实体分辨研究综述

谭明超[1] 刁兴春[1,2] 曹建军[2]
(解放军理工大学指挥信息系统学院 南京 210007)[1] (总参第63研究所 南京 210007)[2]

摘 要 实体分辨是数据集成、信息检索等领域的重要研究内容,目的是发现数据集合中的不同实体和同一实体的不同描述。将实体分辨过程划分为数据分块、记录比较和匹配决策等3个主要步骤。从记录聚集方式的角度介绍了实体分辨的数据分块方法;从字符串划分粒度的角度分析了实体分辨的记录比较方法;从记录关联方式的角度阐述了实体分辨的决策模型。最后对实体分辨研究下一步需要解决的问题进行了展望。

关键词 实体分辨,数据分块,相似度,决策模型
中图法分类号 TP311 **文献标识码** A

Survey on Entity Resolution

TAN Ming-chao[1] DIAO Xing-chun[1,2] CAO Jian-jun[2]
(College of Command Information Systems, PLA University of Science & Technology, Nanjing 210007, China)[1]
(The 63rd Research Institute of the PLA General Staff Headquarters, Nanjing 210007, China)[2]

Abstract Entity resolution (ER) is one of central issues of data integration and information retrieval. Its purpose is to find all real-world entities from a given dataset, and to cluster the references that refer to the same entity. ER process was partitioned into blocking step, records comparison step and matching decision step. The methods of blocking were summarized according to the way that records are clustered together, and the methods of record pair comparison are surveyed according to the size that strings are grained, and the decision models were introduced according to the way that records associate with each other. At last, the future research issues were discussed.

Keywords Entity resolution, Blocking, Similarity, Decision model

以数据为中心的系统当前已经得到广泛的应用,这些系统通常假设所采集的数据是正确无误的,但现实信息系统中存储的数据大多含有各种各样的错误[1,2]。在一个或多个数据库中,同一个现实对象可能具有不同的表示方法,因此,如何分辨同一实体的不同表示便成为数据集成、数据挖掘和信息检索等各类数据应用需要解决的重要问题之一。实体分辨(Entity Resolution, ER)——也称为记录链接(Record Linkage)、冗余检测(Duplicate Detection)或数据匹配(Data Matching),指的是从一个或多个数据库中分辨描述相同现实实体的不同表示方法,正确识别出数据库中所有不同实体的过程[2]。

从Dunn在1946年提出记录链接的概念以来,实体分辨经历了很长的研究历程[3]。19世纪50年代末,Howard Newcombe等人利用计算机自动化实体分辨过程,提出了概率记录链接的基本思想[4]。基于该思想,Ivan Fellegi和Alan Sunter对实体分辨问题进行了形式化,提出了概率决策模型理论,指出在属性条件独立的假设下可以找到一个最优的概率决策规则[5]。从概率决策模型提出到20世纪90年代,诸多对该模型的改进不断提出,最为突出的是在统计学领域Winkler等人提出了利用字符串近似比较函数处理严格匹配带来的问题,定义了基于概率的匹配权重,并提出了利用EM(最大期望)算法对模型中的条件概率参数进行估计的方法[6]。与此同时,在数据库领域中,研究者也在研究检测和清除冗余记录的技术和方法,但并不是基于Fellegi和Sunter提出的概率决策模型,而是通过记录排序和相似度函数计算记录的相似度[7-9]。在最近10年里,计算机科学领域对实体分辨的研究给予了越来越多的重视,尤其是在数据挖掘、机器学习、信息检索等领域。随着大量数据的不断产生,数据质量问题已经成为影响数据应用的最大障碍,实体分辨作为处理数据质量问题的重要手段变得无处不在[10-12]。

1 实体分辨问题概述

1.1 实体分辨的困难和挑战

自从实体分辨的概念出现至今,虽然各个研究领域提出了各种各样的实体分辨方法,但实体分辨主要面对的困难和挑战基本来自以下2个方面[3-12]:

1)参与实体分辨的数据库通常缺少唯一的实体标识或主键,对缺少实体标识的记录进行分辨必须要用到记录的一般属性(非主键),而这些属性中往往充斥着各种"脏"数据,如何在这种情况下准确分辨出描述同一实体的记录是实体分辨面

到稿日期:2013-06-19 返修日期:2013-08-09 本文受国家自然科学基金(61070714),解放军理工大学预研基金(20110604)资助。
谭明超(1979-),男,博士生,工程师,主要研究领域为数据管理、信息质量,E-mail:tanmch@gmail.com;刁兴春(1964-),男,研究员,博士生导师,主要研究领域为数据工程;曹建军(1975-),男,博士,主要研究领域为数据和信息质量控制的理论与方法、进化算法。

临的主要挑战。

2)记录比较是一种计算复杂度很高的操作,当对一个数据库中的记录进行匹配时,每条记录要与该数据库中所有其他记录进行比较。然而由于匹配记录对数量往往远少于不匹配记录对,绝大部分比较浪费在不匹配记录对之间,当数据库中包含的记录量很大时,如何在保证匹配的准确性和完整性的同时尽量减少需要比较的记录数目便成为实体分辨需要解决的另一挑战。

1.2 问题定义和过程划分

给定描述 k 个现实实体的记录集合 I,实体分辨的目的是将记录集合 I 划分为 k 个子集 $I_1, I_2, \cdots, I_k$,满足 $\forall (n, l \in [1,k], n \neq l)\ I_n \cap I_l = \emptyset$,且 $I_1 \cup I_2 \cup \cdots \cup I_k = I$,使得 $\forall (r_1, r_2 \in I_i,\ i \in [1,k])$,$r_1$ 和 r_2 描述现实中同一实体。

依据处理的先后顺序,实体分辨过程可以划分为 3 个相对独立的步骤,分别是数据分块、记录比较和匹配决策。数据分块步骤的目的是将数据划分为规模较小的分块,为记录比较提供尽量少且可能匹配的候选记录对,提高实体分辨的效率;记录比较步骤是对候选记录对进行比较,利用相似度函数计算各个属性相似度,得到比较向量;匹配决策步骤依据比较向量判断两条记录的匹配状态。这 3 个步骤构成了实体分辨的主体过程。

2 实体分辨中的数据分块算法

朴素的实体分辨方法需要对数据库中的记录进行两两比较,其中绝大部分记录比较是在非匹配记录对之间进行,而实体分辨的最终目的是发现匹配记录对,当需要处理的数据规模较大时,将造成大量计算能力的浪费。数据分块的目的就是在尽量不影响实体分辨准确性和完整性的同时尽量减少参与比较的记录数目。数据分块算法通常依据一定的分块主键(或排序主键)对参与实体分辨的记录进行划分,将每条记录划分到一个或多个数据块中(或对数据库进行排序),使得可能匹配的记录尽量出现在同一数据块中,在记录比较时只对同一数据块(或窗口)中的记录进行两两比较,从而在减少候选记录对数目的同时尽量保留可能匹配的记录对[13]。按照记录聚集方式的不同,实体分辨中的记录分块算法可以分为标准分块算法、基于近邻排序的分块算法和基于 q-gram 的分块算法 3 类。

标准分块算法为每条记录产生一个分块键值,将具有相同分块键值的记录划分到同一数据块中,候选记录对由同一数据块中的记录两两组合产生,如果某一键值只有一条记录与之对应则去除该键值和与之对应的记录[5]。标准分块算法只能将一条记录划分到一个数据块中,如果分块主键定义不合理,两条匹配记录可能会划分到两个不同的数据块中,从而遗漏匹配的记录对。文献[14]通过定义不同的分块主键,分别按照不同的分块主键对记录进行多次划分,使得某一分块主键下被划分到两个数据块中的匹配记录在其他分块主键下可能被划分到同一分块中,从而增加匹配记录对被检测到的机会,但这样也增加了需要比较的记录对的总体数目。

基于近邻排序的分块算法依据排序主键对所有记录进行排序,使得相似的记录能够彼此靠近,然后利用滑动窗口在排序后的记录上移动,每滑动一步,对窗口内的记录进行两两比较[11]。较早的近邻排序算法使用固定大小的窗口,窗口大小设置不合理会导致匹配记录丢失或增加候选记录对的数量。文献[15]提出可变窗口的 IA-SNM 和 AA-SNM 算法,IA-SNM 算法递增地增加窗口大小,直到窗口中第一条记录与最后一条记录之间的距离大于某一阈值;AA-SNM 算法在 IA-SNM 算法的基础上保留前一窗口的一条记录,从而可以把多个相邻的窗口进行合并。理论上讲,可变窗口可以获得更好的性能,然而 Uwe Draisbach 等人在实验验证时并没能得出 AA-SNM 算法和 IA-SNM 算法的性能比传统的近邻比较算法更好的结论[16]。文献[13]为排序主键建立倒排索引,窗口在索引值上滑动,使得具有相同索引值的记录都可以划分到相同的窗口中,提高了匹配记录对出现在同一窗口内的概率。

当分块主键(或排序主键)中存在错误或变体时,标准分块算法和基于近邻排序的分块算法可能无法将两条匹配的记录划分到同一个数据块中。文献[17]提出基于 q-gram 的分块算法,将分块键值分解为长度为 q(通常为 2 或 3)的子串,并依据得到的子串生成该分块键值的多种变体,再将具有相同键值变体的记录划分到同一个数据块中,这样一条记录可能被划分到多个不同的数据块中,从而提高了对错误和变体的容忍能力。基于类似的思想,文献[18]提出利用后缀数组生成分块键值变体的方法,并通过合并相似变体的方法减少产生候选记录对的数量。由于一条记录可能被划分到多个数据块中,因此增加了要比较的记录对的数目,文献[19]提出了迭代分块模型,即在实体分辨过程中将匹配记录合并为新的记录,再将合并后的记录与其他记录一起迭代地进行数据块划分,这样在某一个数据块中检测到的匹配信息会传递到其他数据块中,有助于发现更多的匹配记录并减少记录比较次数。

分块主键的定义是数据分块算法的关键,直接影响到候选记录对的数量和实体分辨结果的完整性[13]。文献[20]指出定义分块主键时通常要考虑属性值的缺失情况、属性值的概率分布和数据块大小等问题。文献[21]提出利用机器学习定义分块主键的方法,选择具有最优覆盖率和准确性的分块主键,然而此类方法需要的训练数据在实际应用中难以得到满足。目前大部分数据分块算法中的分块主键仍然是由领域专家进行定义。

3 实体分辨中的记录比较算法

记录比较是实体分辨的核心步骤,主要任务是利用各种相似度算法计算两条记录各个属性的相似度,得到该记录对的比较向量 $V=\langle s_1, s_2, \cdots, s_n \rangle$,其中 s_i 为各个属性的相似度,$i=1,2,\cdots,n$。这些比较向量将作为决策模型判断两条记录是否描述同一个实体的依据。因此,记录比较算法得到的相似度向量的准确性一定程度上决定着实体分辨结果的准确性。现有的属性相似度算法按照字符串划分粒度的不同,通常可以分为基于字符的相似度算法、基于 token 的相似度算法和基于读音的相似度算法 3 类[10]。

基于字符的相似度算法以单个字符为比较单位,度量两个字符串在字符级别上的差异。如 Levenshtein 距离[22]利用两个字符串相互转换所需字符的增、删、改操作的数目计算两个字符串之间的距离。Smith-Waterman 距离[1]在 Levenshtein 距离基础上引入了字符交换操作。LCS 算法[10]通过迭代地查找和移除两个字符串的最大公共子串计算两个字符串

的相似度，可以处理字符串中的单词位置变换产生的错误。基于 q-gram 的相似度算法[23]利用滑动窗口分别将两个字符串划分为长度为 q 的子串，然后利用两个字符串对应的子串集合的 Jaccard 系数计算相似度。Jaro-Winkler 距离[24]结合了 Levenshtein 和 q-gram 算法，利用两个字符串的公共子串和字符顺序计算相似度。基于字符的相似度算法可以很好地处理字符拼写错误，但对单词交换、缩写等问题不能很好捕捉。

基于 token 的相似度算法通过分词技术将字符串分解为多个 token，以 token 为单位计算两个字符串的相似程度，并且可以通过不同的 token 比较方法处理单词缩写、拼写错误等问题。Atomic Strings 算法[10]以标点符号为边界将字符串分解为多个 token，如果两个 token 相同或其中一个为另外一个的前缀，则认为两个 token 匹配，进而计算整个字符串的相似度；WHIRL 算法[10]在对字符串进行分词后，利用信息检索领域中的 TF-IDF 和余弦相似度计算两个字符串的相似程度，但 WHIRL 算法不能处理拼写错误。Soft TF-IDF 算法[25]和 q-gram TF-IDF 算法[26]对 WHIRL 进行了改进：Soft TF-IDF 算法在计算余弦相似度时考虑了“相似”的 token，而 q-gram TF-IDF 算法则用 q-gram 替代单词作为 token，解决了单词拼写、插入、删除等问题。

在信息检索等领域，可能需要计算字符串在读音上的相似程度，基于读音的相似度算法根据单个或多个字符的读音对字符串进行编码，利用得到的编码比较两个字符串读音上的相似程度。如 Soundex 算法根据英文字符的美式发音，将任何一个单词映射为由 1 个字符和 3 个数字组成的编码，使得读音接近的单词能够得到相同的编码[1]；NYSIIS 算法不使用数字编码，而是使用发音相同的字母对原始字符串中的字符进行替换，替换后的单词依然可以作为单词进行处理[10]；ONCA 算法则结合了前两者，先对字符串应用 NYSIIS 进行替换，再利用 Soundex 进行编码，提高了 Soundex 编码的准确性[27]。目前大多数基于读音的相似度算法都是针对英文发音，而对于中文字符读音相似度的计算通常是将中文字符转换为对应的拼音，再利用已有的方法进行编码[28]。

大量相似度算法的提出为记录相似度的计算提供了更多的选择，但同时也带来了问题。文献[23]和文献[24]通过对多种相似度算法的实验比较，指出没有一种相似度算法能够适用于所有数据。基于此，文献[29]对相似度算法选择问题进行了形式化，证明了该问题是 NP 难问题，并提出基于训练数据选择最优相似度算法和最优匹配阈值的启发式算法 SiFi-Greedy，SiFi-Gradient 和 SiFi-Hill，然而在实际应用中仍然难以获得合适的训练数据。

4 实体分辨中的决策模型

通过记录比较得到记录对的比较向量后，匹配决策步骤依据各种决策模型对记录的匹配状态进行判断。根据记录关联方式的不同，现有的决策模型可分为以下 3 类：两两比较模型、聚类模型和基于关系的决策模型。

两两比较模型将实体分辨看作分类问题，依据记录对的比较向量将记录对划分到匹配、不匹配和可能匹配的类别中，然后通过传递闭包获得描述同一现实实体的所有记录。现有的两两比较决策模型主要包括概率决策模型和基于规则的决策模型。概率决策模型[5]在具有训练数据和属性间条件独立的假设下，计算条件概率的比值 $R=\frac{P(\gamma\in\Gamma|r\in M)}{P(\gamma\in\Gamma|r\in U)}$，其中 γ 和 Γ 分别为比较向量和比较空间，r 为当前记录对，M 和 U 分别为匹配和不匹配记录对集，通过将 R 与上下两个阈值进行比较来判断记录对 r 的匹配状态。由于实际应用中常常缺少计算条件概率所需的训练数据，文献[30]利用期望最大算法对条件概率进行估计。文献[31]在概率决策模型的基础上考虑了错判匹配状态的不同代价，提出了基于最小代价的概率决策模型。文献[32]提出基于规则的决策模型，匹配规则由一组布尔表达式及其对应的匹配状态组成，形式为 $P=(term_{1,1}\vee term_{1,2}\vee\cdots)\wedge\cdots\wedge(term_{n,1}\vee term_{n,2}\vee\cdots)\Rightarrow C$，其中 $term_{i,j}$ 是对属性 i 的相似度的一个判断（如 $s_1>0.8$）。如果某个比较向量中的属性相似度满足布尔表达式 P，那么该记录对的匹配状态就为该表达式对应的匹配状态 C。匹配规则通常由领域专家制定，一般难以保证规则集合的完备性。文献[33]为了降低规则过少对实体分辨准确性的影响，将决策规则描述为匹配依赖，并给出了匹配依赖的推理规则，可以根据少量已知规则通过推理获得蕴含的规则集合，从而提高决策准确性。为了减少领域专家的参与程度，文献[34-37]分别提出了利用 CART 决策树、ID3 决策树、支持向量机和主动学习技术从训练数据中学习决策规则的方法。由于决策规则会随着人们对数据、模式和应用的认识程度的加深而发生改变，文献[38]对规则的演化进行了形式化，提出了规则单调和上下文无关两个约束，指出满足这两个约束的规则可以使用增量方式进行处理，在当前决策结果的基础上获得新规则的决策结果，从而减少计算复杂度。

聚类模型将实体分辨过程看作聚类问题，每一个类别对应一个现实实体。文献[39,40]利用多种相似度测度和优先队列对数据库中的记录进行聚类，为每一个实体选出一个代表性的记录保存在优先队列中，将每条记录与优先队列中的代表记录进行比较，依据不同的相似度值将记录划分到一个已有的实体中或建立一个新的实体。文献[41]利用聚类方法对两两比较模型的结果中可能匹配的记录对进行后续处理，在匹配记录对的传递闭包的基础上建立记录之间的匹配关系图，然后通过基于阈值的子图划分将描述同一实体的记录对聚集到一起，解决因传递闭包带来的冲突和误判。该方法的不足之处是记录两两比较的结果决定了记录之间匹配关系图的结构，无法检测比较结果中没有的匹配记录。文献[42]为了解决文献[41]中的问题，在整个比较空间上构建记录匹配关系图，利用紧密集和稀疏近邻的概念对记录对进行聚类。

基于关系的决策模型利用记录间的关联关系构建记录之间的关系图，利用图论知识计算记录间的连接强度，进而判断记录对是否匹配。基于关系的决策模型主要涉及两个问题：1）如何构建关系图；2）如何计算关系图中连接的概率或权重。文献[43]构建了数据库中不同实体间的关联关系图，每个节点代表数据库中的一个实体，每条边表示两个实体间存在关联关系，利用两个节点之间的路径长度估计两个节点的关联程度。文献[44]构建了数据库中的依赖关系图，每个节点表示一对实体间的相似度，节点间的边表示两个实体相似度之间存在依赖关系，利用相似度间的依赖关系迭代地计算两条记录的相似度。文献[45]构建了记录间的关联关系超图，以

每条记录作为节点，用超边连接多个存在关联关系的节点，每条边的权重由属性相似度和关联相似度的加权获得。文献[43-45]的实验结果表明，基于关系的决策模型的准确性整体上优于基于规则和基于聚类的决策模型，但同时也大大增加了计算复杂度。文献[46]为提高此类决策模型的效率，提出划分独立关系子图的方法，切断连接强度小于一定阈值的边，将整个关系图划分为多个关系子图，独立地在每个子图中判断记录的匹配状态，最后合并所有子图的决策结果。

结束语 随着信息技术的发展，越来越多的数据正在不断产生，由于数据质量问题普遍存在，同一现实实体往往具有多种数据描述与之对应，分辨哪些数据描述哪个现实实体对于数据的有效、正确利用有着重要的理论和应用价值。本文介绍了实体分辨的发展历程、面临的挑战和步骤组成，对每个步骤中涉及的具体方法进行了介绍和阐述。基于以上讨论，本文认为对于实体分辨的研究还存在以下问题有待进一步解决。

(1)虽然目前已经提出了大量的记录比较算法，但没有哪一种算法适应所有的数据情况，如何从大量算法中进行选择是一个值得研究的问题。而且，现有的相似度算法大多是针对英文字符串提出的，关于中文字符串相似度计算的研究成果还不多见。

(2)现有的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着多种依赖关系，如函数依赖、多值依赖等，在记录比较中这些依赖关系反映为属性相似度之间的依赖关系，因此可以考虑利用属性之间的依赖关系提高相似度计算的准确性。

(3)很多实体分辨方法需要具有明确匹配状态的训练数据的支持，但实际应用中通常无法或很难获得数据的真实匹配状态，因此如何针对不同的实体分辨方法人工生成合适的训练数据是目前需要解决的一个问题。

(4)描述同一实体的两条记录通常存在一致性问题(如姓名缩写的存在可能会使一个邮箱地址对应两个“不同”的姓名)，数据一致性可以作为启发式信息应用到实体分辨中。关于数据一致性检测的研究已经取得了丰富的研究成果[47]，但关于一致性在实体分辨中的研究还不多见。

参 考 文 献

[1] Christen P. Data Matching[M]. New York, USA: Springer, 2012

[2] 王宏志，樊文飞. 复杂数据上的实体识别技术研究[J]. 计算机学报，2011，34(10)：1843-1852

[3] Dunn H L. Record Linkage [J]. American Journal of Public Health and the Nations Health, 1946, 36(12): 1412-1416

[4] Newcombe H B, Kennedy J M, Axford S J, et al. Automatic linkage of vital records[J]. Science, 1959, 130(3381): 954-959

[5] Fellegi I P, Sunter A B. A theory for record linkage[J]. Journal of the American Statistical Association, 1969, 64(328): 1183-1210

[6] Winkler W E, Thibaudeau Y. An application of the Fellegi-Sunter model of record linkage to the 1990 US decennial census[R]. US Bureau of the Census, 1991: 1-22

[7] Hernández M A, Stolfo S J. The merge/purge problem for large databases[J]. ACM SIGMOD Record, ACM, 1995, 24(2): 127-138

[8] Monge A E. Matching algorithms within a duplicate detection system[J]. IEEE Data Engineering Bulletin, 2000, 23(4): 14-20

[9] Monge A, Elkan C. The field-matching problem: Algorithm and applications[C]// Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1996

[10] Elmagarmid A K, Ipeirotis P G, Verykios V S. Duplicate record detection: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(1): 1-16

[11] Winkler W E. Overview of record linkage and current research directions[R]. Bureau of the Census, 2006

[12] Batini C, Scannapieca M. Data quality: concepts, methodologies and techniques[M]. Springer, 2006

[13] 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

[14] Winkler W E, Yancey W E, Porter E H. Fast record linkage of very large files in support of decennial and administrative records projects[C]// Proceedings of the Section on Survey Research Methods. American Statistical Association, 2010

[15] Yan S, Lee D, Kan M Y, et al. Adaptive sorted neighborhood methods for efficient record linkage[C]// Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries. ACM, 2007: 185-194

[16] Draisbach U, Naumann F, Szott S, et al. Adaptive windows for duplicate detection[C]// IEEE 28th International Conference on Data Engineering (ICDE). IEEE, 2012: 1073-1083

[17] Baxter R, Christen P, Churches T. A comparison of fast blocking methods for record linkage[C]// ACM SIGKDD. 2003, 3: 25-27

[18] De Vries T, Ke H, Chawla S, et al. Robust record linkage blocking using suffix arrays and Bloom filters[J]. ACM Transactions on Knowledge Discovery from Data, 2011, 5(2): 9

[19] Whang S E, Menestrina D, Koutrika G, et al. Entity resolution with iterative blocking[C]// Proceedings of the 35th SIGMOD international conference on Management of data. ACM, 2009: 219-232

[20] Christen P, Goiser K. Quality and complexity measures for data linkage and deduplication[M]// Quality Measures in Data Mining. Springer Berlin Heidelberg, 2007: 127-151

[21] Michelson M, Knoblock C A. Learning blocking schemes for record linkage[J]. Proceedings of the National Conference on Artificial Intelligence, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006, 21(1): 440

[22] Koudas N, Sarawagi S, Srivastava D. Record linkage: similarity measures and algorithms[C]// Proceedings of the 2006 ACM SIGMOD international conference on Management of data. ACM, 2006: 802-803

[23] Cohen W W, Ravikumar P, Fienberg S E. A comparison of string distance metrics for name-matching tasks[C]// Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03). 2003: 47

[24] Snae C. A comparison and analysis of name matching algorithms [J]. International Journal of Applied Science, Engineering and Technology, 2007, 4(1): 252-257

(下转第 20 页)

over-relaxation algorithm[J]. Mathematical Programming Computation,2012,4(4):333-361

[32] Zhang De-bing, Hu Yao, Ye Jie-ping, et al. Matrix completion by truncated nuclear norm regularization[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012:2192-2199

[33] Nie Fei-ping, Huang H, Ding C. Low-rank matrix recovery via efficient schatten p-norm minimization [C] // Proceedings of AAAI Conference on Artificial Intelligence. 2012:655-661

[34] Lee K, Bresler Y. ADMiRA: Atomic decomposition for minimum rank approximation[J]. IEEE Transactions on Information Theory,2010,56(9):4402-4416

[35] Cai T T, Zhou Wen-xin. Matrix completion via max-norm constrained optimization [EB/OL]. http:// arxiv. org/pdf/1303.0341v1. pdf,2013

[36] Recht B, Christopher Re C. Parallel stochastic gradient algorithms for large-scale matrix completion [EB/OL]. http://pages. cs. wisc. edu/~brecht/papers/11. Rec. Re. IPGM. pdf,2013

[37] 史加荣,焦李成,尚凡华. 不完全非负矩阵分解的加速算法[J]. 电子学报,2011,39(2):291-295

[38] Xu Yang-yang, Yin Wo-tao, Wen Zai-wen, et al. An alternating direction algorithm for matrix completion with nonnegative factors [J]. Frontiers of Mathematics in China,2012,7(2):365-384

[39] 尚凡华. 基于低秩结构学习数据表示[D]. 西安:西安电子科技大学,2012

[40] Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices[J]. Journal of Machine Learning Research,2010,11:2287-2322

[41] Mohan K, Fazel M. Iterative reweighted algorithms for matrix rank minimization [J]. Journal of Machine Learning Research, 2012,13:3441-3473

[42] Xin Yu, Tommi Jaakkola T. Primal-Dual methods for sparse constrained matrix completion[J]. Journal of Machine Learning Research-Proceedings Track,2012,22:1323-1331

[43] Liu Ji, Musialski P, Wonka P, et al. Tensor completion for estimating missing values in visual data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(1):208-220

[44] Signoretto M, Van de Plas, R, De Moor B, et al. Tensor versus matrix completion:a comparison with application to spectral data[J]. IEEE Signal Processing Letters,2011,18(7):403-406

[45] 史加荣,焦李成,尚凡华. 张量补全算法及其在人脸识别中的应用[J]. 模式识别与人工智能,2011,24(2):255-261

[46] Julià C, Sappa A D, Lumbreras F, et al. Rank estimation in missing data matrix problems[J]. Journal of Mathematical Imaging and Vision,2011,39:140-160

(上接第 12 页)

[25] Bilenko M, Mooney R, Cohen W, et al. Adaptive name matching in information integration[J]. IEEE Intelligent Systems, 2003, 18(5):16-23

[26] Gravano L, Ipeirotis P G, Koudas N, et al. Text joins in an RDBMS for web data integration[C]// Proceedings of the 12th international conference on World Wide Web. ACM, 2003:90-101

[27] Gill L. OX-LINK: The Oxford Medical Record Linkage System [C]// Proc. Int'l Record Linkage Workshop and Exposition. 1997:15-33

[28] 刁兴春,谭明超,曹建军. 一种融合多种编辑距离的字符串相似度计算方法[J]. 计算机应用研究,2010,27(12):4523-4525

[29] Wang J, Li G, Yu J X, et al. Entity matching: how similar is similar[J]. Proceedings of the VLDB Endowment, 2011, 4(10):622-633

[30] Herzog T N, Scheuren F J, Winkler W E. Data quality and record linkage techniques[M]. Springer, 2007

[31] Verykios V S, Moustakides G V, Elfeky M G. A Bayesian decision model for cost optimal record matching [J]. The VLDB Journal, 2003, 12(1):28-40

[32] Naumann F, Herschel M. An introduction to duplicate detection [J]. Synthesis Lectures on Data Management, 2010, 2(1):1-87

[33] Fan W, Jia X, Li J, et al. Reasoning about record matching rules [J]. Proceedings of the VLDB Endowment, 2009, 2(1):407-418

[34] Cochinwala M, Kurien V, Lalk G, et al. Efficient data reconciliation[J]. Information Sciences, 2001, 137(1):1-15

[35] Christen P. Automatic record linkage using seeded nearest neighbour and support vector machine classification[C]// Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008:151-159

[36] Christen P. Automatic training example selection for scalable unsupervised record linkage[M]// Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2008:511-518

[37] Arasu A, Götz M, Kaushik R. On active learning of record matching packages[C]// Proceedings of the 2010 International Conference on Management of Data. ACM, 2010:783-794

[38] Whang S E, Garcia-Molina H. Entity resolution with evolving rules[J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 1326-1337

[39] Monge A E. Matching algorithms within a duplicate detection system[J]. IEEE Data Engineering Bulletin, 2000, 23(4):14-20

[40] Hassanzadeh O, Miller R J. Creating probabilistic databases from duplicated data[J]. The VLDB Journal—The International Journal on Very Large Data Bases, 2009, 18(5):1141-1166

[41] Hernández M A, Stolfo S J. The merge/purge problem for large databases[C]// ACM SIGMOD Record. ACM, 1995, 24(2):127-138

[42] Chaudhuri S, Ganti V, Motwani R. Robust identification of fuzzy duplicates[C]// Proceedings of 21st International Conference on Data Engineering 2005. IEEE, 2005:865-876

[43] Kalashnikov D V, Mehrotra S. Domain-independent data cleaning via analysis of entity-relationship graph[J]. ACM Transactions on Database Systems (TODS), 2006, 31(2):716-767

[44] 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. ACM, 2005:85-96

[45] Bhattacharya I, Getoor L. Collective entity resolution in relational data[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1):5

[46] Rastogi V, Dalvi N, Garofalakis M. Large-scale collective entity matching[J]. Proceedings of the VLDB Endowment, 2011, 4 (4):208-218

[47] Fan W. Dependencies revisited for improving data quality[C]// Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2008:159-170