基于代表记录的增量实体解析方法研究

高广尚◎著

非外借

科学出版社

基于代表记录的增量实体解析方法研究

高广尚 著

科学出版社

北京

内 容 简 介

本书从记录、代表记录、相似记录、记录簇、传递闭包、并查集、实体、实体解析（entity resolution，ER）、增量实体解析（incremental entity resolution，IER）等概念出发，研究了基于代表记录的增量实体解析方法。本书共 8 章，内容包括绪论、相关研究、基于代表记录的增量实体解析方法研究框架和关键问题、基于优先队列的代表记录产生模型构建方法研究、基于并查集的相似记录聚类模型构建方法研究、基于代表记录的记录簇调整模型构建方法研究、基于代表记录的增量实体解析方法的有效性实验、总结与展望。

本书可供高等院校计算机、数据分析、信息管理等专业的本科生和硕士研究生使用，也可供数据库、数据质量和数据集成领域研究人员和从业者参考。

图书在版编目(CIP)数据

基于代表记录的增量实体解析方法研究/高广尚著. —北京：科学出版社，2019.8

ISBN 978-7-03-060358-6

Ⅰ. ①基… Ⅱ. ①高… Ⅲ. ①信息检索技术-增量-聚类分析法-研究 Ⅳ. ①G254.91

中国版本图书馆 CIP 数据核字（2018）第 302673 号

责任编辑：冯 涛 吴超莉 / 责任校对：赵丽杰
责任印制：吕春珉 / 封面设计：东方人华平面设计部

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

三河市骏走印刷有限公司印刷
科学出版社发行 各地新华书店经销

*

2019 年 8 月第 一 版 开本：B5（720×1000）
2019 年 8 月第一次印刷 印张：9 1/4
字数：200 000

定价：68.00 元

（如有印装质量问题，我社负责调换〈骏杰〉）
销售部电话 010-62136230 编辑部电话 010-62139281

前　言

在大数据时代背景下，数据集上的实体解析正面临着数据更新快、数据规模大和数据质量差的问题。这不仅让此前形成的解析结果很快失效，而且让随之不断演化的相似记录更加难以解析。现有聚类算法并不太适合对数据集中不断演化的相似记录进行有效的解析，因此，如何在不对整个数据集进行重新聚类的情况下对其中不断演化的相似记录进行有效解析，已成为增量实体解析研究中亟待解决的重要问题。

在对国内外现有实体解析、增量实体解析研究中所涉及的诸多聚类方法、聚类算法进行深入研究分析的基础上，本书作者提出通过产生更能代表记录簇的代表记录并基于代表记录集，让与演化的记录紧密相关的记录簇进行自适应调整，以实现增量实体解析这一新思路，同时明确本书要研究的 3 个关键问题：①如何在解析数据集时产生更能代表记录簇的代表记录，以有利于其中潜在相似记录的合并或排除；②如何找出解析过程中遗漏的潜在相似记录，并将它们划分到同一个记录簇，从而让代表记录的代表性进一步增强；③如何在数据集不断演化时基于代表记录集，让相关的记录簇进行快速、有效的自适应调整。本书针对这 3 个关键问题，主要开展了 3 个方面的研究：①在对代表记录产生方法，以及不具代表性的代表记录可能导致潜在相似的记录被排除在记录簇外，而不相似的记录被保留在记录簇内的情况进行分析的基础上，提出基于优先队列的代表记录产生模型，旨在产生更具代表性的代表记录。模型主要关注 3 部分内容，即待比较记录与代表记录间的相似性判定、待比较记录与代表记录间的合并，以及代表记录的产生。②在对相似记录聚类方法，以及因对应属性值彼此存在细微判别而出现相似性异常，或解析方法本身存在局限性而导致潜在相似的记录无法聚类到一起的情况进行分析的基础上，提出基于并查集的相似记录聚类模型，旨在进一步识别出潜在相似记录并将它们聚类到一起，从而让代表记录的代表性进一步得到增强。模型主要关注 3 部分内容，即基于重要属性生成高质量 Key、基于多趟扫描结果计算传递闭包，以及基于并查集合并相似记录。③在对记录簇调整方法，以及数据集上因不断出现新增、删除或更新记录而导致此前解析结果中的记录簇无法进行快速、有效调整进行分析的基础上，提出基于代表记录的记录簇调整模型，旨在让在数据集不断演化时记录簇调整过程更具针对性、有效性、稳定性和快速性。模型主要关注 3 部分内容，即潜在相似代表记录的确定、相关的记录簇自适应调整，以及记录簇的代表记录更新。

本书采用 Cora 数据集分别对提出模型的有效性、可行性进行验证，从整体上将书中方法与相关性聚类方法进行了对比，并进一步将它们的结果和 Cora 数据集中人工划分的结果进行了对比。实验结果表明，书中方法相较于相关性聚类方法在解析效率、解析精度等方面都有一定的优势。

全书共分 8 章。第 1 章介绍研究背景、相关概念、研究目标和意义、研究思路与研究方法等；第 2 章从经典聚类算法下的实体解析方法、一般聚类算法下的实体解析方法、增量聚类算法下的增量实体解析方法 3 个方面对相关研究进行综述，并分析现有研究方法中的不足；第 3 章介绍基于代表记录的增量实体解析方法研究框架和关键问题；第 4 章研究基于优先队列的代表记录产生模型构建方法；第 5 章研究基于并查集的相似记录聚类模型构建方法；第 6 章研究基于代表记录的记录簇调整模型构建方法；第 7 章验证基于代表记录的增量实体解析方法的有效性；第 8 章进行总结与展望。

增量实体解析是一个全新的研究领域，涉及各种算法和技术，撰写本书可谓一项极大的挑战，虽然作者秉承“工匠精神”，在全书的结构、内容和行文等各个方面力求完美，但疏漏仍在所难免，请广大读者批评指正。

目　录

第1章 绪 论

1.1 研究背景

在信息时代，以数据为中心的信息系统正在得到越来越广泛的应用，然而，这些数据并非总是正确无误的，其中可能存在各种各样的错误，如重复、不一致、不正确或不完整等[1]，或这些数据本身存在不同的描述形式。当这些情形出现在记录属性值上时，数据集中相似记录之间就会存在一些细微的差别，而它们却有可能表示同一现实世界实体[2]。例如，邮箱列表中可能包含多条实际上都表示同一物理地址的记录，商品列表中可能包含多条实际上都表示同一实际商品的记录等。因此，如何快速、有效地识别出数据集中那些因属性值上存在细微差别但仍然描述同一实体的 n（$n>1$）条相似记录，一直是实体解析研究中的重点[2-4]。

实体解析这一概念的雏形最初出现在对文件清单目录进行比较的过程中。早在 1969 年，Fellegi 等就在工作中发现他们通常面临这样一个问题：如何对不同的清单进行比较，以除去其中存在的重复目录。之后，他们将解决这个问题的方法称为记录链接（record linkage）[5]。随着关系数据库系统的发展，出现了另一个与之类似的识别问题：如何在数据集中找出并合并代表同一实体的所有相似记录[6]。由于在诸多应用领域中涉及数据集，如人口普查记录的处理与分析[7]、数据清洗[8]、信息集成[9]、模糊关键字查询[10]、诈骗检测[11]、文本聚类[12]、执法和反恐[13]等，因此解决这一识别问题的方法在各个领域中有着不同的名称，包括记录链接[14]、合并/清洗（merge/purge）[6]、重复数据删除（deduplication）[15]、参考协调（reference reconciliation）[16]、对象识别（object identification）[17]和其他名称[4,18,19]。2007 年，一篇由斯坦福大学信息实验室研究人员发表的文章正式将这些类似的识别过程统称为实体解析[20]。

现有大部分实体解析研究主要针对如何在静态数据集中进行实体解析（记录属性值上原本就存在的细微差别。实体解析过程涉及数据集中的所有记录，更具体地说是对数据集中所有记录进行聚类分析。然而，当数据集大部分时间处在动态变化时，即每段时间内数据集中都可能会新增、删除或更新记录，如新签约的客户、新出生的婴儿、新注册的学生和新注册的雇员等（新增），人员离职、个人离世、业务不再激活等（删除），个人姓名、邮箱和地址的变更、学校招生状况或

学生就业层次的变更、操作导致的数据不规范或错误等（更新）[21,22]，仍然采用现有的方法来对其进行解析，在解析速度和解析效率方面将不能满足高质量解析需求。这是因为现有方法每出现一次变化就对整个数据集重新解析一次的解析过程所需时间开销大。因此，针对动态数据集的实体解析更具挑战性，更为重要，也更有意义[4,23]。

为对静态数据集中潜在相似的记录进行聚类，现有研究一般采用“排序&合并”的方法[6,24]。具体过程可概括为：首先将整个数据集按字典序重排（采用 Key 来标志记录，或把整条记录看成一个字符串），这样潜在相似的记录就会被排列在较接近的位置，从而就可以在相对集中的范围内做记录的成对比较（pair-wise）[25]，并计算出记录间的相似性值，进而确定它们是否为相似记录。在这些研究中，Monge[26]提出的基于优先队列（priority queue）算法和 Hernandez 等[6]提出的多趟近邻排序（multi-pass sorted-neighborhood）算法较有影响力。其中，基于优先队列算法的优点是能减少记录比较的次数，提高匹配的效率，而且几乎不受数据规模的影响，因而能很好地适应数据规模的变化；缺点是产生的代表记录往往缺乏代表性，因而代表记录未能有效地帮助聚类静态数据集中潜在相似的记录，从而出现相似记录漏配，这又将影响代表记录的构成。多趟近邻排序算法的优点是精度高，能尽可能多地发现潜在相似的记录；缺点是每趟扫描所使用的 Key 较单一，以及滑动窗口大小较难选取（因为当窗口较大时，所进行的比较次数就会偏多，而有些比较是没必要的；当窗口较小时，可能出现相似记录漏配）。在对静态数据集进行一次解析后，将会得到一个包含诸多记录簇的聚类结果（解析结果）。

由于数据集上的变化过程本质上就是记录层面上的记录增量过程（新增记录、删除记录和更新记录），或更具体地说，是属性层面上的数据演化过程，因此动态数据集上的实体解析又称为增量实体解析[27-29]。事实上，增量实体解析过程的核心是数据集因出现上述变化而成为动态数据集时，原有聚类结果中的记录簇应做出相应调整以反映这种变化，否则这些记录簇将会变得过时，即记录簇内包含的记录可能不都表示同一个实体。

与实体解析相比，增量实体解析的研究相对较少，其中比较有影响的是基于相关性聚类算法的增量实体解析[30]。该增量实体解析方法的优点如下：一是增量解析过程涉及新增记录、删除记录和更新记录 3 种操作，而不仅仅是新增记录操作[28,31,32]；二是对每次操作结果进行解析时不需要对数据集中所有记录进行重新聚类分析，且在准确率上有一定的保障；三是能部分利用此前的解析结果来加快当前的解析过程，并能部分修正此前解析结果中存在的错误。其缺点是增量实体解析过程被视为如何在相似图形上找到一个合适子图的过程（进行子图划分），该

过程涉及的计算复杂度较高，且很难找到一种精确的划分方法，因为它纯粹从图形数据角度来进行考虑。

基于以上背景和现有研究方法中存在的不足，本书提出基于代表记录的增量实体解析方法，旨在形成记录簇的同时产生更具代表性的代表记录，而更具代表性的代表记录又反过来进一步促进记录簇更好地形成，通过研究记录如何产生、更新，以及利用代表记录进行实体解析的方法来增强和补充动态数据集上的增量实体解析体系。为此，本书从3个方面对提出的方法展开研究：①构建基于优先队列的代表记录产生模型，以产生更具代表性的代表记录，使代表记录不仅有助于减少生成的记录簇数量，而且能减少记录比较的次数，最终有利于潜在相似记录的合并或排除；②构建基于并查集的相似记录聚类模型，以聚类那些因对应属性值彼此存在细微判别而出现相似性异常，以及实体解析方法本身存在局限性而无法被划分到一起的潜在相似记录，最终让聚类结果中记录簇的构成更具合理性，进一步增强代表记录的代表性；③构建基于代表记录的记录簇调整模型，以在数据集演化时通过代表记录集让相关的记录簇做出自适应调整，同时让调整过程充分利用此前聚类结果中的相关信息，甚至修正其中存在的错误。

1.2 相关概念

1.2.1 记录

记录（records）是一条包含 n（$n \geqslant 1$）个属性的元组，用 $(V_1, V_2, \cdots, V_n)$ 表示一条记录，其中，V_i 表示第 i 个属性的值。

例如，一条记录可用包含5个属性的符号，即 $(V_1, V_2, V_3, V_4, V_5)$。其中，属性 V_1（name）、V_2（city）、V_3（zipcode）、V_4（phn）和 V_5（represent）的值分别是“Carrefour”“Beijing”“90015”“83950321”“Morgan”。

1.2.2 代表记录

代表记录（canonical records）同样是一条包含 n（$n \geqslant 1$）个属性的元组，如包含 recIDs 属性等，但是其属性上的值不是一个字符串，而是一个字符串列表（即字符串个数不确定），其中每个字符串都对应着其出现的频次。此外，recIDs 属性上的值由其所代表的记录簇内部所有记录的 ID 构成。表 1.1 所示的代表记录中包含两条代表记录，即 r_1 和 r_2。

表 1.1　代表记录

ID	recIDs	first	last	DOB	scode
r_1	5，6	Mary:2	Smith:2	19990921:2	H17:1，G55:1
r_2	10，11，12	Eddie:2，Edgar:1	Jones:3	20001104:3	G34:2，H15:1
…	…	…	…	…	…

1.2.3　相似记录

相似记录（approximately duplicate records）是指在 n（$n \geq 1$）个对应属性上彼此相似的记录。例如，在表 1.2 中，ID 为 5 和 6 的两条记录在 3 个对应属性上（first、last 和 DOB）因字符串相同（或比较时的相似性值大于给定阈值）而彼此相似，因而这两条记录被认为是相似记录。值得一提的是，这两条相似记录可保存在某条代表记录的 recIDs 属性中，如表 1.1 所示。

表 1.2　相似记录

ID	first	last	DOB	scode
5	Mary	Smith	19990921	H17
6	Mary	Smith	19990921	G55

1.2.4　记录簇

记录簇（record clusters）就是由若干相似记录聚类在一起而形成的记录集合。为计算出数据集中哪些记录彼此相似，需要通过笛卡儿积来进行逐记录对比较。例如，表 1.3 中的数据集包含 6 条记录（$r_1 \sim r_6$），因而需要进行 $6\times(6-1)/2=15$ 次记录对比较才能将其中所有记录两两比较完毕。

表 1.3　个人信息数据集

记录 ID	名	姓	年龄	街道名	城区	出生日	出生月	出生年
r_1	John	Smith	18	Miller st	Dickson	12	11	1970
r_2	Jonny	Smith	73	Miller st	Dixon	11	10	1970
r_3	Joan	Smith	73	Dawson cr	Lyneham	11	12	1979
r_4	Max	Miller	73	Dawson cr	Lyneham	11	2	1969
r_5	Sal	Bass	67	Miles rd	Ainslie	28	5	1981
r_6	Sally	Bass	64	Miles rd	Ainslie	23	5	1981

表 1.4 中列出了这些记录间的相似性值（假定通过某公式算出），以及它们与阈值比较后的匹配状态（相似状态）。其中，匹配状态的判定过程称为成对分类技术[4]，即相似性值大于给定阈值（如 5.0）就匹配（相似），否则不匹配（不相似）[33]。在表 1.4 中共有 4 对记录是匹配的。

表 1.4　记录对匹配与否的相关信息

候选记录对	相似性值（函数 SimSum 算出）	匹配状态（阈值β=5.0）
(r_1, r_2)	5.20	匹配
(r_1, r_3)	3.30	不匹配
(r_1, r_4)	1.15	不匹配
(r_1, r_5)	0	不匹配
(r_1, r_6)	0	不匹配
(r_2, r_3)	5.05	匹配
(r_2, r_4)	2.70	不匹配
(r_2, r_5)	0	不匹配
(r_2, r_6)	0	不匹配
(r_3, r_4)	5.25	匹配
(r_3, r_5)	0	不匹配
(r_3, r_6)	0	不匹配
(r_4, r_5)	0	不匹配
(r_4, r_6)	0	不匹配
(r_5, r_6)	6.20	匹配

可依据表 1.4 中记录对的匹配信息和相似性值构造出一个相似图形（similarity graph）[34]，如图 1.1 所示。其中，结点表示一条记录，结点间的实线边表示两条记录相似，实线边上的数值表示相似性值（如 5.25），或称为边的权值。4 条实线边表示存在 4 个匹配。所有由实线相连的子图形看成一个记录簇。

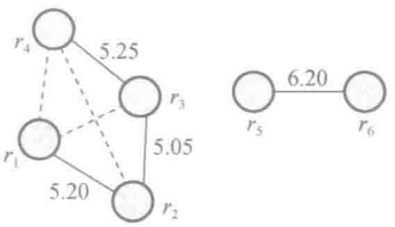

图 1.1　记录簇

1.2.5　传递闭包

传递闭包（transitive closure）是指：如果记录r_i与r_j相似，同时记录r_j与r_k相似，那么记录r_i与r_k也应相似，即记录间传递相似性，使它们都表示同一个实体[24,35]。另一种特殊情况是，如果两条记录彼此相似，那么它们也形成传递闭包。

1.2.6 并查集

在计算机科学中，并查集（union-find sets）是一种树形的数据结构，用于处理一些不相交集合（disjoint sets）的合并与查询问题[24]。其中，Union（合并）操作用来将两条记录各自所在的组进行合并以形成一个组；Find（查找）操作用来查找某记录是否在某个组中。

Union($r_i r_j$)操作：将分别包含任意记录r_i和r_j的两个子组（S_{r_i}和S_{r_j}）合并成一个新的组，即并集$S_{r_i} \cup S_{r_j}$。选择这个新的组$S_{r_i} \cup S_{r_j}$作为代表来替代子组S_{r_i}和S_{r_j}。通常来说，合并之前，应先判断两条记录是否属于同一个组，这可用 Find(r_i)操作来实现。

Find(r_i)操作：返回记录r_i所在组的代表记录[24]。在判断两条记录是否属于同一个组时，只要比较相应的代表记录是否相同即可。

1.2.7 实体

实体（entity）用来描述现实世界的事物或对象，如某个人、地点或物品[36]。实体表现为一系列实体特征，称为属性，所有属性值的组合则提供了关于该特定实体的信息。例如，作为现实中的个人，常见的属性有姓名、家庭住址、生日等；作为现实中的商品，常见的属性通常包括型号、尺寸、生产厂家，或通用产品码等；作为文章的引用，常见的属性为作者、题名、来源（venue）、地址、年份和页码等。因此，数据集中的记录（拥有若干属性）在某种程度上就是实体的数字表示。在同一数据集中可能存在多条记录对应于同一个实体的情形，因为属性值上可能存在由操作引发的字符串修改前后有细微差别、字符串本身具有不同的表示等方面的数据质量问题。

1.2.8 实体解析

实体解析定义为识别并合并那些表示同一现实世界实体的记录的过程[37]。具体来说，实体解析就是将数据集中潜在相似的记录进行分组，使组内的记录尽可能彼此相似，而位于不同组的记录尽可能彼此不相似。从聚类角度来看，分组过程其实就是聚类过程，因而得到的记录组其实就是记录簇或簇。

在同一数据集中，实体解析的过程就是首先计算对应属性值间的相似性值（相似度）并判定它是否大于给定阈值（如α），如果大于给定阈值，则表示两个对应属性值彼此相似，否则不相似；然后汇总被判定为相似的对应属性的个数，并判定它是否大于给定阈值（如β），如果大于给定阈值，则表示两条记录彼此相似，否则不相似；最后被判定为相似的两条记录表示同一个实体。

1.2.9 增量实体解析

增量实体解析定义为识别并合并数据集中那些在新增、删除或修改记录后仍表示同一现实世界实体的记录的过程[27,28,30,38,39]。更具体地说，针对数据集中记录不断变化的问题，增量实体解析着重强调通过利用此前的聚类结果让仅与变化记录相关的记录簇进行自适应调整，而不是对变化后的数据集中所有记录重新进行计算。增量实体解析的目标是显著减少在变化数据集上进行解析时所需的时间，同时无损解析质量，并能利用变化记录中的新证据来修正此前解析结果中存在的错误。

1.3 研究目标和意义

1.3.1 研究目标

本书研究的总体目标就是构建一套行之有效的增量实体解析方法体系，以快速、高效地解析并聚类数据集中存在的相似记录，尤其是在数据集中数据更新快、数据规模大和数据质量差的情况下。针对这一目标和现有研究中存在的问题，本书提出首先通过构建基于优先队列的代表记录产生模型来产生更具代表性的代表记录，以利于潜在相似记录的合并或排除，从而让解析过程变得更加高效，反过来，高效的解析过程又会以更加高效的方式继续产生更具代表性的代表记录；然后通过构建基于并查集的相似记录聚类模型，进一步聚类此前遗漏的潜在相似记录，进而让记录簇中代表记录的代表性进一步得到增强；最后，通过构建基于代表记录的记录簇调整模型，让在数据集不断演化时基于代表记录集的记录簇调整过程更具针对性、有效性、稳定性和快速性，最终实现可适用于大数据环境下的面向数据演化的增量实体解析目标。

本书总体研究目标的实现需要构建以下3个具体模型：

1）代表记录产生模型的构建。代表记录产生模型可使产生的代表记录更具代表性，不仅有助于减少解析过程中生成的记录簇数量、有助于减少记录比较的次数，并且有利于相似记录的合并或排除，最终有助于在整体上让解析过程具有高精度、高效率的特性。因此，构建一个代表记录产生模型，将在整个研究中发挥着重要的作用。

2）相似记录聚类模型的构建。相似记录聚类模型可将使那些因为对应属性值彼此存在细微判别而出现相似性异常，以及实体解析方法本身存在局限性，而无法被划分到同一个记录簇中的潜在相似记录尽可能地聚类到一起。因此，构建一