计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20180691
57(1):175-187, 2020

基于因子图的不一致记录对消歧方法

徐耀丽　李战怀　陈　群　王艳艳　樊峰峰
(西北工业大学计算机学院　西安　710072)
(大数据存储与管理工业和信息化部重点实验室(西北工业大学)　西安　710129)
(yaolixu@mail.nwpu.edu.cn)

An Approach for Reconciling Inconsistent Pairs Based on Factor Graph

Xu Yaoli, Li Zhanhuai, Chen Qun, Wang Yanyan, and Fan Fengfeng
(*School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an* 710072)
(*Key Laboratory of Big Data Storage and Management (Northwestern Polytechnical University), Ministry of Industry and Information Technology, Xi'an* 710129)

Abstract　Entity resolution (ER) is a critical and fundamental problem in data integration and data cleaning systems. Although there have been numerous methods proposed for entity resolution, those approaches explicitly or implicitly depend on ad-hoc assumptions or employ different strategies. Given an ER task, there exist many inconsistent pairs due to conflicting results resolved by these approaches. It is of great challenges of reconciling these pairs without any labeled data: 1) without labeled data, it is impractical to estimate the performance of existing approaches and pick out the best; 2) although an optional way is to reconcile these conflicting results for a better and consistent labeling solution, an effective reconciliation mechanism for combining all hints remains to be investigated. To this end, an approach for reconciling inconsistent pairs based on factor graph is proposed. It firstly achieves inconsistent and consistent pairs through conducting existing entity resolution approaches for a given ER task. Secondly, the features that can indicate the matching status of inconsistent pairs, are extracted by leveraging techniques like kernel density estimation and matching information transfer and so on. Then these features are modeled as factor functions of the factor graph, which represents a joint probability distribution with factor weights. Finally, the weight of each factor is estimated based on the maximum likelihood estimation, and the inconsistent pairs are reconciled according to the distribution represented by the factor graph. Experimental results on real-world datasets show our method is effective and can outperform the state-of-the-art approach.

Key words　data integration; entity resolution; maximum likelihood estimation; inconsistent pair; kernel density estimation (KDE); factor graph

摘　要　实体解析(entity resolution, ER)是数据集成和清洗系统的关键基础问题.尽管有大量实体解析方法提出,但这些方法依赖隐式或显式的假设或采用不同的解析策略.对相同的实体解析任务进行处理

收稿日期:2018-10-12;**修回日期**:2019-06-24
基金项目:国家重点研发计划项目(2018YFB1003403);国家自然科学基金项目(61732014,61672432);陕西省自然科学基础研究计划项目(2018JM6086)
This work was supported by the National Key Research and Development Program (2018YFB1003403), the National Natural Science Foundation of China (61732014, 61672432), and the Natural Science Basic Research Plan in Shaanxi Province of China (2018JM6086).
通信作者:陈群(chenbenben@nwpu.edu.cn)

后,它们的结论存在冲突,产生了大量的不一致记录对.在没有给定标记数据的情况下,进行这类记录对的消歧处理具有很大的挑战:一方面当标签数据缺失时,评估现存方法的解析效果并选出最优的不可行,另一方面尽管可选的方法是协调这些冲突结果以得到一致的标记方案,但有效且融合所有提示信息的消歧策略还有待研究.为此,提出了一种基于因子图的不一致记录对消歧方法.该方法首先对某给定的实体解析任务使用现存的实体解析技术进行实体解析,得到一致或不一致的记录对;接着,用核密度估计、匹配信息传递等方法输出与不一致记录对是否匹配相关的特征,并把这些建模为因子图的因子函数,该因子图是一个带因子权重的联合概率分布;最后基于最大似然估计方法估计出各因子的权重,并基于该分布对不一致记录对进行消歧处理.实验结果表明:在真实的数据集合,该方法有效且优于现存最好的方法.

关键词 数据集成;实体解析;最大似然估计;不一致记录对;核密度估计;因子图

中图法分类号 TP391

大数据时代,信息化进程的迅猛发展促使各行各业都累积了大量的数据.但真实的数据存在各种各样的数据质量问题如数据不完整、不一致、和不满足实体同一性等,使得直接基于这些脏数据的分析和预测不能满足应用场景的需求.为此,大量的研究工作针对这些问题提出一系列清洗算法.其中,针对数据无法满足实体同一性现象,学术界和工业界研究了实体解析问题[1-3].实体解析也称实体匹配,是通过识别出所有描述真实世界同一个实体的记录对,来保证数据的实体同一性.它是数据集成或清洗系统的一个首要问题.

自实体解析问题第1次[4]被提出后,有大量实体解析方法被提出.部分实体解析方法[3,5-6]假设训练数据事先给定.文献[5-6]首先从训练数据中学习实体匹配规则,然后用这些规则来解析记录对是否匹配.文献[3]首先使用embedding技术如fastText[7]将每个属性的单词序列转换为相同维度的向量序列;然后训练双向序列模型Bi_RNN[8](bidirectional recurrent neural network)和序列比对模型来生成属性摘要向量;接着用比较函数构建出记录对的一系列特征;最后使用分类模型如多层感知器模型,训练一个二分类模型,进而进行实体解析.然而,现实场景中,面临一个新的实体解析任务时,事先标记好的数据集并不一定总是可用,而且某些领域数据需要由经验丰富的专家来标记.考虑到依赖专家来标记数据集会带来高昂的成本,本文所提的方法无需标注数据,因而具有更广泛的应用场景.

一些学者[2,9]从数据统计角度分析匹配特征,提出了无监督的实体解析方法.如文献[2]使用离群距离来估算记录对的匹配可能性;文献[9]使用机器学习的聚类算法,把具有类似特征的记录划分到匹配或不匹配组中.由于各种实体解析技术有特定的假设,以及解析任务的复杂性,对同一个实体解析任务处理的结果,存在大量的不一致记录对.例如在文献数据集Cora上,使用11种方法(如Rule,Distance,Cluster等)得到的解析结果中,一致匹配对的数目仅有1 013;而不一致记录对的数目则高达44 909.所谓一致匹配对是指所有的方法一致认为该记录对是匹配的;而不一致记录对是指部分方法(如Rule)认为该记录对是匹配的,而其余方法(如Distance)认为不匹配.本文的工作侧重于解决这些不一致的记录对.

对不一致记录对进行消歧处理面临很大挑战.一方面,在没有标签数据情况下,直接选出所有方法中最好的是不现实的.另一方面,假设能够选出综合表现最好的方法,某些记录对(如已选中的方法无法有效处理,而其他方法可以处理的匹配记录对)的信息就不能够得到充分利用.鉴于此,我们利用因子图把各类匹配特征统一且有效地利用起来,并提出了基于因子图的不一致记录对消歧方法.

与记录对是否匹配相关的信息大致可分为3类:1)记录对自身匹配特征.例如使用字符串相似度,度量记录对属性值的相似程度.2)匹配传递特征.例如一个记录对(r_1,r_2)匹配后,对其他记录对(r_1,r_3)是否匹配的影响.3)外部匹配特征.例如个体方法的解析结果.为便于陈述,本文把现存的实体解析方法称为个体方法.概率图模型[10]的因子图能够利用因子函数灵活地为变量之间的关系建模.它首先把记录对$p_{i,j}$是否匹配视为待推断变量$m(p_{i,j})$,而把与$m(p_{i,j})$相关的其他匹配特征看成是已知变量,$m(p_{i,j})$为二元变量.当$m(p_{i,j})=1$时,表示r_i和r_j是匹配的;当$m(p_{i,j})=0$时,r_i和r_j是不匹配的.接着,使用核密度估计和图的连通性来拟合出

已知变量和未知变量之间的关系,并形式化为因子图的因子函数.这样,我们就能把不一致记录对消歧问题,建模为一个随机变量概率推断问题,其中因子图中因子的权重使用最大似然估计来推算.

本文的主要贡献有3个方面:

1) 首次提出了一个基于因子图的不一致记录对消歧框架FG-RIP.该框架不依赖标记数据,能通过因子图汇总各种异构匹配特征,如记录对自身匹配特征、匹配传递特征和现存实体解析技术的外部匹配特征,来估算不一致记录对的匹配可能性.

2) 设计并实现了基于最大似然估计的因子权重学习算法.该算法可以自动组合匹配特征的权重,并输出最优的匹配特征权重组合.

3) 在真实的数据集上,大量的实验结果表明该算法能明显提升个体方法的解析效果.

1 不一致记录对消歧问题

实体解析(entity resolution, ER)就是给定记录集合 D,识别出所有表示真实世界同一实体的记录对 $p=(r_i,r_j)$,$i\neq j$,其中,$r_i\in D$ 且 $r_j\in D$.如表1所示,文献数据集Cora包括记录的唯一标识 *rID*、论文的作者信息 *author*、标题 *title* 和页码信息 *pages*.实体解析就是找出所有表示同一个实体的记录对如 $p_{2,3}$.给定一个实体解析方法 M,它的输入是候选记录对的特征,输出是所有预测为匹配的记录对,记为 $P(M)$.为了与消歧算法相互区分,本文把现存的实体解析方法称为个体方法.假定一系列个体方法的输出结果如表2所示,*pID* 是记录对的唯一标识符,M_k 列为第 k 个方法的预测结果,其中

Table 1 Dataset

表1 数据集

rID	*author*	*title*	*pages*
r_1	p.auer, n.cesa-bianchi, y.freund, and r.e.schapire.	gambling in a rigged casino: the adversarial multi-armed bandit problem,	pp.322-331.
r_2	nicolo cesa-bianchi, yoav freund, david p.helmbold, david haussler, robert e.schapire, and manfred k.warmuth.	how to use expert advice.	pages 382-391,
r_3	nicolo cesa-bianchi, yoav freund, david p.helmbold, david haussler, robert e.schapire, and manfred k.warmuth.	how to use expert advice.	pages 382-391,
r_4	n.cesa-bianchi, y.freund, d.p.helmbold, and m.warmuth.	on-line prediction and conversion strategies.	pages 205-216,
r_5	n.cesa-bianchi, y.freund, d.p.helmbold, and m.warmuth.	on-line prediction and conversion strategies.	pages 205-216,
r_6	n.cesa-bianchi, y.freund, d.p.helm-bold, and m.warmuth.	on-line prediction and conversion strategies.	pages 205-216,
r_7	n.cesa-bianchi, y.freund, d.p.helmbold, d.haussler, r.e.schapire, and m.k.warmuth.	how to use expert advice.	pages 382-391,
r_8	a.blum, m.furst, m.j.kearns, and richard j.lipton.	cryptographic primitives based on hard learning problems.	pages 24.1-24.10,

Table 2 Output of Individual Methods

表2 个体方法的解析结果

pID	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	GT
$p_{1,2}$	P	N	P	P	P	N	P	P	N	P	N
$p_{1,3}$	N	P	P	P	P	P	P	P	P	P	N
$p_{2,3}$	P	P	N	P	P	P	P	P	N	N	P
$p_{2,5}$	N	P	P	P	P	P	P	P	P	P	P
$p_{3,6}$	N	P	P	P	P	P	P	N	P	P	P
$p_{4,5}$	P	P	P	P	P	P	P	P	P	P	P
$p_{4,6}$	P	P	P	P	P	P	P	P	P	P	P
$p_{5,6}$	P	P	P	P	P	P	P	P	P	P	P
$p_{7,8}$	N	N	N	N	N	N	N	N	N	N	N

Notes: $M_k(k=1,2,\cdots,10)$ represents the predicted label by the k-th method, GT represents the ground truth label of each record pair, N represents the predicated label of a pair is matching, P represents the predicated label of a pair is non-matching.

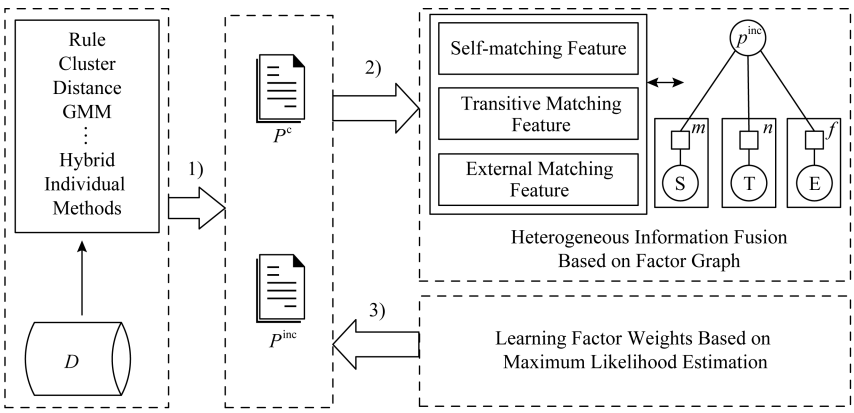
$1 \leqslant k \leqslant 10$,P(positive)(或 N(negative))代表记录对的预测状态为匹配(或不匹配);GT(ground truth)列是记录对的真实状态.不一致记录对 p^{inc} 是只被部分个体方法预测为匹配的记录对,如 $p_{2,3}$.一致记录对 p^{c} 是被个体方法统一预测为匹配的或者不匹配的记录对.p^{c} 包括一致匹配对 p^{cp} 如 $p_{4,5}$,和一致不匹配对 p^{cn} 如 $p_{7,8}$.

所谓不一致记录对的消歧问题,就是给定一系列个体方法和一致记录对集合 P^{c},推断 P^{inc} 中不一致记录对是否匹配.例如表 2 中不一致记录对的消歧问题就是已知个体方法的解析结果 $M_1, M_2, \cdots, M_{10}$ 和一致记录对集合 $P^{c}=\{p_{4,5}, p_{4,6}, p_{5,6}, p_{7,8}\}$,推断 $P^{inc}=\{p_{1,2}, p_{1,3}, p_{2,3}, p_{2,5}, p_{3,6}\}$ 中不一致记录对的匹配状态.

2 不一致记录对消歧框架

本节首先概述基于因子图的不一致记录对消歧框架 FG-RIP,接着详细介绍它的 2 个关键模块:基于因子图的异构信息融合(heterogeneous information fusion based on factor graph)和基于最大似然估计的因子权重学习(learning factor weights based on maximum likelihood estimation).

FG-RIP 的处理流程如图 1 所示:



D is the dataset which consists of records. P^{c} and P^{inc} represent the consistent pair set and inconsistent pair set respectively. p^{inc} is an inconsistent pair. S, T and E represent the variable node of self-matching feature, transitive matching feature and external matching feature of p^{inc} respectively. m, n and f are the numbers of the S, T and E nodes.

Fig. 1 A workflow for reconciling inconsistent pairs

图 1 不一致记录对消歧流程图

1) 利用现存的个体方法(individual methods)解析数据集 D,并得到 P^{c} 和 P^{inc}.

2) 构建自身匹配特征(self-matching feature,缩写为 S),即计算记录对的属性值相似度和使用核密度估计来量化不一致记录对的匹配可能性;利用图的连通性构建一致匹配记录对和不一致记录对之间的匹配传递特征(transitive matching feature,缩写为 T);利用个体方法的解析结果构建外部匹配特征(external matching feature,缩写为 E),进而构建与不一致记录对相关的因子图.

3) 用最大似然估计方法计算因子图中因子的权重.最后使用这些权重估计不一致记录对的边缘概率密度,并判断不一致记录对是否匹配.

2.1 基于因子图的异构信息融合

在概率图模型中,因子图 $G=(V,E)$ 类似于二部图,其中 V(vertices)是顶点集合,E(edges)是顶点之间边的集合.它的顶点集合包括 2 类节点:因子节点 f 和变量节点 v.边只存在于因子节点和变量节点之间,而因子节点(或变量节点)之间没有边.因子节点定义了与它相连的变量节点的联合概率分布.与二部图的不同点是因子图的因子节点定义了一个概率分布,而二部图没有概率分布的含义.

本文利用因子图把与记录对是否匹配相关的异构信息综合起来,以便量化记录对匹配的概率和不匹配的概率.与不一致记录对是否匹配相关的异构特征有 3 类:记录对的自身匹配特征、匹配传递特征和外部匹配特征.因子图模型首先把不一致记录对是否匹配,以及与不一致记录对是否匹配的相关特征看成是随机变量,然后构建这些随机变量的联合概率分布.它的处理过程是:1)把不一致记录对(如 $p_{i,j}$)是否匹配 $m(p_{i,j})$ 看成是一个二元随机变量,把与 $m(p_{i,j})$ 相关的 3 类特征看成是随机变量集合 V,这些随机变量构成因子图的变量节点或因子节点;2)使用指数函数 exp(·)来构建 V 中变量与

$m(p_{i,j})$之间的函数关系,并作为因子节点的因子函数;3)借鉴最大似然估计的思想计算出各个因子的权重;4)计算 $p_{i,j}$匹配的概率 $p(m(p_{i,j})=1|V)$和 $p_{i,j}$不匹配的概率 $p(m(p_{i,j})=0|V)$.如果 $p(m(p_{i,j})=1|V)$大于 $p(m(p_{i,j})=0|V)$,那么 $p_{i,j}$匹配,反之不匹配.为减少不必要的计算,本文所指的概率分布都是未经归一化的概率分布,因为对于某不一致记录对 p^{inc}而言,p^{inc}的匹配状态只取决于匹配概率和不匹配概率的相对大小.对未归一化的 p^{inc}匹配概率和不匹配概率而言,它们的归一化因子是相同的,使用未归一化的概率值不影响匹配状态的预测结果.

2.1.1 自身匹配特征

记录对的自身匹配特征,主要考虑:1)不一致记录对 p^{inc}中相应属性值之间的相似度 $sim(p^{\text{inc}},n^{\text{attr}})$;2)使用核密度估计拟合出属性值的相似度 $sim(p^{\text{inc}},n^{\text{attr}})$与记录对是否匹配 $m(p^{\text{inc}})$之间的概率分布.n^{attr}(attribute name)是属性名,$sim(\cdot)$可以是任意相似度度量.本文使用 Jaccard 相似系数来度量 2 个属性值的相似度.

我们把属性值相似度建模为因子节点.该类因子节点的因子函数 f_{S},是 $m(p^{\text{inc}})$与 $sim(p^{\text{inc}},n^{\text{attr}})$之间的分布律,具体数学描述为

$$f_{\text{S}}(m(p^{\text{inc}}),sim(p^{\text{inc}},n^{\text{attr}}))=\begin{cases}\exp(sim(p^{\text{inc}},n^{\text{attr}})), & m(p^{\text{inc}})=1,\\ \exp(1-sim(p^{\text{inc}},n^{\text{attr}})), & m(p^{\text{inc}})=0,\end{cases}\tag{1}$$

其中,S(self-matching)代表基于相似度的自身匹配特征;n^{attr}是属性名;$sim(p^{\text{inc}},n^{\text{attr}})$是 p^{inc}的 n^{attr}属性值相似度;exp(•)是指数函数;$m(p^{\text{inc}})$是一个布尔变量,如果 $m(p^{\text{inc}})=1$,那么 p^{inc}匹配,否则 $m(p^{\text{inc}})=0$,p^{inc}不匹配.例如表 2 中 $p_{2,3}$的 *title* 属性,有 $sim(p_{2,3},title)=1$,那么 $p_{2,3}$的因子图包含一个因子节点,且该因子节点的因子函数定义为

$$f_{\text{S}}(m(p_{2,3}))=\begin{cases}\text{e}, & m(p_{2,3})=1;\\ 1, & m(p_{2,3})=0.\end{cases}\tag{2}$$

核密度估计(kernel density estimation, KDE)是一种无参数的密度估计技术,它能够依据样本拟合出属性值相似度和记录对是否匹配之间的概率密度函数.对每个属性 n^{attr},我们使用了所有的一致匹配对集合 P^{cp}和一致不匹配对集合 P^{cn}的子集(也就是在 n^{attr}上与 P^{inc}相似的)作为核密度估计的输入样本.对于任意 p^{inc}和 n^{attr},依据拟合出的密度函数和属性值相似度 $sim(p^{\text{inc}},n^{\text{attr}})$,我们就可以计算 p^{inc}匹配/不匹配的概率值.匹配的概率值越高,表明 p^{inc}匹配的可能性越大.

本文使用 scikit-learn 提供的核密度估计算法 KernelDensity[11].它的核心思想是以给定的一系列样本值作为观察值集合,以一个非线性函数作为核函数,对于某待判断的样本 x_{q},它的概率密度值由观察值集合 X 的样本与 x_{q} 的差值决定,具体计算为

$$p(x_{\text{q}})=n\Big(\sum_{i=1}^{N}K((x_{\text{q}}-x_i)/h)\Big),\tag{3}$$

其中,N 是样本的数目;$n(\cdot)$是一个归一化函数,保证 $p(x_{\text{q}})\in[0,1]$.h 是一个平滑因子,用来权衡偏差和方差.h 越大,密度函数 $p(x_{\text{q}})$具有高的偏差,也越光滑;反之,h 越小,$p(x)$具有高的方差,即越不光滑.为了避免核密度估计遇到不连续性问题,本文采用平滑的高斯函数作为核函数:

$$K(x;h)\propto\exp\left(-\frac{x^2}{2h^2}\right).\tag{4}$$

类似属性值相似度特征,我们把由 KDE 拟合的记录匹配特征看成是另一种自身匹配特征.这类因子节点的因子函数 f_{SKDE}定义为

$$f_{\text{SKDE}}(m(p^{\text{inc}}),p(v))=\begin{cases}\exp(p(v)), & m(p^{\text{inc}})=1,\\ \exp(1-p(v)), & m(p^{\text{inc}})=0,\end{cases}\tag{5}$$

$$p(v)=p(sim(p^{\text{inc}},n^{\text{attr}})).\tag{6}$$

式(6)是把 $sim(p^{\text{inc}},n^{\text{attr}})$作为式(3)的输入得到的,SKDE(self-matching KDE)代表基于核密度估计的自身匹配特征.

2.1.2 匹配传递特征

假设把每个记录看成是无向图的顶点,如果某个记录对被预测为匹配,那么相关的记录之间存在 1 条实线边,如图 2(a)中 r_1 和 r_2 所示.对于待判断记录对如 $p_{2,5}$,该记录对的相关记录之间存在 1 条虚线边,表示待预测状态.在消歧过程中,处于待预测状态的记录对 $p_{i,j}$可分为 2 种情况:

情况 1. r_i 和 r_j 分别属于不同的连通图,记为 $c=1$,如图 2(a)的 r_2 和 r_5,c 表示记录对所属情况.

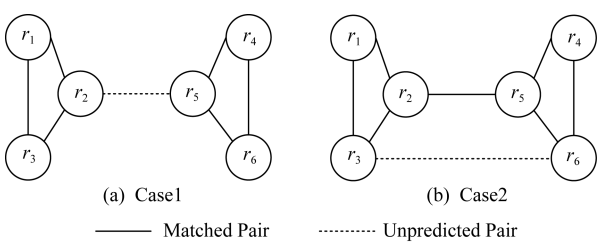


Fig. 2 Relationships among matched pairs and unpredicted pairs

图 2 匹配的记录对与待预测的记录对之间的关系

情况 2. r_i 和 r_j 属于同一个连通图,记为 $c=2$,如图 2(b)的 r_3 和 r_6.

情况 1 如图 2(a)所示.假设已知记录对 $p_{1,2}$,$p_{1,3}$,$p_{2,3}$,$p_{4,5}$,$p_{4,6}$,$p_{5,6}$处于匹配状态,那么记录对 $p_{2,5}$是否匹配的信息可以从 r_2 和 r_5 各自所在的连通图中得到.由实体解析的定义可知,处于匹配状态的记录对 $p_{2,3}$,表明 r_2 和 r_3 是现实世界中同一个实体 e_{rec} 的 2 个不同描述.对某个属性 n^{attr} 来说,$r_2[n^{\text{attr}}]$和 $r_3[n^{\text{attr}}]$是 $e_{\text{rec}}[n^{\text{attr}}]$的近似描述.当判断 r_2 和 r_5 是否匹配时,可以用 $sim(r_3[n^{\text{attr}}],r_5[n^{\text{attr}}])$作为匹配传递特征.传递特征的因子函数定义与自身匹配特征的因子函数定义类似.

情况 2 如图 2(b)所示.由于 $p_{2,5}$处于匹配状态,待预测记录对 $p_{3,6}$的 r_3 和 r_6 处于同一个连通图.我们在因子图中添加传递变量节点 v_{T} 和传递因子节点 f_{T}.f_{T} 的因子函数为

$$f_{\text{T}}(m(p_{3,6}))=\begin{cases}\text{e}, & m(p_{3,6})=1,\\ 1, & m(p_{3,6})=0,\end{cases} \tag{7}$$

其中,$m(p_{3,6})$是待预测的变量节点,且 $m(p_{3,6})=1$ 表示 $p_{3,6}$匹配,而 $m(p_{3,6})=0$ 表示 $p_{3,6}$不匹配.为便于陈述,本文把与匹配传递特征相关的节点称为传递变量节点或传递因子节点,并用 T(transitive)表示匹配传递特征.

综上所述,对于匹配传递特征,我们按照算法 1 构建与 p^{inc} 相关的传递变量节点和传递因子节点.

算法 1. 匹配传递特征的因子节点和变量节点的构建.

输入:不一致记录对 p^{inc}和一致匹配对集合 P^{cp};

输出:p^{inc}的变量节点、传递变量节点集 V_{T} 和传递因子节点集 F_{T}.

① 根据 P^{cp}构建无向图.

② 对于待预测状态的记录对 p^{inc},利用图的连通性,判断 p^{inc}是属于情况 1 还是情况 2.

③ 若 p^{inc}属于情况 1,首先构建一个待预测变量节点 $m(p^{\text{inc}})$,并为它的每个属性 n^{attr}构建一个传递变量节点 $v_{\text{T}}\in V_{\text{T}}$ 和一个传递因子节点 $f_{\text{T}}\in F_{\text{T}}$.考虑到某些待预测状态的记录对,会产生较多的匹配传递特征.例如在图 2(a)中,$p_{2,5}$的 n^{attr}属性的匹配传递特征包括 $sim(r_3[n^{\text{attr}}],r_5[n^{\text{attr}}])$,$sim(r_1[n^{\text{attr}}],r_5[n^{\text{attr}}])$,$sim(r_2[n^{\text{attr}}],r_4[n^{\text{attr}}])$,$sim(r_2[n^{\text{attr}}],r_6[n^{\text{attr}}])$.鉴于此,对于每个属性 n^{attr},本文只选择这些匹配特征中 $sim(\cdot)$最小的,记为 $ms(p^{\text{inc}},n^{\text{attr}})$,作为 p^{inc}在 n^{attr}上传递变量节点 v_{T} 的变量值.f_{T} 的因子函数定义为

$$f_{\text{T}}(m(p^{\text{inc}}),ms(p^{\text{inc}},n^{\text{attr}}))=\begin{cases}\exp(ms(p^{\text{inc}},n^{\text{attr}})), & m(p^{\text{inc}})=1;\\ \exp(1-ms(p^{\text{inc}},n^{\text{attr}})), & m(p^{\text{inc}})=0.\end{cases} \tag{8}$$

④ 若 p^{inc} 属于情况 2,首先构建变量节点 $m(p^{\text{inc}})$,并为它的每个属性构建一个传递变量节点 $v_{\text{T}}\in V_{\text{T}}$ 和一个传递因子节点 $f_{\text{T}}\in F_{\text{T}}$,其中 f_{T} 的因子函数定义为

$$f_{\text{T}}(m(p^{\text{inc}}))=\begin{cases}\text{e}, & m(p^{\text{inc}})=1;\\ 1, & m(p^{\text{inc}})=0.\end{cases} \tag{9}$$

2.1.3 外部匹配特征

与传统的实体解析方法不同,本文的消歧方法是对传统实体解析方法的结果中不一致记录对进行预测.这些不一致记录对具有个体方法的投票信息.本文把这些投票信息归类为外部匹配特征,并用 E(external)表示外部匹配特征.p^{inc}的外部匹配特征包括:1)p^{inc}获得的投票比例;2)个体方法关于 p^{inc}的投票信息.

依据 p^{inc}获得的投票比例,可构建与源自投票比例的外部匹配特征相关的因子节点和变量节点.假设使用 $k(k>1)$个不同的个体方法,且有 $N^{p^{\text{inc}}}$ 个方法预测 p^{inc} 为匹配的,那么在因子图中添加因子节点 f_{E}、变量节点 v_{E} 和待预测变量节点 $m(p^{\text{inc}})$.其中 $v_{\text{E}}=N^{p^{\text{inc}}}/k$,$f_{\text{E}}(m(p^{\text{inc}}),v_{\text{E}})$是 f_{E} 的因子函数,其数学形式为

$$f_{\text{E}}(m(p^{\text{inc}}),N^{p^{\text{inc}}}/k)=\begin{cases}\exp(N^{p^{\text{inc}}}/k), & m(p^{\text{inc}})=1;\\ \exp(1-N^{p^{\text{inc}}}/k), & m(p^{\text{inc}})=0.\end{cases} \tag{10}$$

依据个体方法的投票信息,可构建与源自个体方法的外部匹配特征相关的因子节点和变量节点.例如,对于某一个体方法 M_k 和 p^{inc},可构建个体方法层面的因子节点 f_{E}、变量节点 v_{E}^k 和待预测变量节点 $m(p^{\text{inc}})$,其中 $v_{\text{E}}^k=1$ 表示 M_k 认为 p^{inc}是匹配的,而 $v_{\text{E}}^k=0$ 表示 M_k 认为 p^{inc}是不匹配的.f_{E} 的因子函数定义为

$$f_{\text{E}}(m(p^{\text{inc}}),v_{\text{E}}^k)=\begin{cases}\exp(v_{\text{E}}^k), & m(p^{\text{inc}})=1;\\ \exp(1-v_{\text{E}}^k), & m(p^{\text{inc}})=0.\end{cases} \tag{11}$$

2.2 基于最大似然估计的因子权重学习

对于某不一致记录对 p^{inc},假设已经得到与它相关的所有因子函数如 $f_1,f_2,\cdots,f_l$,其中 l 是因子节点的数目,那么 p^{inc}是否匹配的变量 $m(p^{\text{inc}})$与所有相关的变量 $V=\{v_1,v_2,\cdots,v_l\}$之间的联合概率密度可定义为

$$p(m(p^{\text{inc}}),v_1,v_2,\cdots,v_l;w_1,w_2,\cdots,w_l)=\exp(w_1)\times f_1\times\exp(w_2)\times f_2\times\cdots\times\exp(w_l)\times f_l, \quad (12)$$

其中 w_i 是因子 f_i 的权重.

最大似然估计是一种估计总体分布中未知参数的方法,它的核心思想是概率最大的事件最有可能出现.由于因子节点的权重是未知的,为了估计这些权重,我们采用最大似然估计的思想,极大化观察数据,也就是最大化变量 V 关于参数 W 的对数似然函数:

$$L(W)=\sum_{i=1}^{n}\ln\sum_{m(p^{\text{inc}})}p(m(p^{\text{inc}}),V;W), \quad (13)$$

其中 $V=\{v_1,v_2,\cdots,v_l\}$ 是除了 $m(p^{\text{inc}})$ 以外的所有变量集合;$W=\{w_1,w_2,\cdots,w_l\}$ 是所有权重的集合;n 是待消歧的不一致记录对的数目 $|P^{\text{inc}}|$.

为了能够得到这些因子函数之间的相对重要性,我们增加了一个约束即 $w_1+w_2+\cdots+w_l=1$.这样,因子权重的估计问题就可以转换为一个有约束的最大似然估计问题.假设能够事先获得各个因子函数的权重信息或者因子函数权重之间的关系,就可以直接把这些先验知识形式化为似然函数 $L(W)$ 的约束条件.而最优的权重就是 $\hat{W}=\arg\max(L(W))$.

我们使用 scipy 提供的信任区域约束算法(trust region constrained algorithm)[12]求解有约束的最优化问题.

3 实验与结果

本节概述了实验的运行环境,并在真实数据集 Cora 和 Song 上验证算法的有效性.所有实验的运行环境配置为 Intel® Core™ i7-4710MQ 2.50 GHz 处理器、16 GB 内存和 Ubuntu 16.04 64 位的操作系统.编程语言是 Python 3.服务器端数据库是 MongoDB.

3.1 度量标准

本文采用实体解析文献[5,13-14]广泛使用的查准率、查全率和 F_1 来评价算法的有效性.由于个体方法是从候选记录对开始处理,所以,在与个体方法的对比中,使用的是全部候选记录对的查全率、查准率和 F_1,而其他实验使用不一致记录对集合的查全率、查准率和 F_1.所谓查准率是指预测为匹配且真正匹配的记录对数目,与预测为匹配的记录对数目的比值,记为 P_{pre}.查全率是指预测为匹配且真正匹配的记录对数目,与所有真正匹配的记录对数目的比值,记为 R_{rec}.F_1 是查准率和查全率的调和平均值,具体定义为

$$F_1=\frac{2\times R_{\text{rec}}\times P_{\text{pre}}}{R_{\text{rec}}+P_{\text{pre}}}. \quad (14)$$

3.2 数据集

本文在 Cora 和 Song 数据集上测试提出的方法.下面我们从数据集特点和记录对方面介绍这 2 个数据集.

数据集 Cora[15]是一个文献数据.它包含 1 295 个记录,而这些记录隶属于 112 个实体的某一个.每个记录由 12 个属性描述如文献的作者列表和标题等.我们将这些记录对两两比较,得到的候选记录对数目为 837 865.本文处理的对象是不一致记录对.对 Cora 数据而言,不一致记录对的数目为 44 909,一致匹配对的数目是 1 013,而一致不匹配对的数目是 791 943.

数据集 Song[16]是一个歌曲数据.它包含 100 000 个记录,每个记录由 7 个属性描述,例如歌曲的专辑名和发布时间等.我们抽取了其中的 20 744 个记录进行实验.这些记录对经过 blocking 技术过滤后,得到的候选记录对的数目是 260 181,其中不一致记录对的数目为 115 258,一致匹配对的数目为 651,一致不匹配对的数目为 144 272.

3.3 对比方法概述

本文采用的个体方法总共有 11 个,包括:1)5 个无监督的解析方法,分别是基于 RR 规则[6]的方法 Rule、基于离群距离的方法 Distance[2]、基于 k-means 的 Cluster[9]、基于高斯混合模型的 GMM[10]、基于狄利克雷过程的变分贝叶斯高斯混合模型 DPBGM[17];2)6 个基于学习的解析方法,分别是基于支持向量机的 SVM[18]、基于决策树模型的 CART[19]、基于随机森林的 ERT[20]、基于高斯朴素贝叶斯模型的 GNB[11]、基于多层感知器的 MLP[8]和基于深度学习技术的 Hybrid[3].

各个无监督方法的解析过程逐一概述为:Rule 方法是使用事先给定的匹配规则来判断记录对是否匹配,其规则形式与文献[6]提出的 RR 规则相同.Distance 方法[2]首先计算离群距离,然后依据离群距离和匹配约束来判断记录对是否匹配.Cluster 方法是使用开源的机器学习库 scikit-learn 来复现文献[9]中提到的 k-means 聚类解析方法.GMM 和 DPBGM 方法是把实体解析问题等价于将候选记录对划分为匹配组和不匹配组的聚类问题.GMM 即

高斯混合模型[10],假设匹配记录对和不匹配记录对分别服从2个参数未知的高斯分布,而观测数据来自这2个高斯分布的混合模型.GMM通过EM算法[18]学习该模型的未知参数,进而使用训练好的模型将记录对划分为匹配组和不匹配组.DPBGM是高斯混合模型的一种变体,即高斯混合的变分贝叶斯估计模型.DPBGM与GMM的不同点是,DPBGM使用变分推断估计模型的参数.

基于学习的解析方法SVM,CART,ERT,GNB,MLP是机器学习领域的分类模型.这些模型的主要思想是构建样本特征,训练二分类模型,并预测记录对是否匹配,其中样本的特征是记录对的相应属性值的相似度构成的向量.这些模型的不同点在于模型构建原理.本实验的SVM是非线性支持向量机,其模型构建原理是在特征空间中搜索一个超平面,用来把记录对集合划分为匹配组和不匹配组.CART的模型构建原理是使用特征和阈值构建二叉树,其中每个树节点都具有最大的信息收益.ERT的模型构建原理是首先使用训练集的子样本构建一系列随机决策树,然后使用这些决策树的解析结果均值来进行最终的判断.GNB算法与朴素贝叶斯算法的相同点是基于贝叶斯定理和类条件独立性假设;两者的不同点是GNB假设在给定类标签后,每个特征服从高斯分布.MLP模型是一个二分类的前馈神经网络,本实验中MLP的输入层是记录对的属性相似度特征,输出层是记录对的匹配状态.隐藏层包括2层:第1层的神经元数目是5;第2层的神经元数目是2.它的损失函数是交叉熵损失函数,优化算法是拟牛顿方法L-BFGS.本实验调用scikit-learn库中这些算法的API实现接口[11].本实验的Hybrid算法,首先使用预训练的embedding模型把每个属性的单词序列转换为固定维度的向量序列;然后训练并融合双向序列模型Bi_RNN和序列比对模型来构建属性摘要向量;接着用比较函数计算记录对的属性相似度描述向量;最后多层感知器以训练集的属性相似度描述向量为输入,训练出二分类模型进行实体解析.所有监督或无监督的现存实体解析方法,都可以作为消歧框架的个体方法,但必须有至少2个不依赖标签数据,且有差异的个体方法.这样基于学习的解析方法就可以用有差异的无监督方法的输出结果中的一致的部分作为训练集.有差异的无监督个体方法越多,训练集的纯度越高.所谓训练集的纯度是指训练集中标签正确的记录对所占的比例.

消歧方法GL-RF[21]是针对Clean-Clean ER场景[22]下消歧算法.本文将该算法的匹配约束去掉,修改为可以处理Dirty-Dirty ER场景下的消歧算法,并进行了对比实验.

3.4 实验和结果

本节中,我们进行4组实验来验证FG-RIP的有效性.

实验1. 与个体方法进行了对比,验证了FG-RIP算法能自动组合出在F_1指标上最好的方法.这些个体方法的实验结果如表3,4所示,最大值已用黑体标出.

Table 3 Comparison with Individual Methods on Song

表3 Song数据集上与个体方法的对比

Individual Method	Recall	Precision	F_1
Rule	**0.998 2**	0.754 4	0.859 4
Distance	0.311 9	0.084 8	0.133 3
Cluster	0.978 5	0.894 0	0.934 4
GMM	0.993 1	0.065 9	0.123 6
DPBGM	0.394 6	0.037 5	0.068 4
SVM	0.476 6	0.646 7	0.548 7
CART	0.992 9	0.976 1	0.984 4
ERT	0.572 3	0.963 5	0.718 1
GNB	0.403 6	**0.999 0**	0.574 9
MLP	0.942 6	0.974 6	0.958 4
Hybrid	0.660 4	0.911 2	0.765 8
FG-RIP	0.979 9	0.997 3	**0.988 5**

Note: The maximum values are in bold.

Table 4 Comparison with Individual Methods on Cora

表4 Cora数据集上与个体方法的对比

Individual Method	Recall	Precision	F_1
Rule	0.922 4	0.566 3	0.701 8
Distance	0.064 1	**0.854 8**	0.119 2
Cluster	0.819 6	0.762 5	0.790 0
GMM	0.929 9	0.154 4	0.264 9
DPBGM	**0.991 6**	0.033 4	0.064 7
SVM	0.895 4	0.756 4	0.820 1
CART	0.888 2	0.754 8	0.8161
ERT	0.902 9	0.758 6	0.824 5
GNB	0.680 2	0.64	0.659 5
MLP	0.895 2	0.755 4	0.819 4
Hybrid	0.949	0.651 3	0.772 5
FG-RIP	0.945 6	0.758 6	**0.841 9**

Note: The maximum values are in bold.

由表 3,4 可以得出:

1) 在相同的数据集上,个体方法各有所长.①对 Song 数据而言,在所有的个体方法中,Rule 具有最高的查全率和相对较高的查准率,这说明了基于属性值相似度和阈值的领域规则能够有效识别出记录对,但存在准确率欠缺的不足.这也表明,查全率高而查准率低的个体方法如 Rule,GMM 等,有助于过滤掉不匹配的记录对,同时保障真正匹配的记录对以较高的概率落入不一致记录对集合中.另外,尽管 GNB 具有最高的查准率,但查全率较低.这表明被 GNB 预测为匹配的记录对,具有较高的可信性.这也表明对于查准率高而查全率低的方法如 GNB,可有效保证真正匹配记录对以较高概率落入一致匹配记录对集合中.CART 和 MLP 算法的查准率、查全率和 F_1 均达到 90%以上.这表明属性值相似度特征和较少的模型参数就能够为 Song 数据集训练较好的解析模型.②对 Cora 数据而言,Distance 具有最高的查准率和极低的查全率.这是由于樊峰峰等人[2]提出了一个基于主成分分析的离群距离.对某个记录 r_i,该离群距离能够找到与该记录匹配概率最高的记录对 r_j,其中 $i \neq j$.对于数据集中某实体 e_{rec},假设只有 2 个记录 r_i 和 r_i 描述该实体 e_{rec},那么 Distance 的解析结果较好.而 Cora 数据集中 e_{rec} 对应多个记录,使得 Distance 识别为匹配的记录对,有较高概率是真正匹配的.而其他匹配的记录对被解析为不匹配的,导致了极低的查全率.DPBGM 具有最高的查全率,且 GMM 具有较高的查全率.这些说明在没有标签数据的情况下,选择一个适合所有数据集的方法具有很大的挑战.

2) 在不同的数据集上,大部分个体方法的查准率、查全率和 F_1 有差异.例如 SVM 在 Cora 数据上 F_1 值与最高的 F_1 值相差 2.18%,却在 Song 数据上相差 43.98%.这是由于 SVM 模型依赖于来自数据集的支持向量,若这些支持向量无法有效分类样本,则模型的分类效果较差.Hybrid 在 Cora 数据上有较高的查全率,而在 Song 数据集上有较高的查准率.这是由于文献[3]融合了 embedding 技术、双向序列模型、序列比对模型和多层感知器来训练出一个二分类模型.由于该模型的参数多(如在 Song 和 Cora 数据集上,需要训练的参数数目分别为 9 210 006 和 26 545 814),但训练样本集合(即一致记录对集合)有限,且训练样本集合(一致记录对集合)和测试样本集合(不一致记录对集合)属于不同的分布,导致训练的模型过拟合即在训练数据集上查准率、查全率和 F_1 高达 90%以上,而在 Song 数据集的测试集合上,仅查准率较高;而在 Cora 数据集上,仅查全率最高.另外,某些个体方法不受数据集差异的影响.比如 GMM 在 Cora 和 Song 数据上均有高达 90%以上的查全率和较低的查准率,这说明混合高斯模型对数据的差异不敏感,具有较好的健壮性.

3) 平均来看,监督模型的解析效果优于无监督的.例如对 Song 数据而言,6 个监督模型的平均查全率、查准率和 F_1,依次为 0.674 7,0.911 9,0.756 9;而 5 个无监督模型的平均查全率、查准率和 F_1 依次为 0.735 3,0.367 3,0.423 8.对 Cora 数据而言,6 个监督模型的平均查全率、查准率和 F_1 依次为 0.868 5,0.719 4,0.785 4;而 5 个无监督模型的平均查全率、查准率和 F_1 依次为 0.745 5,0.474 3,0.388 1.这表明,尽管在消歧问题上,训练集(一致记录对集合)和测试集(不一致记录对集合)不满足独立同分布,且在理论上,当训练样本和测试样本是独立同分布时,监督模型才有最好的效果,但由于部分监督模型如 CART 和 MLP,具有较好的健壮性,能在非独立同分布场景下,训练出较好的二分类模型.如在 Song 数据的消歧问题上,CART 和 MLP 的查准率、查全率和 F_1 高达 90%.这也表明通过分析一致记录对的特征有助于更好地预测不一致记录对的匹配状态.

4) FG-RIP 能够更准确地解析出更多的匹配记录,具有最好的综合指标 F_1.由表 3,4 可知,与其他个体方法相比,FG-RIP 具有较高的查准率和查全率.这是由于 FG-RIP 把个体方法的解析结果作为外部匹配特征,并综合记录对的自身匹配特征和匹配传递特征来进行消歧.在 Cora 数据集上,虽然 FG-RIP 的查准率没有 Distance 高,但它具有最好的综合指标 F_1.类似地,在 Song 数据集上,FG-RIP 也具有最高的 F_1 值.这些表明,基于因子图的消歧算法 FG-RIP 能自动组合各类特征,获得最优的综合指标.

实验 2. 与已存在的消歧方法 GL-RF 进行了对比.我们分别从 Song 和 Cora 数据集的不一致记录对集合中抽取了 1 000 个和 2 000 个进行对比实验.如表 5,6 所示:

Table 5 Comparison with GL-RF on Song

表 5 Song 数据集上与 GL-RF 的对比

Individual Method	Recall	Precision	F_1
GL-RF	1.0	0.991	0.995 5
FG-RIP	1.0	0.991	0.995 5

Table 6 Comparison with GL-RF on Cora

表 6 Cora 数据集上与 GL-RF 的对比

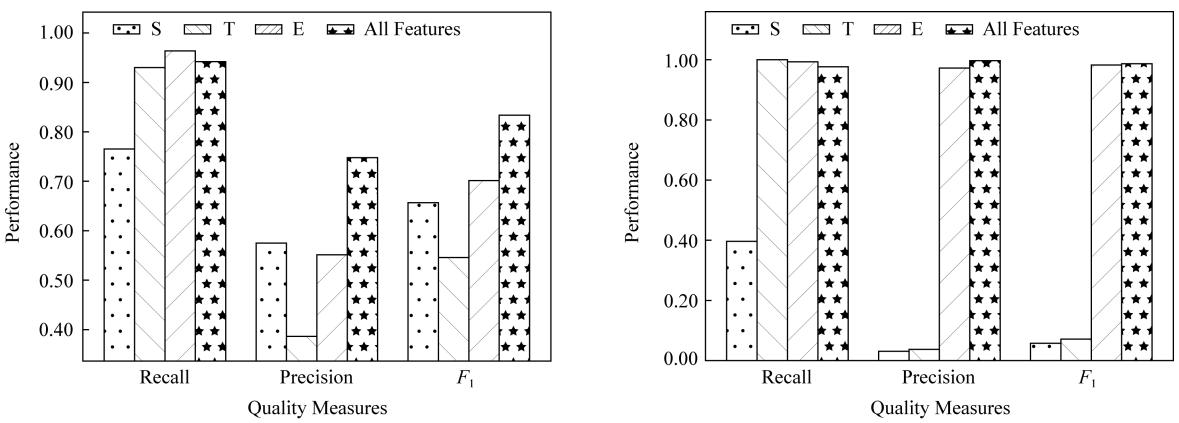
Individual Method	Recall	Precision	F_1
GL-RF	0.015 9	0.324 3	0.030 3
FG-RIP	1.0	0.754 0	0.859 7

表 5,6 所示的实验对比结果可得出:

1) 改造后的 GL-RF 可以有效处理部分 Dirty-Dirty ER 场景下的不一致记录对消歧问题.由表 5 可知,在 Song 数据集上,GL-RF 和 FG-RIP 的实验结果相同.这是由于虽然 Song 数据属于 Dirty-Dirty ER 场景,但 Song 的不一致记录对集中匹配记录对与 Clean-Clean ER 场景的大体吻合.具体来说,在 Clean-Clean ER 场景下,某个记录最多有 1 个与之相匹配的记录.Song 数据上,92.81%的记录最多有 1 个与之匹配的记录;7.19%的记录与 2 个以上的其他记录相匹配,因而 GL-RF 能有效地处理这类消歧问题.

2) FG-RIP 算法在 Cora 数据集上优于 GL-RF.这是由于:①从数据集的角度看,Cora 的不一致记录对集中匹配记录对与 Clean-Clean ER 场景差异较大.具体来说,Cora 数据中,仅 6.88%的记录最多有 1 个与之匹配的记录;93.12%的记录与 2 个以上的记录相匹配;有的记录甚至有 83 个与之匹配的记录.这导致改进后的 GL-RF 在 Cora 数据中消歧效果有限.②从方法的角度看,GL-RF 只考虑了一致记录对和不一致记录对之间的距离关系,适合识别与某个记录最匹配的记录.而 FG-RIP 把各类匹配特征建模为特征因子,自动组合最优消歧模型,因而能更准确地识别匹配记录对.

实验 3. 本文使用查准率、查全率和 F_1 指标,分析了不同的因子特征对 Cora 和 Song 数据的消歧结果的影响.在图 3(a)(b)中,如第 2 节所述,S,T,E 分别代表记录对自身匹配特征、匹配传递特征、外部匹配特征.我们可以观察到:1)匹配传递特征 T 在 2 个数据集上均具有较高的查全率.这表明特征 T 有助于识别更多的匹配记录对.2)记录对自身匹配特征 S 消歧效果表现不稳定.在 Cora 数据上,S 具有较高的 F_1,而在 Song 数据上具有较低的 F_1.3)外部匹配特征 E 具有较好的消歧效果.如 E 的查全率在 Cora 数据上最高,且 F_1 值高于记录对自身匹配特征 S 和匹配传递特征 T.这表明融合个体方法的特征能有效地发挥个体方法识别匹配记录对的能力.4)所有特征表现稳定,在所有数据集上均能取得较好的效果.以 F_1 指标为例,所有特征的解析效果在 Song 和 Cora 数据上达到最高.



(a) Cora (b) Song

Fig. 3 Performance comparison on different factors

图 3 不同因子的消歧效果对比

实验 4. 分析基于相似度的核密度估计技术对 FG-RIP 方法的影响.如图 4 所示,KDE 代表 FG-RIP 算法只使用基于相似度的核密度估计特征;-KDE 代表 FG-RIP 算法使用去掉核密度估计特征后所有其他特征.本质上,在相似度值的基础上,用核密度估计技术计算概率值,相当于把原来的相似度特征从线性空间变换到非线性空间.变化后的特征对非线性可分的数据集有效,而对线性可分的数据集效果不明显.由图 4 的实验结果可知:1)当变化后的特征空间能有效划分候选记录对集合的匹配记录对和不匹配记录对时,基于核密度估计技术的特征可获得较好的消歧效果.如对于 Cora 而言,与-KDE 相比,变化后的特征空间(KDE)有较高的查全率、查准率和 F_1 值,即能提供更好的分类效果;

而对于 Song 而言,原有特征空间(-KDE)的消歧质量指标均高于 KDE 的消歧质量.2)基于全部特征的消歧算法具有一定的鲁棒性.在核密度估计特征有效时,它的消歧效果略低于核密度估计特征的效果;在核密度估计特征无效时,它受其消歧效果的影响小.①当变化后的特征空间能有效区分匹配对和不匹配对时,仅仅使用基于相似度的核密度估计特征,FG-RIP 就能获得最好的消歧效果,甚至在某些质量指标上略高于所有特征.如在 Cora 数据集上,KDE 的查全率和 F_1 值最高.所有特征的查全率和 F_1 值低于 KDE 的相应值,表明在 Cora 数据中增加非 KDE 特征后,质量指标有所下降,但降低的幅度不大,如所有特征的 F_1 值仅减低了 0.49%.②当变化后的特征空间不能有效地区分匹配对和不匹配对时,KDE 的消歧效果较差.如在 Song 数据集上,KDE 的所有质量指标最低.但所有特征的消歧质量受 KDE 特征的影响小.由于缺少标签数据,事先评估 KDE 的消歧效果不可行.鉴于此,在实际应用场景中,建议使用全部的特征,以便消歧算法具有较好的鲁棒性.

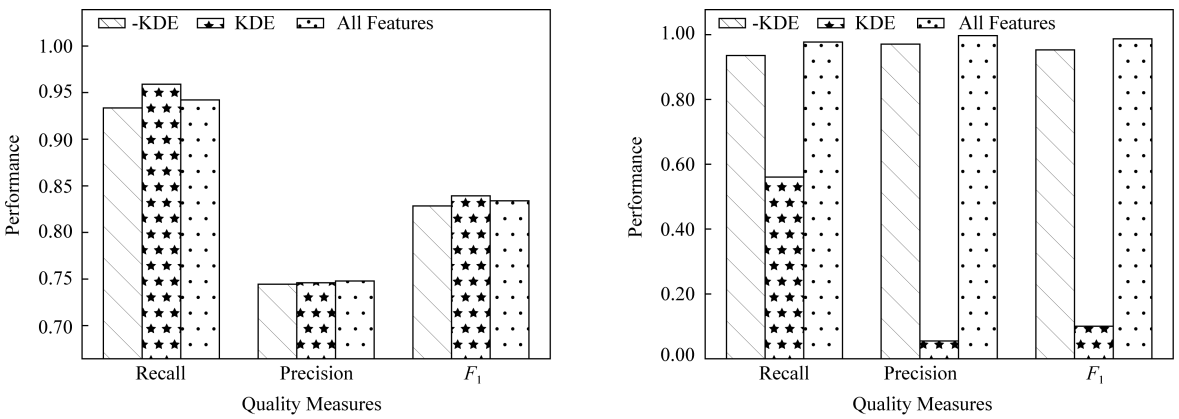


Fig. 4 Performance comparison on kernel density estimation

图 4 核密度估计的消歧效果对比

4 相关工作

由于实体解析问题是数据集成和清洗系统的核心基础问题,且在很多领域有广泛的应用,领域专家和学者提出了一系列的解析技术.这些相关的工作可划分为三大类:基于学习的[3,5-6]、基于统计的[2,9,23-25]、和基于人机配合的[26-27].

基于学习的实体解析方法[3,5-6]首先使用训练数据集来学习实体匹配的模式或者规则,接着用训练好的模式或规则判定记录对是否匹配.文献[3]提出了 4 种属性级摘要表示方法(分别是聚合模型(SIF)、基于循环神经网络的序列模型(RNN)、注意力模型(Attention)和基于序列和注意力的混合模型(Hybrid)),并使用神经网络来训练实体解析模型.文献[6]从已知的匹配(和不匹配)记录对集合中学习属性级别的匹配规则集(ARs),并组合这些规则成一系列记录级别的规则集(RRs).文献[5]提出了一种描述记录和实体之间匹配关系的规则(ER-rule),并设计了从训练集中自动学习 ER-rule 的算法和高效的在线实体解析算法.这些方法都依赖标签数据,且不能解决不一致记录对消歧问题,而我们的方法假设标签数据不存在,侧重于标签数据缺失场景下消歧问题.

基于统计的实体解析方法[2,9,23-25]是通过统计分析记录对的匹配特征,并形式化为某个合适度量,进而选取合适的阈值来将记录对划分为匹配的或者不匹配的.比如文献[23]分别从单词的简写和全写,前缀关系和字符近似匹配方面提出了 3 种字段匹配度量算法.文献[24]提出了 Footrule Distance,用来输出与给定元组最相关的 top k 个记录值.针对大多数算法未体现关键属性重要性的不足,文献[25]利用信息增益或统计概率的方法计算属性权重,并提出了基于这些属性权重的最终相似度来提升实体解析的准确率.文献[2]提出了离群距离,并证明了离群距离与记录对匹配的可能性是正相关的.这类方法的优点是不需要训练数据集和额外的训练过程,只需要估算出合适的匹配度量和阈值.但由于实体解析应用场景的复杂性,匹配度量和阈值的确定很难做到适用于所有的场景.文献[9]使用统计机器学习

的聚类算法如 k-means 方法,把候选记录对分成两组:匹配组和不匹配组.这类不依赖标签数据的方法都可以作为消歧框架的个体方法,用来对数据集进行预先处理,以得到一致或不一致记录对.

针对全自动的实体解析方法不能彻底解决实体解析问题,基于人机配合的实体解析方法[26-27]提出借用用户的知识以人机配合的方式进行解析.文献[26]将实体解析过程分为设计阶段和执行阶段.设计阶段就是用户在样本数据上使用现成的工具灵活地构建出一个实体解析工作流;在执行阶段,用户可调用支持大数据处理的工具来执行设计好的工作流.文献[27]提出了基于规则的候选记录对生成阶段和基于 crowd 细化匹配记录对 2 个阶段.这 2 个阶段都有人和自动算法的共同参与,这类方法需要人的参与,无法自动完成不一致记录对的消歧处理.

与本文的消歧算法最相关的是 GL-RF[21].该算法的核心思想是首先基于 TF.IDF 计算记录对的向量表示;接着分析与每个不一致记录对距离最近的前 k 个一致匹配对和一致不匹配对,对于该不一致记录对的影响;最后当前 k 个一致匹配对的影响大于前 k 个一致不匹配对时,该不一致记录对为匹配,反之为不匹配.文献[22]依据数据集是否存在重复记录,把实体识别问题区分为 3 种场景:Clean-Clean ER, Dirty-Clean ER, Dirty-Dirty ER.其中 Clean-Clean ER 是指左数据源和右数据源都没有重复记录.FG-RIP 与 GL-RF 的区别有 2 点:1)GL-RF 是针对 Clean-Clean ER 场景,而 FG-RIP 则是针对 Dirty-Dirty ER 场景;2)GL-RF 没有考虑个体方法的解析结果的可信程度,而 FG-RIP 使用因子权重来区分个体方法的解析结果.

5 总结和展望

本文研究了在没有标签数据场景下不一致记录对消歧问题,并首次提出了基于因子图的不一致记录对消歧框架.该框架利用因子图融合与不一致记录对相关的特征(包括自身匹配特征、匹配传递特征和外部匹配特征),并使用最大似然估计计算因子图中因子的权重.实验结果表明:该算法能够有效地学习到合适的权重,并自动组合出最优的消歧方案.在没有标签的场景下,自动估计不同特征的解析效果很有挑战性,也具有深远的现实意义.因为自动估计不同特征的解析结果,并选择最优的特征组合,可进一步提升解析的效果.比如在 Cora 数据集上只使用记录对的基于相似度的核密度估计特征就能获得更好的解析效果.鉴于此,我们把无标签场景下,不同特征消歧结果的质量估计问题作为将来的研究问题.

参 考 文 献

[1] 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

[2] Fan Fengfeng, Li Zhanhuai, Chen Qun, et al. An outlier-detection based approach for automatic entity matching [J]. Chinese Journal of Computers, 2017, 40(10): 2197-2211 (in Chinese)
(樊峰峰, 李战怀, 陈群, 等. 一种基于离群点检测的自动实体匹配方法[J].计算机学报, 2017, 40(10): 2197-2211)

[3] Mudgal S, Li H, Rekatsinas T, et al. Deep learning for entity matching: A design space exploration [C] //Proc of the 44th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2018: 19-34

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

[5] Li Lingli, Li Jianzhong, Gao Hong. Rule-based method for entity resolution [J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(1): 250-263

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

[7] Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information [J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146

[8] Goodfellow I, Bengio Y, Courville A. Deep Learning [M]. Cambridge, MA: MIT Press, 2016

[9] Christen P. Febrl: A freely available record linkage system with a graphical user interface [C] //Proc of the 2nd Australasian Workshop on Health Data and Knowledge Management. Darlinghurst: Australian Computer Society, Inc, 2008: 17-25

[10] Bishop C M. Pattern Recognition and Machine Learning [M]. Berlin: Springer, 2006

[11] Buitinck L, Louppe G, Blondel M, et al. API design for machine learning software: Experiences from the scikit-learn project [OL]. [2018-07-01]. https://arxiv.org/abs/1309.0238

[12] Lalee M, Nocedal J, Plantenga T. On the implementation of an algorithm for large-scale equality constrained optimization [J]. SIAM Journal on Optimization, 1998, 8(3): 682-706

[13] Köpcke H, Thor A, Rahm E. Evaluation of entity resolution approaches on real-world match problems [J]. Proceedings of the VLDB Endowment, 2010, 3(1): 484-493

[14] Chen Zhaoqiang, Chen Qun, Fan Fengfeng, et al. Enabling quality control for entity resolution: A human and machine cooperation framework [C] //Proc of the 34th IEEE Int Conf on Data Engineering. Los Alamitos, CA: IEEE Computer Society, 2018: 1156-1167

[15] Bilenko M. Cora [OL]. [2018-09-24]. http://www.cs.utexas.edu/users/ml/riddle/data/cora.tar.gz

[16] Doan A H. Song [OL]. [2018-09-24]. http://pages.cs.wisc.edu/~anhai/data/falcon_data/songs.tar.gz

[17] Blei D M, Jordan M I. Variational inference for Dirichlet process mixtures [J]. Bayesian Analysis, 2006, 1(1): 121-143

[18] Li Hang. Statistical Learning Method [M]. Beijing: Tsinghua University Press, 2012 (in Chinese)
(李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012)

[19] Breiman L, Friedman J, Olshen R, et al. Classification and Regression Trees [M]. Boca Raton, FL: CRC, 1984

[20] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees [J]. Machine Learning, 2006, 63(1): 3-42

[21] Xu Yaoli, Li Zhanhuai, Chen Qun, et al. GL-RF: A reconciliation framework for label-free entity resolution [J]. Frontiers of Computer Science, 2018, 12(5): 1035-1037

[22] Papadakis G, Koutrika G, Palpanas T, et al. Meta-blocking: Taking entity resolution to the next level [J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(8): 1946-1960

[23] Monge A E, Elkan C. The field matching problem: Algorithms and applications [C] //Proc of the 2nd Int Conf on Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI, 1996: 267-270

[24] Guha S, Koudas N, Marathe A, et al. Merging the results of approximate match operations [C] //Proc of the 30th Int Conf on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann, 2004: 636-647

[25] Zhen Lingmin, Yang Xiaochun, Wang Bin, et al. An entity resolution approach based on attributes weights [J]. Journal of Computer Research and Development, 2013, 50(增刊1): 281-289 (in Chinese)
(甄灵敏, 杨晓春, 王斌, 等. 基于属性权重的实体解析技术[J]. 计算机研究与发展, 2013, 50(Suppl1): 281-289)

[26] Konda P, Das S, Suganthan G C P, et al. Magellan: Toward building entity matching management systems [J]. Proceedings of the VLDB Endowment, 2016, 9(12): 1197-1208

[27] Li Guoliang. Human-in-the-loop data integration [J]. Proceedings of the VLDB Endowment, 2017, 10(12): 2006-2017

Xu Yaoli, born in 1987. PhD candidate. Student member of CCF. Her main research interests include data repairing, entity resolution and machine learning.

Li Zhanhuai, born in 1961. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include database theory, data management, big data analysis and data quality.

Chen Qun, born in 1976. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include data management, big data analysis and data quality.

Wang Yanyan, born in 1991. PhD candidate. Her main research interests include opinion mining, sentiment analysis and deep learning.

Fan Fengfeng, born in 1986. PhD. His main research interests include data quality, entity resolution and machine learning.