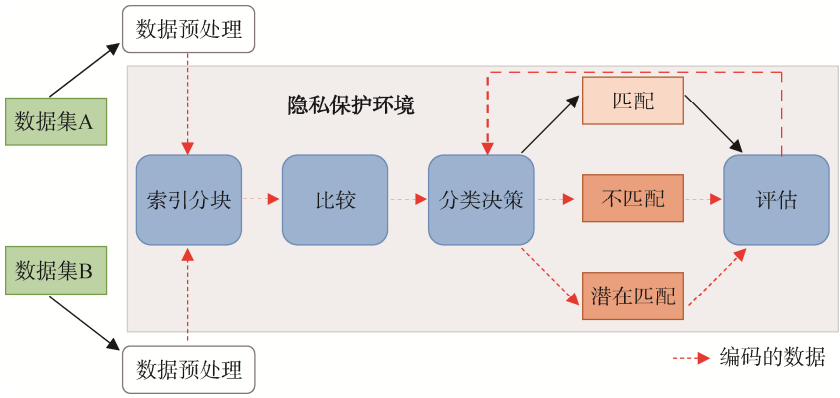

图 7 隐私保护下的实体解析过程[61]

实体解析中的隐私保护主要涉及两个方面: 如何保证数据在应用过程中不泄露隐私; 如何更有利于数据的应用。鉴于此, 现有研究主要从三个方面开展: 基于数据扰乱、基于数据重构和基于数据加密[63]。

8.1　基于数据扰乱

数据扰乱技术主要采用 *k*-匿名(*k*-Anonymity)和差分隐私(Differential Privacy)两种关键方法, 都是通过对原始数据添加噪声实现隐私保护。

(1) *k*-匿名

k-匿名最早由 Sweeney[64-65]提出, 其基本思想是, 数据在准标识符(Quasi Identifier, QI)上至少存在 *k* 个不可区分的记录, 这使攻击者不能判别出隐私信息所属的具体个体, 从而保护个人隐私。*k*-匿名通过参数 *k* 指定用户可承受的最大信息泄露风险, 即每条记录被泄露的风险为 1/*k*。类似地, Kantarcioglu 等[66]提出以 *k*-匿名方式来显示特定病人的人口统计数据。*k*-匿名方法的缺陷是被动式地防止隐私泄露, 并依据单一数据集上的攻击假设制定相应的匿名策略。然而, 大数据的大规模性、多样性使其顾此失彼。

(2) 差分隐私

差分隐私是另一种强大的数据隐私保护模型, 不依赖于任何假设, 例如攻击者的背景知识等, 其基本思想是通过对原始数据进行变换或对统计结果添加噪声来实施数据隐私保护[67]。差分隐私可以保证在数据集中添加或删除一条记录时, 不影响实体解析结果, 因此即便在最坏情况下, 攻击者已知除一条记录之外的所有敏感数据, 仍可以保证这条记录的敏感信息不会被泄露[68-69]。事实上, 差分隐私与大数据之间有天然的匹配性, 因为大数据的大规模性和多样性导致在数据集中添加或删除记录时对整体数据的影响非常小, 这一特性与差分隐私的内涵相吻合[70]。差分隐私的优点是加入的噪声与数据集大小无关, 因此对于大型数据集来说, 仅通过添加极少量的噪声就能达到高级别的隐私保护。其缺点是无法主动控制隐私参数, 从而导致隐私性偏低或偏高。此外, 大数据之间的关联性也有可能弱化差分隐私保护效果。

8.2　基于数据重构

数据重构技术指将记录信息转换为其他数值形式, 同时保留某些统计学特征而不保留真实数值。现有研究通常采用布隆过滤器(Bloom Filter)将属性值集合转换为位数组以实现数据重构[71]。布隆过滤器是 Bloom[72]提出的用于有效检查集合成员的数据结构, 也可以用来确定两个集合是否近似匹配[73]。尽管通过布隆过滤器得到的位数组, 在一定程度上代表转换前的记录并保护了记录隐私, 但转换后的位数组并不是绝对安全的, 无法抵御基于频率的密码学分析。鉴于此, Durham 等[74]提出一种利用布隆过滤器的可以抵御基于频率的密码学分析的实体解析方法。

类似地, Vatsalan 等[75]提出一种可扩展的基于“与”运算, 并结合安全合计(Secure Sum)与 Dice 相似度函数(Dice Coefficient Similarity)的多方隐私保护下的实体解析方法, 但该方法在处理存在质量问题的数据(Data with Quality Issues, DQI)时, 会丢失较多真实匹配的记录, 从而导致查全率过低, 进而使其实际应用价值偏低。韩姝敏等[62]认为大部分多方隐私保护下的实体解析方法采用精确匹配方式, 因而不具有容错性, 而那些少部分具有容错性的方法, 却在处理存在质量问题的数据时由于容错性差和时间代价大而不能有效找出数据源间的共同实体。鉴于此, 作者提出一种结合布隆过滤器、安全合计、动态阈值、检查机制和改进 Dice 相似度函数的改进方法, 以解决存在的可扩展性差、容错性差等问题。此外, Randall 等[76]提出一种使用布隆过滤器来加密个人信息的解析方法。

8.3　基于数据加密

数据加密技术常用的一种方法是安全多方计算(Secure Multiparty Computation, SMC)[77-78], 与零知识(Zero Knowledge)概念密切相关[79-80], 例如, 可以在不透露两位百万富翁净资产的情况下计算哪一个更富有。安全多方计算的基本思想是, 在分布式环境下基于多方参与者提供的数据计算出相应函数值, 并确保除参与者的输入以及输出信息外, 不会额外地暴露参与者的任何其他信息。简单来说, 它是指一组互不信任的参与者在不泄露各自隐私信息的前提下进行多方合作计算。这一计算过程如图 8 所示: 假如有三个参与方 A、B、C, 要对其中的数据进行安全合计, 首先将扰乱数据 R 传入 A, 并与 A 中的数据进行加和运算, 然后将结果传入 B, 继续进行加和运算后传入 C, 参与方 C 无法得知 A、B 中的数据, 继续进行加和运算后传回参与方 A, 减去扰乱数据 R, 即得到三者之和, 该过程中任意一方均不知道其他参与方的数据。该方法

的缺点是大数据环境下计算过程易陷入循环怪圈。

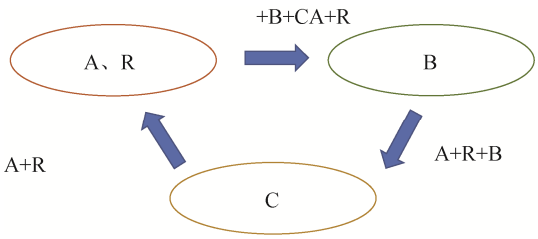

图 8　安全多方计算过程[78]

此外, Lindell 等[81]研究了安全多方计算的基本范式和概念, 并讨论了它们与保密数据挖掘领域的关系。

9　结　语

对大数据进行充分开发和利用以发挥大数据的大价值, 是人工智能发展过程中不可避免的趋势, 也是社会进步的重要推力。随着数据量和数据种类的增加, 实体解析与人工智能等技术之间的联系也会越来越紧密。鉴于此, 本文从某种程度上探讨和研究了实体解析理论中经典的实体解析方法及逻辑思路, 并总结了各方法的性能特点和局限性, 如表 1 所示, 以期为大数据环境下的实体解析研究提供进一步的理论指导和实践参考。

表 1　文中实体解析基本方法的详细比较

实体解析方法	逻辑思路	性能特点	局限性
基于概率决策的实体解析方法	通过计算对应属性值之间的相似度, 并结合属性级阈值和记录级阈值, 来决定两条记录是否匹配。	1.字符串相似度计算过程相对简单; 2.启发式动态调整属性权重。	1.没有一种字符串相似度算法能够适用于所有数据, 需要根据具体情况进行调整; 2.字符串相似度算法通常默认属性之间相互独立, 但现实数据中属性之间存在多种依赖关系。
面向通用扩展的实体解析方法	仅从可扩展层面设计算法, 以大大减少调用通用的“黑盒”函数的次数, 并且不考虑用作匹配、合并记录的“黑盒”函数的具体实现。	1.将精确性和性能分开处理, 并专注于性能; 2.具有良好的可扩展性和较强的灵活性。	如果当前两条记录不匹配, 则在算法完成之前不能确定它们真的不匹配。
基于协同分析的实体解析方法	以迭代的方式对共同出现的多种类型或同种类型的数据对象协同而非独立地进行解析。	能充分考虑已经发现的证据, 从而提高实体解析精度。	协同解析过程是一个不断迭代的过程, 因此其计算代价要远大于传统的基于属性相似度的解析过程。
基于众包模式的实体解析方法	通过整合计算机和互联网上未知的大众(工作者), 以完成计算机难以单独完成的人类智能任务, 从而有效提高实体解析准确率。	1.充分利用人的推理和联想能力, 挖掘出隐藏的信息; 2.运用到实体解析的整个流程中; 3.减少对设计实体解析过程的专业人员的依赖, 也能进一步提高算法对不同领域数据的适应性。	1.需要花费一定的时间和金钱; 2.不能保证数据安全和隐私。
基于主动学习的实体解析方法	通过主动学习器从未标注数据对象集中有策略性地选出某些数据对象以让专家进行标注, 标注后的数据对象有助于以最快的速度加强分类器。	1.减轻了人工标注的负担, 而且在准确性和可靠性上有一定保障; 2.在标注代价较小的情况下获得较高的性能; 3.可作为获得训练数据的有效手段。	1.需要交互式指导; 2.匹配和非匹配的数据对象通常非常不平衡; 3.需要少量训练数据。
面向实时应用的实体解析方法	通过索引技术从数据集中选取一个记录块(记录子集), 用该记录块解析到来的查询记录流, 这一过程的持续时间(响应时间)为亚秒级。	1.可使用已有的分类、比较函数; 2.能实现较好的查询时间和匹配质量; 3.块键算法能在一定程度上满足大数据环境下的实时解析需要。	1.必须考虑所选择分类、比较函数的复杂性; 2.数据集通常不包含提供丰富特征空间的文档, 只包含简短的字符串或数字等。
隐私保护下的实体解析方法	确定两个或多个组织之间的数据集中是否包含描述同一现实世界实体的记录, 同时又不会向对方或其他任何组织泄露除匹配记录以外的任何信息。	1.只有最终匹配的记录可被各数据源之间共享, 其他未匹配的记录信息均未被泄露; 2.可满足大数据环境下的数据隐私保护需求。	1.不容易合理计算多条记录之间的相似度; 2.在处理存在质量问题的数据时, 查全率偏低; 3.计算成本较高。

展望未来, 随着大数据应用需求持续加大, 以及研究不断深入和拓展, 实体解析技术在数据混杂性、隐私保护和分布式环境等方面仍存在一些开放性的研究方向。

(1) 数据混杂性下的实体解析

在真实数据集中, 数据对象的属性值可能存在诸

多问题, 例如错误、缺失、重复、表示多样化和随时间演化等[82]。然而, 传统上用于实体解析的一些字符串相似度函数却并不能很好地克服这些问题, 即不能合理计算出存在诸多问题的属性值之间的相似度。显然, 为计算出数据混杂性环境下属性值之间的相似度, 需要一种解析能力更强的数据模型, 以充分发掘出海量数据中蕴藏的丰富内在信息和发现更好的特征, 继而据此计算出相似度。事实上, 深度学习模型本质上是通过构建具有很多隐层的机器学习模型和海量的训练数据来学习文本中更有用的特征, 从而最终提升分类或预测的准确性[83-84]。因此如何基于深度学习模型中的词嵌入(Word Embedding, 将记录对表示为 N-gram 嵌入)和深度神经网络(Deep Neural Networks, 将记录对分类为匹配和不匹配)技术进行数据混杂性下的实体解析是一个有意义的研究问题。

(2) 隐私保护下的实体解析

大数据时代下, 人们在享有大数据共享带来便捷化、精准化的同时, 也逐渐关注随之而来的数据安全、个人信息保护等问题。考虑到实体解析是实现大数据共享的一种关键技术, 因此研究如何在数据隐私保护的前提下有效进行实体解析将具有广泛而深远的现实意义[85]。事实上, 现有隐私保护下的实体解析方法存在一些局限, 严重阻碍了其在现实世界中的应用, 例如可扩展性差导致其无法应用于大数据集、容错性差导致其会丢失较多真实匹配的记录或比较能力差导致其无法应用于多方数据源之间多记录的比较。因此解决现有解析方法中存在的上述问题, 使其更好地应用到现实世界中, 将成为未来研究的热点和发展方向。

(3) 分布式环境下的实体解析

大数据时代的数据量达到 PB 或以上级别, 而且数据演化速度也较快, 因此如何通过分布式计算来高效利用这些数据, 是实体解析研究的另一主要任务, 例如基于 Spark 或 MapReduce 的分布式实体解析[86]。分布式实体解析通常基于分块技术, 然而现实世界中非均匀的数据分布, 往往会导致其产生大小不一的、冗余的分块, 这给分布式实体解析带来了负载均衡的挑战(涉及在多台机器上并行计算)。因此如何更合理地分配子解析任务, 以及提出有效的冗余去除技术, 以进一步提高分布式环境下的实体解析效率将是未来研究中的重点。

参考文献:

[1] 孟小峰, 杜治娟. 大数据融合研究: 问题与挑战[J]. 计算机研究与发展, 2016, 53(2): 231-246. (Meng Xiaofeng, Du Zhijuan. Research on the Big Data Fusion: Issues and Challenges[J]. Journal of Computer Research and Development, 2016, 53(2): 231-246.)

[2] 李建中, 王宏志, 高宏. 大数据可用性的研究进展[J]. 软件学报, 2016, 27(7): 1605-1625. (Li Jianzhong, Wang Hongzhi, Gao Hong. State-of-the-Art of Research on Big Data Usability[J]. Journal of Software, 2016, 27(7): 1605-1625.)

[3] Dong X L, Srivastava D. Big Data Integration[C]// Proceedings of the 29th International Conference on Data Engineering. 2013: 1245-1248.

[4] Getoor L, Machanavajjhala A. Entity Resolution: Theory, Practice & Open Challenges[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2018-2019.

[5] Dong X L, Rekatsinas T. Data Integration and Machine Learning: A Natural Synergy[C]// Proceedings of the 2018 International Conference on Management of Data. 2018: 1645-1650.

[6] Newcombe H B, Kennedy J M, Axford S J, et al. Automatic Linkage of Vital Records[J]. Science, 1959, 130(3381): 954-959.

[7] Talburt J R. Entity Resolution and Information Quality[M]. Elsevier, 2011.

[8] Fellegi I P, Sunter A B. A Theory for Record Linkage[J]. Journal of the American Statistical Association, 1969, 64(328): 1183-1210.

[9] Cohen W W, Ravikumar P, Fienberg S E. A Comparison of String Metrics for Matching Names and Records[A]// KDD Workshop on Data Cleaning & Object Consolidation[M]. 2003, 4(2): 73-78.

[10] Naumann F, Herschel M. An Introduction to Duplicate Detection[J]. Synthesis Lectures on Data Management, 2010, 2(1): 1-87.

[11] Doan A, Halevy A, Ives Z. Principles of Data Integration[M]. Elsevier, 2012.

[12] Levenshtein V I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals[J]. Doklady Akademii Nauk SSSR, 1965, 163(4): 845-848.

[13] Jaro M A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida[J]. Journal of the American Statistical Association, 1989, 84(406): 414-420.

[14] Winkler W E, Thibaudeau Y. An Application of the Felligi Sunter Model of Record Linkage to the 1990 US Decennial Census[R]. Washington: US Bureau of the Census, 1991.

[15] On B W, Koudas N, Lee D, et al. Group Linkage[C]// Proceedings of the 23rd International Conference on Data Engineering. 2007: 496-505.

[16] Bilenko M, Mooney R J, Cohen W W, et al. Adaptive Name Matching in Information Integration [J]. IEEE Intelligent

Systems, 2003, 18(5): 16-23.

[17] Cohen W W, Ravikumar P, Fienberg S E. A Comparison of String Distance Metrics for Name-Matching Tasks[C]// Proceedings of the 2003 International Joint Conference on Artificial Intelligence. 2003: 73-78.

[18] Monge A E, Elkan C. The Field Matching Problem: Algorithms and Applications[C]// Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. 1996: 267-270.

[19] Whang S E, Garcia-Molina H. Developments in Generic Entity Resolution[A]// Bulletin of the Technology Committee on Data Engineering[M]. IEEE Computer Society, 2011.

[20] Benjelloun O, Garcia-Molina H, Kawai H, et al. Generic Entity Resolution in the Serf Project[R]. Stanford InfoLab, 2006.

[21] Benjelloun O, Garcia-Molina H, Menestrina D, et al. Swoosh: A Generic Approach to Entity Resolution[J]. The VLDB Journal, 2009, 18(1): 255-276.

[22] Kawai H, Garcia-Molina H, Benjelloun O, et al. P-Swoosh: Parallel Algorithm for Generic Entity Resolution[R]. 2006.

[23] Kim H S, Lee D. Parallel Linkage[C]// Proceedings of the 16th ACM Conference on Information and Knowledge Management. 2007: 283-292.

[24] Kawai H, Garcia-Molina H, Benjelloun O, et al. Bufoosh: Buffering Algorithms for Generic Entity Resolution[R]. 2006.

[25] Benjelloun O, Garcia-Molina H, Gong H, et al. D-Swoosh: A Family of Algorithms for Generic, Distributed Entity Resolution[C]// Proceedings of the 27th International Conference on Distributed Computing Systems. IEEE, 2007.

[26] Bhattacharya I, Getoor L. Collective Entity Resolution in Relational Data[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 5-18.

[27] Malin B, Airoldi E, Carley K M. A Network Analysis Model for Disambiguation of Names in Lists[J]. Computational & Mathematical Organization Theory, 2005, 11(2): 119-139.

[28] 高学东, 黄月. 异质对象协同实体解析的联合聚类算法[J]. 系统工程理论与实践, 2015, 35(4): 997-1004. (Gao Xuedong, Huang Yue. Co-Clustering Algorithm for Collective Entity Resolution of Multi-Typed Objects[J]. Systems Engineering - Theory & Practice, 2015, 35(4): 997-1004.)

[29] Dong X, Halevy A, Madhavan J. Reference Reconciliation in Complex Information Spaces[C]// Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. 2005: 85-96.

[30] 孙琛琛, 申德荣, 寇月, 等. 面向关联数据的联合式实体识别方法[J]. 计算机学报, 2015, 38(9): 1739-1754. (Sun Chenchen, Shen Derong, Kou Yue, et al. A Related Data Oriented Joint Entity Resolution Approach[J]. Chinese Journal of Computers, 2015, 38(9): 1739-1754.)

[31] Kalashnikov D V. Domain-Independent Data Cleaning via Analysis of Entity-Relationship Graph [J]. ACM Transactions on Database Systems, 2006, 31(2): 716-767.

[32] 冯剑红, 李国良, 冯建华. 众包技术研究综述[J]. 计算机学报, 2015, 38(9): 1713-1726. (Feng Jianhong, Li Guoliang, Feng Jianhua. A Survey on Crowdsourcing[J]. Chinese Journal of Computers, 2015, 38(9): 1713-1726.)

[33] Lee J, Cho H, Park J W, et al. Hybrid Entity Clustering Using Crowds and Data[J]. The VLDB Journal, 2013, 22(5): 711-726.

[34] Wang J, Kraska T, Franklin M J, et al. CrowdER: Crowdsourcing Entity Resolution[J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1483-1494.

[35] Wang S, Xiao X, Lee C H. Crowd-Based Deduplication: An Adaptive Approach[C]// Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. 2015: 1263-1277.

[36] Demartini G, Difallah D E, Cudré-Mauroux P, et al. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking[C]// Proceedings of the 21st International Conference on World Wide Web. 2012: 469-478.

[37] Liu X, Lu M, Ooi B C, et al. CDAS: A Crowdsourcing Data Analytics System[J]. Proceedings of the VLDB Endowment, 2012, 5(10): 1040-1051.

[38] Settles B. Active Learning Literature Survey[R]. University of Wisconsinmadison, 2009.

[39] 杨文柱, 田潇潇, 王思乐, 等. 主动学习算法研究进展[J]. 河北大学学报：自然科学版, 2017, 37(2): 216-224. (Yang Wenzhu, Tian Xiaoxiao, Wang Sile, et al. Recent Advances in Active Learning Algorithms[J]. Journal of Hebei University: Natural Science Edition, 2017, 37(2): 216-224.)

[40] Sarawagi S, Bhamidipaty A. Interactive Deduplication Using Active Learning[C]// Proceeding of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002: 269-278.

[41] Tejada S, Knoblock C A, Minton S. Learning Domain-Independent String Transformation Weights for High Accuracy Object Identification[C]// Proceeding of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002: 350-359.

[42] Bellare K, Iyengar S, Parameswaran A G, et al. Active Sampling for Entity Matching[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012: 1131-1139.

[43] Arasu A, Kaushik R. On Active Learning of Record Matching Packages[C]// Proceeding of the 10th ACM SIGMOD International Conference on Management of Data. 2010: 783-794.

[44] Qian K, Popa L, Sen P. Active Learning for Large-Scale Entity Resolution[C]// Proceedings of the 2017 ACM Conference on Information and Knowledge Management.

2017: 1379-1388.

[45] Fisher J, Christen P, Wang Q. Active Learning Based Entity Resolution Using Markov Logic[C]// Proceedings of the 2016 Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2016: 338-349.

[46] Fu Y, Zhu X, Li B. A Survey on Instance Selection for Active Learning[J]. Knowledge and Information Systems, 2013, 35(2): 249-283.

[47] Ramadan B. Indexing Techniques for Real-Time Entity Resolution[D]. Canberra: The Australian National University, 2016.

[48] Ramadan B, Christen P, Liang H, et al. Dynamic Sorted Neighborhood Indexing for Real-Time Entity Resolution[J]. Journal of Data and Information Quality (JDIQ), 2015, 6(4): Article No.15.

[49] Bayardo R J, Ma Y, Srikant R. Scaling up All Pairs Similarity Search[C]// Proceedings of the 16th International Conference on World Wide Web. 2007: 131-140.

[50] Broder A Z, Carmel D, Herscovici M, et al. Efficient Query Evaluation Using a Two-Level Retrieval Process[C]// Proceedings of the 12th International Conference on Information and Knowledge Management. 2003: 426-434.

[51] Zobel J, Moffat A. Inverted Files for Text Search Engines[J]. ACM Computing Surveys(CSUR), 2006, 38(2): 6.

[52] Christen P. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(9): 1537-1555.

[53] Christen P, Gayler R, Hawking D. Similarity-Aware Indexing for Real-Time Entity Resolution[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. 2009: 1565-1568.

[54] Ramadan B, Christen P, Liang H, et al. Dynamic Similarity-Aware Inverted Indexing for Real-Time Entity Resolution[C]// Proceedings of the 2013 Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2013: 47-58.

[55] Ramadan B, Christen P. Unsupervised Blocking Key Selection for Real-Time Entity Resolution[C]// Proceedings of the 2015 Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2015: 574-585.

[56] Papadakis G, Koutrika G, Palpanas T, et al. Meta-Blocking: Taking Entity Resolutionto the Next Level[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1946-1960.

[57] Efthymiou V, Papadakis G, Papastefanatos G, et al. Parallel Meta-Blocking: Realizing Scalable Entity Resolution over Large, Heterogeneous Data[C]// Proceedings of the 2015 IEEE International Conference on Big Data. 2015: 411-420.

[58] Dragut E C, Ouzzani M, Elmagarmid A K. Query-Time Record Linkage and Fusion over Web Databases[C]// Proceedings of the 31st International Conference on Data Engineering. IEEE, 2015: 42-53.

[59] Gruenheid A, Dong X L, Srivastava D. Incremental Record Linkage[J]. Proceedings of the VLDB Endowment, 2014, 7(9): 697-708.

[60] Whang S E, Marmaros D, Garcia-Molina H. Pay-as-you-go Entity Resolution[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(5): 1111-1124.

[61] Hall R, Fienberg S E. Privacy-Preserving Record Linkage[C]// Proceedings of the 2013 International Conference on Privacy in Statistical Databases. 2010: 269-283.

[62] 韩姝敏，申德荣，聂铁铮，等. 一种基于隐私保护下的多方记录链接方法[J]. 软件学报，2017, 28(9): 2281-2292. (Han Shumin, Shen Derong, Nie Tiezheng, et al. Multi-Party Privacy-Preserving Record Linkage Apprpoach[J]. Journal of Software, 2017, 28(9): 2281-2292.)

[63] Kuzu M, Kantarcioglu M, Inan A, et al. Efficient Privacy-Aware Record Integration[C]// Proceedings of the 16th International Conference on Extending Database Technology. ACM, 2013: 167-178.

[64] Sweeney L. K-Anonymity: A Model for Protecting Privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[65] Sweeney L. Achieving K-Anonymity Privacy Protection Using Generalization and Suppression[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 571-588.

[66] Kantarcioglu M, Jiang W, Malin B. A Privacy-Preserving Framework for Integrating Person-Specific Databases[C]// Proceedings of the 2008 International Conference on Privacy in Statistical Databases. Springer, 2008: 298-314.

[67] Dwork C. Differential Privacy: A Survey of Results[C]// Proceedings of the 5th International Conference on Theory and Applications of Models of Computation. 2008: 1-19.

[68] Inan A, Kantarcioglu M, Ghinita G, et al. Private Record Matching Using Differential Privacy[C]// Proceedings of the 13th International Conference on Extending Database Technology. ACM, 2010: 123-134.

[69] 张啸剑，孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报，2014, 37(4): 927-949. (Zhang Xiaojian, Meng Xiaofeng. Differential Privacy in Data Publication and Analysis[J]. Chinese Journal of Computers, 2014, 37(4): 927-949.)

[70] 孟小峰，张啸剑. 大数据隐私管理[J]. 计算机研究与发展，2015, 52(2): 265-281. (Meng Xiaofeng, Zhang Xiaojian. Big Data Privacy Management[J]. Journal of Computer Research and Development, 2015, 52(2): 265-281.)

[71] Schnell R, Bachteler T, Reiher J. Privacy-Preserving Record Linkage Using Bloom Filters[J]. BMC Medical Informatics and Decision Making, 2009, 9(1): 41.

[72] Bloom B H. Space/Time Trade-offs in Hash Coding with Allowable Errors[J]. Communications of the ACM, 1970,

13(7): 422-426.

[73] Jain N, Dahlin M, Tewari R. Using Bloom Filters to Refine Web Search Results[C]// Proceedings of the 8th International Workshop on the Web and Databases. 2005: 25-30.

[74] Durham E A, Kantarcioglu M, Xue Y, et al. Composite Bloom Filters for Secure Record Linkage[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12): 2956-2968.

[75] Vatsalan D, Christen P. Scalable Privacy-Preserving Record Linkage for Multiple Databases[C]// Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. ACM, 2014: 1795-1798.

[76] Randall S M, Ferrante A M, Boyd J H, et al. Privacy-Preserving Record Linkage on Large Real World Datasets[J]. Journal of Biomedical Informatics, 2014, 50: 205-212.

[77] Jurczyk P, Xiong L. Towards Privacy-Preserving Integration of Distributed Heterogeneous Data[C]// Proceedings of the 2nd PhD Workshop on Information and Knowledge Management. ACM, 2008: 65-72.

[78] Sheikh R, Mishra D K, Kumar B. Secure Multiparty Computation: From Millionaires Problem to Anonymizer[J]. Information Security Journal: A Global Perspective, 2011, 20(1): 25-33.

[79] Goldwasser S, Micali S, Rackoff C. The Knowledge Complexity of Interactive Proof Systems[J]. SIAM Journal on Computing, 1989, 18(1): 186-208.

[80] Quisquater J, Quisquater M, Quisquater M, et al. How to Explain Zero-Knowledge Protocols to Your Children[C]// Proceedings of the 1989 Conference on the Theory and Application of Cryptology. 1989: 628-631.

[81] Lindell Y, Pinkas B. Secure Multiparty Computation for Privacy-Preserving Data Mining[J]. Journal of Privacy & Confidentiality, 2009, 25(2): 761-766.

[82] 维克托·迈尔-舍恩伯格, 肯尼思·库克耶. 大数据时代：生活、工作与思维的大变革[M]. 杭州: 浙江人民出版社, 2013. (Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution That will Transform How We Live, Work, and Think[M]. Hangzhou: Zhejiang People's Publishing House, 2013.)

[83] Kooli N, Allesiardo R, Pigneul E. Deep Learning Based Approach for Entity Resolution in Databases[C]// Proceedings of the 2018 Asian Conference on Intelligent Information and Database Systems. 2018: 3-12.

[84] Schmidhuber J. Deep Learning in Neural Networks: An Overview[J]. Neural Networks, 2015, 61: 85-117.

[85] Vatsalan D, Sehili Z, Christen P, et al. Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges[A]// Zomaya A, Sakr S. Handbook of Big Data Technologies[M]. 2017: 851-895.

[86] Altowim Y, Mehrotra S. Parallel Progressive Approach to Entity Resolution Using MapReduce[C]// Proceedings of the 33rd International Conference on Data Engineering. 2017: 909-920.

利益冲突声明:

作者声明不存在利益冲突关系。

收稿日期: 2018-12-07
收修改稿日期: 2019-01-17

Reviewing Basic Methods of Entity Resolution

Gao Guangshang
(Business School, Guilin University of Technology, Guilin 541004, China)

Abstract: [Objective] This paper discusses the classical entity resolution methods and logical thinking in entity resolution theory. **[Coverage]** Google Scholar and CNKI were respectively used to search literatures with the keywords "Entity Resolution", "Collective Analysis", "Crowdsourced", "Active Learning", "Privacy-Preserving" and "Entity Resolution" in Chinese. I then obtained a total of 86 representative literatures in conjunction with topic screening, intensive reading and retrospective method. **[Methods]** For each entity resolution method, the paper first summarizes and analyzes the basic idea of the method, and presents the resolution process through illustration, and then focuses on analyzing the key strategies, algorithms or techniques adopted by the existing research in the process of implementation of the method. **[Results]** Entity resolution is the basic operation of data quality management, and the key step to find the value of data. **[Limitations]** There is no in-depth analysis of the evaluation indicators and application of each entity resolution method. **[Conclusions]** Although existing entity resolution methods can meet the requirements of most applications to some extent, they still face challenges in data heterogeneity, privacy protection and distributed environment in the big data environment.

Keywords: Entity Resolution Collective Analysis Crowdsourced Active Learning Privacy-Preserving