

文章编号: 1000-5641(2018)05-0041-15

异构网络中实体匹配算法综述

李 娜, 金冈增, 周晓旭, 郑建兵, 高 明

(华东师范大学 数据科学与工程学院, 上海 200062)

摘要: 互联网、物联网和云计算技术的不断融合,使得各行各业信息化程度越来越高,但同时也带来了数据碎片化的问题。数据碎片化的海量性、异构性、隐私性、相依性和低质性等特征,导致了数据可用性较差,利用这些数据难以挖掘出准确而完整的信息。为了更有效地利用数据,实体匹配、融合和消歧变得尤为重要。主要对异构网络中实体匹配算法进行了综述,对实体相似度度量和数据预处理技术进行了梳理;特别针对海量数据,概述了可扩展实体匹配方法的研究进展,综述了运用监督学习和非监督学习两类技术的实体匹配算法。

关键词: 数据融合; 实体匹配; 记录链接; 实体解析

中图分类号: TP391 文献标志码: A DOI: 10.3969/j.issn.1000-5641.2018.05.004

A survey of entity matching algorithms in heterogeneous networks

LI Na, JIN Gang-zeng, ZHOU Xiao-xu, ZHENG Jian-bing, GAO Ming

(School of Data Science and Engineering, East China Normal University, Shanghai 200062, China)

Abstract: The continuous integration of Internet, Internet of Things, and cloud computing technologies has been improving digitization across different industries, but it has also introduced increased data fragmentation. Data fragmentation is characterized by mass, heterogeneity, privacy, dependence, and low quality, resulting in poor data availability. As a result, it is often difficult to obtain accurate and complete information for many analytical tasks. To make effective use of data, entity matching, fusion, and disambiguation are of particular significance. In this paper, we summarize data preprocessing, similarity measurements, and entity matching algorithms of heterogeneous networks. In addition, particularly for large datasets, we investigate scalable entity matching algorithms. Existing entity matching algorithms can be categorized into two groups: supervised and unsupervised learning-based algorithms. We conclude the study with research progress on entity matching and topics for future research.

Keywords: data fusion; entity matching; record linkage; entity resolution

收稿日期: 2018-07-04

基金项目: 国家重点研发计划项目(2016YFB1000905); 国家自然科学基金广东省联合重点项目(U1401256); 国家自然科学基金(61672234, 61502236, 61472321); 上海市科技兴农推广项目(T20170303)

第一作者: 李 娜, 女, 硕士研究生, 研究方向为数据挖掘. E-mail: nali0606@foxmail.com.

通信作者: 郑建兵, 男, 博士, 研究方向为信息处理技术. E-mail: zhengjb@js.chinamobile.com.

0 引言

信息技术的迅猛发展产生了海量的数据,无论是社交网络、医疗领域、教育领域,还是社会治理^[1],人们对数据挖掘的重要性和必要性越发明确。然而,在大数据时代,网络平台上的数据表现出零散、低质、异构等碎片化特征,特别是用户产生内容(User Generated Content, UGC)的碎片化特征更加显著。一方面,企业内部存在的“数据孤岛”和“数据烟囱”问题^[2],即不同部门之间的数据相互独立、部门之间数据不共享;另一方面,不同企业间的数据存在天然的壁垒,相互间的信息存在低关联性。尽管数据碎片化问题比较严重,但是这些数据中却蕴含着巨大的商业价值。例如,综合运用微博平台上的社交数据和淘宝等电子商务平台上的消费数据,可以推测用户的购买能力、兴趣偏好和好友圈子等,这些信息有助于更为精准的商品推荐。因此,匹配不同数据源的实体是融合互联网数据以便更好地进行数据挖掘的前提,只有将碎片化数据进行拼接,才可以帮助相关企业更好地理解用户的行为、兴趣、爱好等信息。此外,通过实体匹配不仅可以减小数据的冗余,而且拼接碎片化数据还可以提高数据质量。

实体匹配具有广泛的应用场景,可应用于数据管理、信息检索、机器学习、数据挖掘等多个关键领域。其主要应用场景如下。

- 推荐系统:通过实体匹配,可以更好地分析用户的偏好和兴趣,以实现更为准确的个性化推荐。
- 智慧医疗:通过将碎片化的病人信息进行匹配,可以获取病人更全面、更准确的得病史、身体情况以及家族病史等。借助这些较完整的病人信息,医生可以更精准地定位病源,从而实现更为精准的治疗。
- 用户画像:通过用户实体匹配,可以将某个人在多个数据源上的信息进行融合,以便更全面地了解其整体情况,对用户精准画像、多维度画像提供依据。

随着数据规模的迅速扩大,实体匹配面临的挑战也越来越多,数据固有的特点为实体匹配问题增加了难度,主要挑战如下。

- 数据海量性:不断扩大的用户规模所产生的数据越来越多。例如,微信的用户量已超过10亿^①;Facebook的月活跃用户数破20亿^②,占近乎世界人口的1/4。同时,数据的类型不断增多,相对于传统的数字型数据和文本数据,视频、图片等非结构化数据不断增加。
- 数据异构性:数据具有异构化的特征。数据的异构性主要体现在两个方面:第一,数据源的异构,即数据来源于不同的平台,例如,用户社交数据来源于不同的社交网络(微博、微信等);第二,数据分布的异构,即UGC中包含结构化、半结构化和非结构化数据,如文字、图片、音频和视频数据。
- 数据隐私性:网络数据涉及信息安全问题,随着信息技术的飞速发展,安全问题越来越受到个人用户和企业的重视。隐私信息需要进行加密和脱敏处理,如身份证件信息、真实

① http://www.xinhuanet.com/politics/2018-03/05/c_1122488991.htm

② http://www.xinhuanet.com/info/2017-07/06/c_136421691.htm

姓名、真实家庭地址等。加密和脱敏处理会使得数据质量降低、数据完整性缺失,这就大大增加了实体匹配的难度。

- 数据相依性: 实体之间存在相依性关系, 传统数据独立性的假设不再成立。数据的相依性虽然可以更好地相互佐证改进数据的质量, 但是这种相依关系会增加实体匹配的难度。
- 数据低质性: 从单个数据源看, 数据存在信息缺失、不准确或者信息过时等问题, 甚至存在数据真实性问题; 从多个数据源看, 数据之间存在冲突、不一致和难关联等问题。在低质数据下的实体匹配是一项非常具有挑战性的任务。

实体匹配作为提高数据质量的有效途径, 可以将多数据源的实体结合起来。随着时代的发展, 实体匹配的任务不再局限于算法匹配结果的正确性。实体匹配作为数据集成的关键任务之一^[3], 面临着新的任务需求。首先, 提高准确性仍然是实体匹配算法优先考虑的问题。如何在低质数据、海量数据、异构数据情况下设计高效的算法, 如何挖掘数据的内在联系以获得最大的算法准确性, 是研究的一个重要任务。其次, 需要提高算法的可扩展性。传统的实体匹配中, 候选匹配集的个数是 $|S| \times |T|$, 其中, $|\bullet|$ 表示•数据集中的实体数。在数据集小的情况下, 不需要考虑候选集的规模; 但是随着数据量的增大, 候选集的规模越来越大, 带来的资源消耗越来越多、算法性能越来越差。因此, 如何解决大规模数据下的性能瓶颈问题、如何高效地进行数据匹配是研究中心必须考虑的。最后, 如何降低人力成本也是需要考虑的问题。降低人力成本主要指降低标注数据成本。现有的大多数算法都是基于监督学习, 都需要有标注的数据。对于实际应用来说, 数据几乎没有标注或者标注数据很少。现有的大量的研究工作需要人工标注数据, 非常浪费人力物力, 带来的成本消耗非常大。

本文结构如下: 第1节给出实体匹配的问题定义; 第2节论述实体匹配的基础知识, 主要总结实体匹配中数据预处理的主流方法, 并概括经典的距离度量方式, 特别针对海量数据, 综述可扩展实体匹配方法的研究进展; 第3节对现有的实体匹配算法进行分类, 主要分为基于监督学习、非监督学习两类; 第4节对算法进行总结, 指出各类别算法的优点及不足; 第5节对全文进行总结并对未来的研究热点做出展望。

1 问题定义

实体匹配(Entity Matching, EM), 又称为记录链接(Record Linkage)^[4]、实体解析(Entity Resolution)^[5-6]、对象识别(Object Identification)^[7], 在单个数据源中又叫做重复检测(Duplicate Detection)^[8]。1946年, Dunn等人^[9]定义了实体匹配的任务: 是在多个数据源中找到指向同一个实体的记录。实体匹配主要包括数据源内的实体匹配和多数据源之间的实体匹配。同一数据源内的匹配主要针对数据源内有重复实体记录的情况; 数据源之间的匹配主要在多数据源之间匹配实体, 这个层级的匹配可以更好地关联不同数据源之间的实体, 为更全面地挖掘用户行为和用户特征打下基础。以用户实体匹配为例, 数据源内部和数据源之间的实体匹配示意图如图1和图2所示, 图中虚线连接的实体表示的实际是同一实体。

下面以2个数据源为例给出实体匹配的定义: 假设 R 和 S 分别是两数据源中相同类型的实体集合, 其中 $R = \{r_1, r_2, \dots, r_m\}$, $S = \{s_1, s_2, \dots, s_n\}$, 判断 r_i 和 s_j 是否匹配便是实体匹配问题。假设实体对匹配的函数用 $f(r_i, s_j)$ 表示, 一般规定当 $f(r_i, s_j)$ 大于等于阈值 τ 意味

着实体 r_i 和 s_j 是匹配的, 反之则不匹配. 形式化表示为

$$p = \begin{cases} \text{不匹配}, & f(r_i, s_j) < \tau, \\ \text{匹配}, & f(r_i, s_j) \geq \tau, \end{cases} \quad (1)$$

其中, r_i 与 s_j 共有若干属性, p 表示实体 r_i 和 s_j 的匹配结果. 所以实体匹配的任务便是如何定义匹配函数 $f(r_i, s_j)$ 以及阈值 τ . 实际生产环境中, 很难出现两个实体的特征值完全匹配的情况, 所以如何在异构的、低质量的数据中进行实体匹配是实体匹配面临的一个重大挑战.

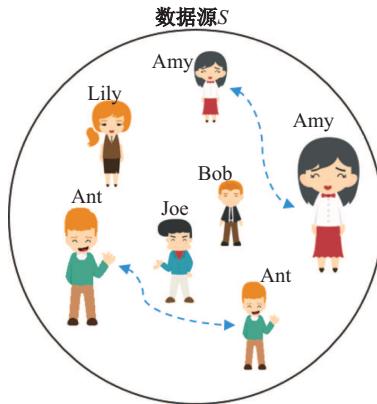


图 1 单数据源实体匹配

Fig. 1 Entity matching within a single data source

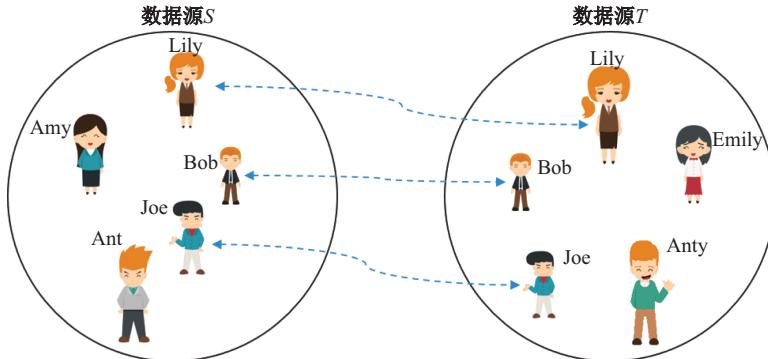


图 2 多数据源实体匹配

Fig. 2 Entity matching between multiple data sources

2 实体匹配基础

2.1 数据预处理方法

在进行实体匹配之前, 首先需要进行数据的预处理, 即对缺失数据和噪声数据的预处理. 表 1 简单总结了在实体匹配问题中针对缺失数据和噪声数据的处理方法以及方法的优缺点.

2.2 海量数据下的性能优化方法

随着数据规模的不断增大, 候选集规模显著增大, 导致算法效率降低. 识别匹配的实体, 首先需要构建候选集合, 若不对生成候选集合的算法进行优化, 那么候选集合会是不同数据

源中实体的笛卡儿积操作的结果。比如,对于两个实体数量分别为 M 、 N 的数据源,候选集合中元素的数量是 $M \times N$ 。面对海量的数据,基于笛卡儿积操作的候选对生成方法带来的代价是无法接受的,为了降低复杂度,减少候选集合元素的个数是一个重要的数据筛选过程。因此,如何减小候选集的规模是研究的一个重点,是实体匹配性能优化的关键点。目前解决此问题的主流技术是分块技术和索引技术。

表 1 数据预处理类型、方法及优缺点

Tab. 1 Types, methods, advantages, and disadvantages of data processing

数据类型	处理方法	优点	缺点
	含有缺失值的数据记录直接删除	简单方便	会去除很多有用的数据, 包括很多可能匹配的数据
	数值型元素按该列平均值补全, 字符型用null补全	数据质量较高, 应用范围最广	损失了一部分精度
缺失数据	聚类方法, 基于其他特性判断该样本的整体情况, 维持其整体情况不变并补全 比如: 一个人在新浪微博和人人网上的好友量都有top3, 且在开心网上的好友数据缺失, 那么使用平均值补全其数据是不合理的, 因为人在开心网上好友排名仍然会相对较高。 可以根据这个信息进行缺失值补全 ^[10]	数据质量最高, 数据内在特性得以保留	复杂度较高, 效率低
	聚类之后去除离群点	思想简单, 易操作	会去除很多有用的数据
噪声数据	根据特定规则, 基于统计, 将少数出现的数据变成多数 比如: Modiano不是一些特殊数据且出现多次, Medianoy只出现一次, 极有可能Medianoy是拼错的, 将Medianoy换成Modiano ^[11]	针对特定属性进行特别处理, 得到数据质量较高	不同的属性可能有不同的规则, 比如姓氏Lin出现多次, Ling出现一次, 但是不可能将Ling更改为Lin, 所以替换场景比较难确定

分块技术从候选匹配集中去除那些明显不匹配的实体对,保留相对匹配可能性较高的实体对,以此来减小候选匹配集的规模。算法流程如图 3 所示,具体为:①对来自不同数据源的实体使用相同的分块技术(如哈希函数等),分成若干实体块;②块内两两匹配生成候选集;③针对每一个候选对运行实体匹配算法。

Kong 等人^[12]使用基于字符串的局部敏感哈希(Locality-Sensitive Hashing, LSH)技术对实体进行分块,以降低数据的规模。其主要思想是通过使用 LSH,将相似的实体分到同一个桶中,如果来自不同数据源的 N 个实体不在 LSH 的同一个桶中,则这个包含 N 个实体的组合从候选集中剔除。文献 [13] 中使用基于三叉树的属性聚类(Attribute Clustering, AC)算法进行分块,AC 算法基于三元组的相似性得分将相似的实体分成一块。除此之外,文献 [14-18] 都是基于分块技术减小候选集的规模,并取得了很好的效果。

Zhang 等人^[19]也提出了减小匹配候选对个数的 3 种方案:基于内容的方法、基于结构的方法以及二者混合的方法。基于内容的方法是根据用户的昵称等一些文本数据,只有相似度较高的实体才生成实体对。基于结构的方法是在基于内容的实体对候选集基础上,根据网络结构扩大候选集。二者混合的方法便是基于内容和基于结构单独生成实体对候选集之后

取并集.

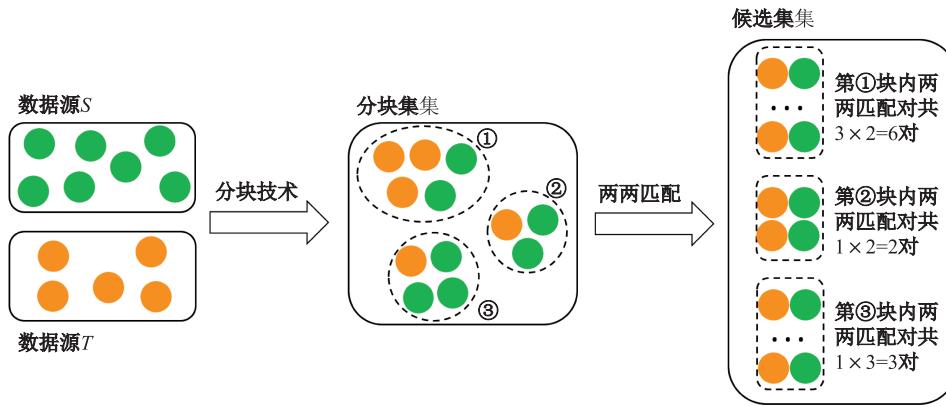


图 3 使用分块技术的候选集生成过程

Fig. 3 The process of generating pair-wise candidate sets using blocking technology

索引技术可分为分块索引、近邻排序索引、基于 Q-Gram 的索引、后缀阵列索引、冠层聚类、字符串映射索引等 6 大类. 这几类索引都在实际生产中有着广泛的应用, 可参见 Christen 在 2012 年发表的综述^[20].

2.3 相似度度量方式

相似度度量是实体匹配算法的基础. 早期的相似度度量主要基于字符串相似度. 随着算法研究的不断深入, 基于字符串相似度的度量不足以捕捉到实体之间的联系, 基于网络拓扑结构和基于网络向量的相似度度量方法则渐渐流行起来. 下面详细介绍几种相似度度量方法.

(1) 基于字符串的相似度

基于字符串相似度的实体匹配算法主要是利用实体的字符串信息, 如用户名、商品名称等计算候选对之间的相似度. 常用的相似度度量如下.

- 欧氏距离^[21]、余弦相似度^[22-23] 这两种相似度的度量都是基于向量空间的, 即需要先将字符串转化为向量空间. 比如, 两个字符串转化为向量空间之后的值为 $x = (x_1, x_2, \dots, x_n)$, $y = (y_1, y_2, \dots, y_n)$, 则欧氏距离和余弦相似度的计算分别是

$$E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2},$$

$$C(x, y) = \frac{x \cdot y}{|x||y|} = \frac{\sum_{i=1}^n (x_i \times y_i)}{\sqrt{\sum_{i=1}^n (x_i)^2} \times \sqrt{\sum_{i=1}^n (y_i)^2}}.$$

- Levenshtein 编辑距离^[24]、最长公共子串^[25]、N-gram 相似度^[26]、Jaro-Winkler 距离^[26-27] 这 4 种相似度的度量都是基于字符串文本计算的. Levenshtein 编辑距离用来计算从原串转换到目标串所需要的最少的插入、删除和替换的次数. 最长公共

子串是指两个字符串之间最长公共子串的长度. N -gram 是字符串长度为 N 的子串, N -gram 相似度是两个字符串 N -gram 相同的子串个数. Jaro-Winkler 距离是 Jaro 距离的扩展, Jaro 距离的定义为

$$d_j = \begin{cases} 0, & \text{若 } m = 0, \\ \frac{1}{3} \left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m} \right), & \text{其他,} \end{cases}$$

其中, $|s1|$ 和 $|s2|$ 是两字符串的长度, m 是匹配的字符数, t 是一个字符串转化为另一个字符串换位字符的数目.

基于 Jaro 距离, Jaro-Winkler 距离公式为

$$d_w = d_j + lp(1 - d_j),$$

其中, l 是两字符串公共前缀的字符个数, p 是常量比例因子, $p \leq 0.25$.

- Jaccard 相似度^[28]是基于集合进行度量的, 因此在使用 Jaccard 之前, 可以使用单词或者 N -grams 方法转换等构造字符集合. 假设原串产生的集合是 S , 目标串产生的集合是 T , 则 Jaccard 的计算公式为

$$J(S, T) = \frac{|S \cap T|}{|S \cup T|} = \frac{|S \cap T|}{|S| + |T| - |S \cap T|},$$

当 S 和 T 都为空时, $J(S, T) = 1$.

- Soundex 距离^[11,26]关注的是发音的相似度, 而上述几种相似度度量都是基于字符串文本的相似度. 首先, 该算法会将两个字符串分别通过一定的哈希算法转换成哈希值, 该值由 4 个字符构成. 进行字符串转化的哈希算法并非随机选取, 而是利用了该拉丁文字符串的读音近似值. 例如, “Michael”和“Mickel”具有相同的 Soundex 距离.

(2) 基于网络拓扑结构的相似度

随着实体匹配研究的发展, 基于简单的字符串的相似度算法趋于成熟, 基于网络拓扑结构的相似度研究不断增多^[11-29,30]. 基于网络拓扑结构的相似度是指利用实体之间网络的直接或间接关系, 挖掘出实体网络之间的潜在特征.

常用的基于网络拓扑结构的相似度度量算法有基于邻居节点的 Jaccard 度量^[31]、基于节点的出入度^[32-33]、基于元路径的度量^[11,34]、基于图的最短路径^[35]以及定义的其他度量^[29,36].

除基于字符串和基于网络拓扑结构的相似度之外, 还有一些工作是基于信息熵或者特定领域规则来进行数据相似性的衡量的, 比如, 对于不同地区的姓名, 按照各地区的姓名习惯来进行匹配和筛选等^[11,37].

通常情况下, 不会单一地考虑某种相似度, 而是将基于字符串的相似度和基于拓扑结构的相似度结合. 实体 i 和实体 j 的距离定义是

$$d_{ij} = \sum_{k=1}^K w_k d_{ijk},$$

公式中,

$$\sum_{k=1}^K w_k = 1$$

其中, 实体共有 K 个特征, 记作 Y_k , $k = 1, 2, \dots, K$, w_k 表示 Y_k 的权重, d_{ijk} 表示实体 i 和实体 j 在 Y_k 特征上的距离.

(3) 基于网络向量的相似度

基于网络向量的相似度是通过网络嵌入技术对节点进行向量表示, 然后通过向量求节点间的相似度. 网络嵌入旨在学习网络中顶点的潜在表示, 学习到的向量包含网络节点间的潜在信息. DeepWalk^[38]、Node2Vec^[39]和 LINE^[40]是典型的网络表示学习方法. DeepWalk 使用 SkipGram 的方法进行网络中的节点学习, 随机游走均匀地选取网络节点. Node2Vec 在 DeepWalk 的基础上, 改进了随机游走策略, 同时考虑了局部和宏观的信息. LINE 使用一阶近邻和二阶近邻学习网络向量. 这些网络表示方法也被逐渐应用到了实体匹配领域, 如 Zhou 等在文献 [41] 中借助随机游走的思想, 使用自然语言处理中基于负采样技术的 CBOW(Continuous Bag-of-Words) 模型^[42]来学习网络向量.

3 实体匹配算法

3.1 基于监督学习的实体匹配算法

基于监督学习的实体匹配算法的思想是: 首先对不同数据源的数据进行预处理操作, 接着将不同数据源的数据两两配对生成实体对集合, 对训练集中的实体对使用监督学习算法进行训练后得到算法模型; 然后利用训练好的模型对新的实体对进行分类, 得到匹配结果. 若 S 和 T 表示两个数据源, $S = \{s_1, s_2, \dots, s_m\}$, $T = \{t_1, t_2, \dots, t_n\}$, 若不使用分块/索引技术, 通过两两匹配构造候选集合 $P = \{(s_1, t_1), (s_1, t_2), \dots, (s_m, t_n)\}$, 共有 $m \times n$ 个候选对, 这个实体对规模带来的计算代价是巨大的. 因此, 在海量数据场景中, 一般在生成实体对候选集之前会进行数据分块/索引处理, 将所有数据使用同一个哈希函数进行映射等方法减小实体对的规模. 其具体流程如图 4 所示.

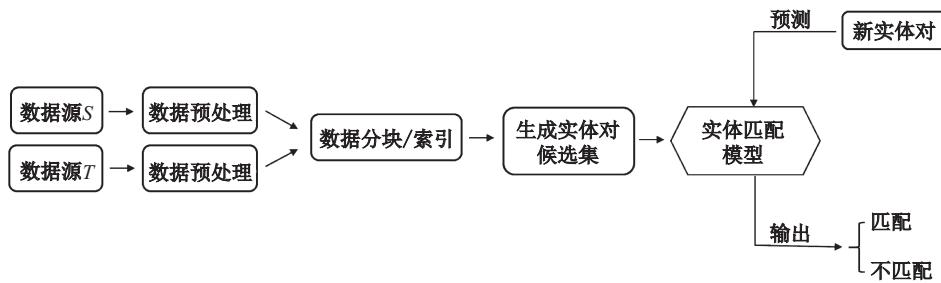


图 4 基于监督学习的匹配算法流程图

Fig. 4 Flowchart for supervised learning-based entity matching algorithms

支持向量机^[42]、Logistic 回归^[44-45]、AdaBoost^[13]、梯度提升树^[27,37]、贝叶斯网络^[46]等机器学习分类算法都可以用来训练最终的决策模型.

基于监督学习算法训练的模型能更好地拟合数据, 反映数据真实的特征. 但是基于监督学习算法的训练过程需要大量的数据, 训练模型有较高的复杂度, 尤其是一些集成学习的方法, 训练过程对硬件要求比较高.

3.2 基于非监督学习的实体匹配算法

3.2.1 基于规则的实体匹配算法

基于规则的实体匹配算法一般基于人为定义的规则来确定两个实体是否匹配, 这类算法思想比较简单, 也很直观。规则可分为精准匹配规则和近似匹配规则, 如“性别相同”是精准匹配规则, “姓名相似”是近似匹配规则。简单的规则举例如下: 若两个人姓名字符串编辑距离小于 2 且性别相同, 则判定这两个人属于真实的同一个人, 这便是基于简单规则的候选对结果的判定。

在文献 [47] 中, 作者定义了规则: $\forall o$, 若 o 满足条件 A , 则 o 指向 B 。比如“(name=‘wei wang’) \cap (‘kum’ \in coa) \Rightarrow e_1”表示若 name 是“wei wang”, 且 e_1 属性的其中一个取值是“kum”, 则将实体将“wei wang”指向实体 e_1 。Zhong 等人^[37]在算法中构造了事件规则, 若当前作者姓名是大写的, 且在当前论文中共同作者的姓名是缩写, 则将此定义为一个事件。Chin 等人^[48]定义了一套复杂的规则来判断作者的姓名是否是中文名。Zafarani 等人^[45]使用复杂的规则来生成相似的用户名。文献 [49-51] 也是基于规则的匹配算法。

基于规则的实体匹配算法较直观, 代码的复杂性较低。但是实体匹配规则的制定需要依赖于专家的先验知识, 制定规则需要消耗大量的人力和时间, 并且规则的合理性、有效性将直接决定算法的性能。除此之外, 相似度是多少才算比较相似也是一个很难定义的问题^[21]。一般而言, 纯粹基于规则的实体匹配算法已经无法满足实际需求, 简单规则的制定也被应用到其他类别的实体匹配算法, 如使用聚类算法进行实体匹配, 但是可以在某一特征上进行规则的制定, 以进行特征数据的量化。

3.2.2 基于聚类的实体匹配算法

基于聚类的算法也是解决实体匹配问题的一类常用算法, 使用这种算法不需要事先生成候选集。聚类算法是无监督学习的算法, 也就是说, 其可以解决没有样本标签的实体匹配问题, 所以更适用于大多数的实际情况, 是研究的一个热点。聚类是对大量未知标注的数据集, 按照数据内部存在的数据特征将数据集划分为多个不同的类别, 使类别内的数据比较相似, 类别之间的数据差异性比较大, 即算法将比较相似的实体聚集到一起, 将不同类别的实体分离。聚类的个数表示真实情况的实体数, 比如, 有 N 个用户名(可能存在一个真实用户有多个用户名的情况), 现在使用聚类算法将其划分为 K 个簇($K \leq N$), 则说明, 这 N 个用户名实际代表的是 K 个真实的用户。在聚类算法中, 每个簇至少包含一个实体对象, 且每个实体对象仅属于唯一的簇, 为了达到目标函数的最优化, 聚类算法是一个迭代的过程。

早期的基于聚类的实体匹配算法需要人工指定 K 值。典型的算法是 K -Means^[52], 其算法执行流程为: ①选择 K 个样本作为初始簇心; ②计算每个样本到簇心的距离, 并将样本划分到最小距离簇心所在的簇中; ③更新簇心; ④迭代执行②、③直到算法收敛。Monika 等人^[53]便利用 K -Means 来解决不同社交网络间用户链接的问题。因为大多数的 K 值在数据处理之前用户也是未知的, 并且同一实体匹配任务中, 每个实体对应的 K 值也不一样, 因此 K 值的设定具有随机性, 而这种随机性将对结果产生重要的影响。人工指定的 K 值很难是最优值。随着聚类技术的发展, 自动确定 K 值的算法被提出, 通过制定停止准则使其不断学习数据中的信息, 来决定最终的聚类个数, 这种方法更加流行并且取得了较好的效果。聚类算法大致可分为两类: 自上而下聚类算法和自下而上聚类算法。

(1) 自上而下聚类算法

在自上而下聚类算法中, 初始时刻所有的实体都属于同一个簇, 算法根据目标规则

不断将不属于同一类别的实体进行分裂, 直到最终算法收敛. Tang^[29]使用贝叶斯信息准则(Bayesian Information Criterion, BIC)自动确定 K 值. 其规则是: 最开始将全集当做一个聚簇, 聚簇结果记为 M_1 ; 接着在 M_1 中的每个聚簇中找到两个子聚簇, 聚簇结果记为 M_2 , 若发现 $BIC(M_2) > BIC(M_1)$, 则将此聚簇分裂, 否则不进行分裂. 将聚类结果不断迭代计算直到最终的聚簇数趋于稳定. 此算法实质上是 K -Means 的改进, 和 K -Means 不同的是采用的目标函数不是平方差而是针对实际意义的最大似然估计函数.

(2) 自下而上聚类算法

在自下而上聚类算法中, 初始时刻每一个实体都属于一个类, 然后根据目标规则寻找与其相似的实体并不断合并, 直到最终算法收敛. Cen 等人^[54] 使用基于自适应停止准则的层次聚类算法, 将停止准则构建成一个核岭回归问题, 通过 EM(Expectation-Maximum) 算法确定核岭回归模型的参数, 从而可以针对不断变化的聚类结果自适应停止准则, 这种方法比指定固定的停止阈值更有效. 除此之外, 文献 [23,55-56] 也使用了自下而上的聚类算法.

3.2.3 基于概率图的实体匹配算法

基于概率图的实体匹配算法主要是利用概率模型来实现实体的匹配. 早期的基于概率模型的算法将特征值拟合成某种概率分布, 分布的参数决定判定的结果, 因此其主要任务便是求解分布的参数. 求解最终匹配得分的经典模型是 Fellegi-Sunter 模型^[57], 该方法使用生成概率模型表示观测数据的联合概率分布, 通过最大似然函数学习模型的参数, 并利用最终学习到的参数确定实体匹配的最终得分. 常使用 EM 算法^[58] 学习概率模型的参数或将目标转化为对偶问题^[59]. 文献 [12,60] 将特征值的分布拟合成指数族分布, 利用 EM 算法学习指数族分布的参数, 最后使用 Fellegi-Sunter 模型确定最终的匹配得分. Gao 等人在文献 [60] 中提出的算法的主要思想是: 若合并两个实体后, 它们在某些人们公认的行为上表现更好, 则这两个实体更可能是一个整体, 其匹配的可能性更大. 其利用 EM 算法学习混合模型的参数以使似然函数取得最优值, 并利用参数计算出匹配得分. 除此之外, 文献 [54,62-63] 也使用了 EM 算法求解参数的思想, 文献 [10] 则将问题转化为对偶问题进行求解.

随着文本技术的发展, 基于自然语言处理的方法和思想也逐渐被迁移到实体匹配中, 比如主题模型、随机游走等算法可以很好地考虑到实体属性的内容以及实体之间的联系, 而且对人为因素依赖较小.

文献 [64] 中引入了条件随机场算法, 其考虑了不同实体匹配之间的复杂的相互作用关系. 表 2 所示的是一个参考文献数据库的例子.

表 2 参考文献数据库

Tab. 2 Reference database

记录	标题	作者	审稿单位
b1	Object Identification using CRFs	Linda Stewart	Proc. AAAI-05
b2	Object Identification using CRFs	Linda Stewart	Proc. 20th AAAI
b3	Learning Boolean Formulas	Bill Johnson	Proc. AAAI-05
b4	Learning of Boolean Expressions	William Johnson	Proc. 20th AAAI

在这个例子中, 因为 b1 的标题和作者与 b2 的标题和作者一致, 可以认定 b1 和 b2 是同一个论文实体, 即 $b1=b2$, 由此可以推导出 Proc.AAAI-05 和 Proc.20th AAAI 是同一个审稿单位实体(这个实体的判定是从 b1 和 b2 实体判定的结果中得到的信息); 随后利用这个结论去判定 b3 和 b4 是否是同一个实体. 为了利用这种传递的思想, 作者使用了条件随机场模型来实现, 这

种模型能够捕捉到实体匹配结果之间复杂的联系, 而后把这种联系运用到其他实体的匹配中。这个算法使得准确度在一定程度上有所提升, 其将预测实体之间的相互关系变为有效输入信息为之后的实体匹配提供决策支撑。文献[65]也是基于条件随机场的算法。

也有很多基于主题模型的实体匹配算法, 典型的主题模型有 LDA^[66]以及 LDA 的变体 Twitter-LDA^[67]。基于主题模型的实体匹配算法大多借助 LDA 的思想并根据自身任务需求进行建模。Yang 等人^[68]于 2015 年提出了基于概率主题模型的跨异构网络实体匹配算法, 此算法主要包括主题提取和实体匹配两阶段。因为异构网络可能出现文本重叠较少的现象(如: 一个数据源是中文, 另一个数据源是英文), Yang 等人利用“相似实体虽然表面上有不同的主题表示, 但是在主题隐空间中却比较相似, 即隐空间中主题相似的实体最可能是匹配实体”的思想对 LDA 概率图模型进行了改进。经典的 LDA 模型, 因其本质是生成概率模型, 因此若基于 LDA 进行实体匹配, 必须要求文本之间是有交叉信息的, 面对几乎无重复文本的情况, Yang 等人基于主题模型的方法很好地解决了这个问题。文献[55]中, 作者基于 PLSA 和 LDA 主题模型, 在概率图中加入“人”作为隐变量, 利用其确定文档中的姓名和文档主题之间的关系, 并将其作为后续聚类进行实体匹配的特征。Bhattachary 等人在文献[69]中提出了一种基于 LDA 的生成式模型, 并将实体是否匹配作为隐变量, 相似的实体隐变量相同。文献[70-71]也是基于主题模型的算法。

基于马尔科夫模型和贝叶斯分类器的算法也被大家广泛关注。Tang^[29]提出了一种用于姓名消歧的概率模型, 此方法将实体的属性和实体之间的关系紧密结合, 使用了隐马尔科夫随机场来建模不同研究人员匹配的难题, 此算法在一定意义上也可以看做是聚类算法。除此之外, 文献[72-73]也利用了概率图模型的思想并使用了马尔科夫链蒙特卡洛(Markov Chain Monte Carlo, MCMC)算法。

4 算法总结

第3节主要对两大类实体匹配算法进行了总结, 将算法分为基于监督学习和基于非监督学习两类。表3是对两类算法的整理, 主要从算法思想、涉及的算法以及算法的优缺点这4个方面来说明。在基于监督学习的实体匹配算法中, 准确性的提升必须以人工标注标签为代价; 基于非监督学习的算法不需要有标签的数据, 但是一般准确率低于监督学习; 非监督学习中的基于概率图的算法能够很好地挖掘内在文本数据信息以及实体间的联系, 但是其复杂性较高。

5 总结和展望

随着大家对数据重视程度的提高, 越来越多的数据可以被获取, 数据之间的匹配将变得愈发重要, 也将为数据的融合打下基础。实体匹配使得行为分析、行为预测、推荐系统、智慧医疗、智慧教育、用户画像等更加准确和可靠。因此, 在大数据时代, 实体匹配有着举足轻重的作用。本文主要基于对异构网络中实体匹配算法的调研, 对算法进行了分类和总结。

实体匹配也面临着很多要点和难点, 面对大规模数据, 其挑战也不断增多。主要面临的待解决的困难和研究点梳理如下。

(1) 虽然现在有很多技术针对数据量过大的问题提出了一些解决方案, 比如通过筛选减小训练集大小等, 但是数据量大导致的效率过低仍然是一个比较严重的问题。如何在算法效率和算法效果之间寻找最好的权衡是未来实体匹配研究的热点和难点。

(2) 真实数据缺少标注是一个很重要的问题, 单纯依赖人为干预成本过高, 即便使用非监督学习算法, 测试集也需要少量有标签的数据。如何自动标注或者使用非监督的算法进行实体匹配也是未来的主要研究热点。

(3) 目前的研究工作大多基于某一种语言, 虽然提出的算法在两种语言上效果都不错, 但是跨语言的实体匹配(比如, 两个数据源, 一个数据源是中文, 一个是英文)仍然是一个重要的研究

点。这个问题将来可能需要结合自然语言处理知识或精准翻译算法来进行解决。

(4) 随着隐私保护越来越受到重视, 脱敏数据下的实体匹配是一个研究热点; 除了显式可以挖掘的特征外, 如何更好地衡量用户行为的特征也是未来的一个研究方向。

(5) 除此之外, 大数据下的精准实时匹配也是未来研究的一个难题。精准匹配对算法的效果具有非常高的要求, 实时匹配对算法的效率也有非常高的要求。如何做到精准性和实时性二者兼顾是未来的需求和方向。

表 3 算法优缺点总结

Tab. 3 Advantages and disadvantages of different algorithms

类别	主要算法思想	涉及算法	优点	缺点
基于监督学习的实体匹配算法	使用数据集训练算法模型, 新的实体对直接使用此模型得到匹配结果。	支持向量机 ^[42] 、logistic回归 ^[44] 、AdaBoost ^[13] 、梯度提升树 ^[27,37] 等	准确性相对较高	若想得到较好的结果, 需要对分类算法的参数进行训练, 需要大量的有标签数据, 这在实际中难以满足需求
基于规则的算法主要通过人为制定匹配规则。	文献[45,47-49,51]	思路简单, 比较容易理解	人为定义规则, 非常消耗人力且定义规则存在主观性; 很难达到最优值。	
基于非监督学习的实体匹配算法	基于聚类的思想是类别内的数据比较相似, 类别之间的数据差异性比较大	K-Means算法 ^[29,53] 、层次聚类算法 ^[54-56]	不需要标注数据, 人力成本低	算法效果一般较差; 簇的个数若人为指定, 匹配效果很难最优; 若采用自动确定方式, 需要引用停止条件, 则会增加算法复杂度
基于概率图的算法	概率模型 ^[12,54,62-63,58] 、主题模型 ^[55,68-71] 、条件随机场、马尔建模成概率模型	科夫随机场 ^[29] 等	根据概率分布、文本内在的主题信息、关联信息等挖掘潜在信息, 考虑实体之间的联系, 效果比较好; 可利用隐空间信息, 这在没有交叉文本的匹配任务上占有非常大的优势	模型建立难度大; 算法复杂性较高

[参 考 文 献]

- [1] WU X, ZHU X, WU G Q, et al. Data mining with big data[J]. IEEE transactions on knowledge and data engineering, 2014, 26(1): 97-107.
- [2] 赵国栋. 大数据时代的历史机遇: 产业变革与数据科学[M]. 北京: 清华大学出版社, 2013.
- [3] GU B, LI Z, ZHANG X, et al. The interaction between schema matching and record matching in data integration[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 186-199.
- [4] LI C, JIN L, MEHROTRA S. Supporting efficient record linkage for large data sets using mapping techniques[J]. World Wide Web, 2006, 9(4): 557-584.
- [5] WHANG S E, GARCIA-MOLINA H. Incremental entity resolution on rules and data[J]. The VLDB Journal, 2014, 23(1): 77-102.
- [6] GETOOR L, MACHANAVAJJHALA A. Entity resolution: theory, practice & open challenges[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2018-2019.
- [7] TEJADA S, KNOBLOCK C A, Minton S. Learning Object Identification Rules for Information Integration[J]. Information Systems, 2001, 26(8):607-633.
- [8] LEITAO L, CALADO P, HERSCHEL M. Efficient and effective duplicate detection in hierarchical data[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(5): 1028-1041.

- [9] DUNN H L. Record linkage[J]. American Journal of Public Health and the Nations Health, 1946, 36(12): 1412-1416.
- [10] LIU S, WANG S, ZHU F, et al. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling[C]//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. ACM, 2014: 51-62.
- [11] LIU J, LEI K H, LIU J Y, et al. Ranking-based name matching for author disambiguation in bibliographic data [C]//Proceedings of the 2013 KDD Cup 2013 Workshop. ACM, 2013: Article No 8.
- [12] KONG C, GAO M, XU C, et al. Entity matching across multiple heterogeneous data sources[C]//International Conference on Database Systems for Advanced Applications. Cham: Springer, 2016: 133-146.
- [13] KEJRIWAL M, MIRANKER D P. Semi-supervised instance matching using boosted classifiers[C]//European Semantic Web Conference. Cham: Springer, 2015: 388-402.
- [14] KONDA P, DAS S, SUGANTHAN GC P, et al. Magellan: Toward building entity matching management systems[J]. Proceedings of the VLDB Endowment, 2016, 12(9): 1197-1208.
- [15] WHANG S E, MENESTRINA D, KOUTRIKA G, et al. Entity resolution with iterative blocking[C]//Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009: 219-232.
- [16] KARAPIPERIS D, VERYKIOS V S. An LSH-based blocking approach with a homomorphic matching technique for privacy-preserving record linkage[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(4): 909-921.
- [17] GRUENHEID A, DONG X L, SRIVASTAVA D. Incremental record linkage[J]. Proceedings of the VLDB Endowment, 2014, 9(7): 697-708.
- [18] KIM H, LEE D. HARRA: Fast iterative hashed record linkage for large-scale data collections[C]//Proceedings of the 13th International Conference on Extending Database Technology. ACM, 2010: 525-536.
- [19] ZHANG Y, TANG J, YANG Z, et al. Cosnet: Connecting heterogeneous social networks with local and global consistency[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1485-1494.
- [20] 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.
- [21] WANG J, LI G, YU J X, et al. Entity matching: How similar is similar[J]. Proceedings of the VLDB Endowment, 2011, 10(4): 622-633.
- [22] SONG Y, KIMURA T, BATJARGAL B, et al. Cross-language record linkage using word embedding driven metadata similarity measurement[C/OL]//International Semantic Web Conference (Posters & Demos). (2016-09-28)[2018-06-01]. <http://ceur-ws.org/vol-1690/paper90.pdf>.
- [23] YANG K H, PENG H T, JIANG J Y, et al. Author name disambiguation for citations using topic and web correlation [C]//International Conference on Theory and Practice of Digital Libraries. Berlin: Springer, 2008: 185-196.
- [24] VAN DER LOO M P J. The stringdist package for approximate string matching[J]. The R Journal, 2014, 6(1): 111-122.
- [25] LI C L, SU Y C, LIN T W, et al. Combination of feature engineering and ranking models for paper-author identification in KDD Cup 2013[C]//Proceedings of the 2013 KDD Cup 2013 Workshop. ACM, 2013: Article No 2.
- [26] PELED O, FIRE M, ROKACH L, et al. Entity matching in online social networks[C]//Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013: 339-344.
- [27] LI J, LIANG X, DING W, et al. Feature engineering and tree modeling for author-paper identification challenge[C]// Proceedings of the 2013 KDD Cup 2013 Workshop. ACM, 2013: Article No 5.
- [28] PELED O, FIRE M, ROKACH L, et al. Matching entities across online social networks[J]. Neurocomputing, 2016, 210(C): 91-106.
- [29] TANG J, FONG A C M, WANG B, et al. A unified probabilistic framework for name disambiguation in digital library[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(6): 975-987.
- [30] 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.
- [31] VESDAPUNT N, GARCIA-MOLINA H. Identifying users in social networks with limited information[C]//Data Engineering (ICDE), 2015 IEEE 31st International Conference on. IEEE, 2015: 627-638.
- [32] LEE J Y, HUSSAIN R, RIVERA V, et al. Second-level degree-based entity resolution in online social networks[J/OL]. Social Network Analysis and Mining, 2018, 8: 19. (2018-03-16)[2018-06-01]. <https://link.springer.com/content/pdf/10.1007%2Fs13278-018-0499-9.pdf>.
- [33] ZHOU X, LIANG X, ZHANG H, et al. Cross-platform identification of anonymous identical users in multiple social media networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(2): 411-424.

- [34] ZHANG J, YU P S, ZHOU Z H. Meta-path based multi-network collective link prediction[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 1286-1295.
- [35] LEVIN F H, HEUSER C A. Evaluating the use of social networks in author name disambiguation in digital libraries[J]. Journal of Information and Data Management, 2010(2): 183-197.
- [36] POOJA K M, MONDAL S, CHANDRA J. An unsupervised heuristic based approach for author name disambiguation[C]//Communication Systems and Networks (COMSNETS), 2018 10th International Conference on. IEEE, 2018: 540-542.
- [37] ZHONG E, LI L, WANG N, et al. Contextual rule-based feature engineering for author-paper identification[C]//Proceedings of the 2013 KDD Cup 2013 Workshop. ACM, 2013: Article No 6.
- [38] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 701-710.
- [39] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2016: 855-864.
- [40] TANG J, QU M, WANG M, et al. Line: Large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web, International World Wide Web, Conferences Steering Committee, 2015: 1067-1077.
- [41] ZHOU X, LIANG X, DU X, et al. Structure based user identification across social networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6): 1178-1191.
- [42] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint, arXiv:1301.3781v3 [cs.CV] 7 Sep 2013.
- [43] LIU S, WANG S, ZHU F. Structured learning from heterogeneous behavior for social identity linkage[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(7): 2005-2019.
- [44] DEY D. Entity matching in heterogeneous databases: A logistic regression approach[J]. Decision Support Systems, 2008, 44(3): 740-747.
- [45] ZAFARANI R, LIU H. Connecting users across social media sites: a behavioral-modeling approach[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2013: 41-49.
- [46] ZHOU Y, HOWROYD J, DANICIC S, et al. Extending naive bayes classifier with hierarchy feature level information for record linkage[C]//Workshop on Advanced Methodologies for Bayesian Networks. Cham: Springer, 2015: 93-104.
- [47] LI L, LI J, GAO H. Rule-based method for entity resolution[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(1): 250-263.
- [48] CHIN W S, ZHUANG Y, JUAN Y C, et al. Effective string processing and matching for author disambiguation[J]. The Journal of Machine Learning Research, 2014, 15(1): 3037-3064.
- [49] JIANG Y, LIN C, MENG W, et al. Rule-based deduplication of article records from bibliographic databases[J]. Database, 2014: Article ID bat086. DOI: 10.1093/database/bat086.
- [50] SETOGUCHI S, ZHU Y, JALBERT J J, et al. Validity of deterministic record linkage using multiple indirect personal identifiers: linking a large registry to claims data[J]. Circulation: Cardiovascular Quality and Outcomes, 2014, 7(3): 475-480.
- [51] EKTEFA M, JABAR M A, Sidi F, et al. A threshold-based similarity measure for duplicate detection[C]//Open Systems (ICOS), 2011 IEEE Conference on. IEEE, 2011: 37-41.
- [52] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967: 281-297.
- [53] MONIKA S, ANAND C, GNANAMURTHY R K. Analyzing the User Profile Linkage across Different Social Network Platforms[J]. International Journal of Computer Science and Engineering, 2016, 4(2): 1378-1383.
- [54] CEN L, DRAGUT E C, SI L, et al. Author disambiguation by hierarchical agglomerative clustering with adaptive stopping criterion[C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2013: 741-744.
- [55] SONG Y, HUANG J, COUNCILL I G, et al. Efficient topic-based unsupervised name disambiguation[C]//Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. ACM, 2007: 342-351.
- [56] QIAN Y, HU Y, CUI J, et al. Combining machine learning and human judgment in author disambiguation[C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM, 2011: 1241-1246.
- [57] FELLEGI I P, SUNTER A B. A theory for record linkage[J]. Journal of the American Statistical Association, 1969, 64(328): 1183-1210.

- [58] WINKLER W E. Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage[C]//Proceedings of the Section on Survey Research Methods, American Statistical Association. 1988: 667-671.
- [59] BOYD S, VANDENBERGHE L. Convex Optimization[M]. Cambridge: Cambridge University Press, 2004.
- [60] GAO M, LIM E P, LO D, et al. CNL: Collective network linkage across heterogeneous social platforms[C]//IEEE International Conference on Data Mining. IEEE Computer Society, 2015: 757-762.
- [61] YAKOUT M, ELMAGARMID A K, Elmeleegy H, et al. Behavior based record linkage[J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 439-448.
- [62] FANG Y, CHANG M W. Entity linking on microblogs with spatial and temporal signals[J]. Transactions of the Association for Computational Linguistics, 2014, 2(1): 259-272.
- [63] HAN H, XU W, ZHA H, et al. A hierarchical naive Bayes mixture model for name disambiguation in author citations[C]//Proceedings of the 2005 ACM Symposium on Applied Computing. ACM, 2005: 1065-1069.
- [64] SINGLA P, DOMINGOS P. Collective object identification[C]// International Joint Conference on Artificial Intelligence. [S.l.]: Morgan Kaufmann Publishers Inc, 2005: 1636-1637.
- [65] BALAJI J, MIN C, JAVED F, et al. Avatar: Large scale entity resolution of heterogeneous user profiles[C]//Proceedings of the 2nd Workshop on Data Management for End-To-End Machine Learning. ACM, 2018: Article No 2.
- [66] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine Learning Research, 2003(3): 993-1022.
- [67] ZHAO W X, JIANG J, WENG J, et al. Comparing twitter and traditional media using topic models[C]//European Conference on Information Retrieval. Berlin: Springer, 2011: 338-349.
- [68] YANG Y, SUN Y, TANG J, et al. Entity matching across heterogeneous sources[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1395-1404.
- [69] BHATTACHARYA I, GETOOR L. A latent dirichlet model for unsupervised entity resolution[C]//Proceedings of the 2006 SIAM International Conference on Data Mining, [S.l.]: Society for Industrial and Applied Mathematics, 2006: 47-58.
- [70] SEN P. Collective context-aware topic models for entity disambiguation[C]//Proceedings of the 21st International Conference on World Wide Web. ACM, 2012: 729-738.
- [71] TANG J, FANG Z, SUN J. Incorporating social context and domain knowledge for entity recognition[C]//Proceedings of the 24th International Conference on World Wide Web. [S.l.]: International World Wide Web Conferences Steering Committee, 2015: 517-526.
- [72] STEORTS R C, HALL R, FIENBERG S E. A bayesian approach to graphical record linkage and deduplication[J]. Journal of the American Statistical Association, 2016, 516(111): 1660-1672.
- [73] MAURYA A, TELANG R. Bayesian multi-view models for member-job matching and personalized skill recommendations[C]//Big Data, 2017 IEEE International Conference on. IEEE, 2017: 1193-1202.

(责任编辑: 李 艺)