软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(3):695−709 [doi: 10.13328/j.cnki.jos.005900]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

时间约束的实体解析中记录对排序研究*

孙琛琛[1,2], 申德荣[3], 李玉坤[1,2], 肖迎元[1,2], 马建红[4]

[1](计算机视觉与系统教育部重点实验室(天津理工大学),天津 300384)

[2](天津市智能计算及软件新技术重点实验室(天津理工大学),天津 300384)

[3](东北大学 计算机科学与工程学院,辽宁 沈阳 110189)

[4](河北工业大学 人工智能与数据科学学院,天津 300401)

通讯作者: 孙琛琛, E-mail: dustinchenchen_sun@163.com

摘 要: 实体解析是数据集成和数据清洗的重要组成部分,也是大数据分析与挖掘的必要预处理步骤.传统的批处理式实体解析的整体运行时间较长,无法满足当前(近似)实时的数据应用需求.因此,研究时间约束的实体解析,其核心问题是基于匹配可能性的记录对排序.通过对多路分块得到的块内信息与块间信息分别进行分析,提出两个基本的记录匹配可能性计算方法.在此基础上,提出一种基于二分图上相似性传播的记录匹配可能性计算方法.将记录对、块及其关联关系构建二分图;相似性沿着二分图不断地在记录对结点与块结点之间传播,直到收敛.收敛结果可以通过不动点计算得到.提出近似的收敛计算方法来降低计算代价,从而保证实体解析的实时召回率.最后,在两个数据集上进行实验评价,验证了所提出方法的有效性,并测试方法的各个方面.

关键词: 实体解析;记录对排序;时间约束;数据集成

中图法分类号: TP18

中文引用格式: 孙琛琛,申德荣,李玉坤,肖迎元,马建红.时间约束的实体解析中记录对排序研究.软件学报,2020,31(3):695−709. http://www.jos.org.cn/1000-9825/5900.htm

英文引用格式: Sun CC, Shen DR, Li YK, Xiao YY, Ma JH. Research on record pair ranking for entity resolution with time constraint. Ruan Jian Xue Bao/Journal of Software, 2020,31(3):695−709 (in Chinese). http://www.jos.org.cn/1000-9825/5900.htm

Research on Record Pair Ranking for Entity Resolution with Time Constraint

SUN Chen-Chen[1,2], SHEN De-Rong[3], LI Yu-Kun[1,2], XIAO Ying-Yuan[1,2], MA Jian-Hong[4]

[1](Key Laboratory of Computer Vision and System of Ministry of Education (Tianjin University of Technology), Tianjin 300384, China)

[2](Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology (Tianjin University of Technology), Tianjin 300384, China)

[3](School of Computer Science and Engineering, Northeastern University, Shenyang 110189, China)

[4](School of ArtificialIntelligence, Hebei University of Technology, Tianjin 300401, China)

Abstract: Entity resolution (ER) is an important aspect of data integration and data cleaning, and is also a necessary pre-process step of big data analytics and mining. Traditional batch based ER's overall runtime is costly, and cannot satisfy current (nearly) real-time data applications' requirements. Therefore, time constraint entity resolution (TC-ER) is focused on, while core problem is record pair ranking

* 基金项目: 国家重点研发计划(2018YFB1003404); 国家自然科学基金(61672142, 61472070, 61602103); 天津市自然科学基金(17JCYBJC15200)

Foundation item: National Key Research and Development Program of China (2018YFB1003404); National Natural Science Foundation of China (61672142, 61472070, 61602103); Natural Science Foundation of Tianjin of China (17JCYBJC15200)

本文由人工智能赋能的数据管理、分析与系统专刊特约编辑李战怀教授、于戈教授和杨晓春教授推荐.

收稿时间: 2019-07-15; 修改时间: 2019-09-10; 采用时间: 2019-11-25; jos 在线出版时间: 2020-01-10

CNKI 网络优先出版: 2020-01-10 14:29:55, http://kns.cnki.net/kcms/detail/11.2560.TP.20200110.1429.015.html

according to match probability both information inner blocks and information across blocks are analyzed from multi-pass blocking respectively, and two basic recordsmatch probability methods are proposed. The basic methods are improved by proposing an advanced record match probability method based on similarity flowing over a biparitite graph.A bipartite graph is constructed according to record pairs, blocks, and relations between them. Similarities iteratively flow between pair nodes and block nodes over the bipartite graph until convergence. The convergence result is computed with fixpoint iterations. An approximate convergence computation mehod is proposed to reduce cost, and it improves real-time recall in TC-ER. Finally, the proposed methods are evaluated on two datasets, which shows their effectiveness and also tests different aspects of the proposed methods.

Key words: entity resolution; record pair ranking; time constraint; data integration

实体解析(entity resolution,简称 ER)是数据集成和数据清洗的重要组成部分,它将数据源中描述相同实体的记录分到同一组[1−15].大数据具有多样性的特点,描述同一实体的记录可能以多种形式出现,成为大数据可用性的一个瓶颈,因此,ER 是大数据分析与挖掘的必要预处理操作[15].传统的 ER 包括分块、相似性计算和匹配决定等步骤,将整个脏数据集作为输入,批处理之后整体输出解析结果[1,3].在大数据时代,一方面,数据产生的速度和更新的频率比以往更快;另一方面,大量(近似)实时的数据分析应用出现要求有限的时间内解析出尽量多的匹配记录,称为时间约束的实体解析(entity resolution with time constraint,简称 TC-ER),传统的批处理 ER 无法满足这种新需求.

当前有很多时间约束的 ER 应用,例如犯罪侦查应用中要求近似实时的实体解析,希望在较短时间内解析出一部分嫌疑人记录来,以便及时地布署侦查行动.尽管短时间内解析的结果不完整,但及时的解析结果可以大大增加抓捕到嫌疑人的可能性.再例如网购比价服务(如一淘网)中,互联网用户搜索了一件商品后,系统将尽快返回一部分匹配的商品条目,并逐渐优化搜索结果,这样可以提升用户体验,因为众所周知,互联网用户是没有耐心的.

时间约束的 ER 希望在给定的短时间(远少于批处理运行时间)内将解析结果最大化.TC-ER 的关键在于实体解析过程中的记录对选择,即优先选择匹配可能性大的记录对进行解析.Whang 等人提出了 3 个基于“线索”的启发式 Pay-as-you-go ER 方法,其中的“线索”分别是排序的记录对列表、记录集合的层次划分和排序的记录列表[6].Papenbrock 等人提出一组基于排序的记录列表的渐进式 ER 方法,其中,渐进式滑动窗口方法将变化的窗口多次滑过排序列表生成候选对;渐进式分块方法将排序列表划分成等规模小块,然后渐进地扩大分块范围[7].Papenbrock 等人提出的基于排序列表的方法要优于 Whang 等人提出的基于“线索”的方法[7].这些方法都假定已知最优分块键或排序键,并且无法对记录对进行全局排序,因此可用性和实时召回率都比较受限.由此可见,已有的时间约束的 ER 方法有较大的改进空间.

本文研究时间约束的实体解析中记录对排序,通过优先选择匹配可能性高的记录对进行解析,来保证实时的召回率.分块是 ER 中降低计算代价的基本的、有效的手段[16−26],然而单凭分块方法无法实现时间约束的 ER.整体而言,将分析和挖掘分块信息来估计记录对的相似性.将脏数据集进行多路分块后生成有交叠的块集合,如果一个块包含的记录越多,那么块内记录的匹配可能性越小;如果两条记录共同出现的块数目越多,那么它们的匹配可能性越大.首先,基于这些直观的思想,提出两个基本的记录对相似性估计方法,分别利用了块内信息和块间信息.接下来,通过考虑记录对的相似性与块的质量之间的相互影响来改进基本的相似性估计方法.将记录对、块及其关联关系映射成二分图;然后相似性在二分图上迭代地传播,直到收敛,获得最终的相似性.基于图传播的相似性估计充分挖掘了分块的隐藏信息,从而更有效.提出了基于不动点迭代的收敛结果计算方法,然而其计算代价较大;进一步提出了近似的收敛结果计算方法,力求在不影响记录对相似性估计有效性的前提下降低计算代价,从而保证时间约束的 ER 的实时召回率.通过实验评估,证明了提出方法的有效性.

本文的主要贡献总结如下:

- 提出两种基本的记录对相似性估计方法,分别利用了块的质量(块内信息)和记录与不同块的隶属关系(块间信息);
- 提出了基于相似性传播的记录对相似性估计方法,利用二分图上可收敛的相似性传播来衡量记录对

的相似性,通过不动点迭代来计算收敛结果,并提出了近似方法来降低计算代价;

- 在两个数据集上,通过与已有方法的对比测试,证明了本文提出方法的有效性;此外,对比了不同的相似性估计方法的表现,并测试了迭代次数对基于相似性传播的记录对相似性估计方法的影响.

本文第 1 节定义研究的问题,并概括地介绍研究框架.第 2 节介绍两种基本的记录对相似性估计方法.第 3 节提出基于二分图上相似性传播的记录对相似性估计方法,并通过近似方法降低计算代价.第 4 节在两个数据集上评价本文提出的方法,验证其有效性.第 5 节介绍相关工作.最后总结全文,并指出下一步可能的研究方向.

1 研究概述

定义 1(实体解析). 给定一个脏数据集 $R=\{r\}$,ER 将描述相同实体的记录分到一组,$C=\{c_k|\forall r_i\in c_k, r_i\in R\wedge \varphi(r_i)=e_k\wedge\nexists r_j\in c_l\wedge\varphi(r_j)=e_k\}$,其中,$\varphi(\cdot)$是从记录到实体的映射函数,$e_k$ 表示分组 c_k 对应的实体,c_l 为不同于 c_k 的一个分组.

如图 1 所示,ER 传统上是批处理操作,通常包括 3 个步骤:分块、相似度计算和匹配决定,其中,前者是可选步骤,后两者是必要步骤[1].

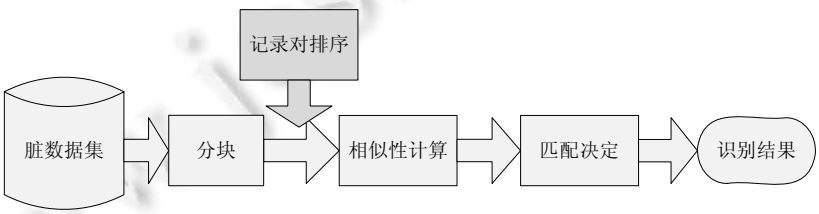


Fig.1 Entity resolution model

图 1 实体解析模型

(1) 相似性计算

利用记录相似性函数计算两条记录的相似性,通常,相似性表示为[0,1]范围内的数值.两条记录的相似性越大,匹配可能性越大,0 表示不可能匹配,1 表示完全匹配.记录通常包括多个属性,比如,一条个人信息的记录包括姓名、年龄、工作单位、城市、省份和邮编等,不同的属性需要使用不同的相似性函数来计算相似性.记录属性以文本型为主,以数字型为辅.针对文本属性,目前已有多种字符串相似性函数,如 TF-IDF、Q-gram、Jaccard、编辑距离等[27].针对数字属性,则需要采用专门的函数进行比较,比如差值、汉明距离等.记录相似性函数选择多个属性,分别选择适合的相似性函数来计算属性相似性,最后将多个属性相似性聚集得到记录相似性,聚集方式包括线性组合、非线性组合等,与匹配决定的策略相关.

(2) 匹配决定

根据记录的相似性来决定记录是否匹配有两类方法:分类和聚类.基于分类的匹配决定使用支持向量机(support vector machine,简称 SVM)、遗传算法、主动学习和决策树等方法来决定记录对是否匹配[3].一部分分类方法是监督的,需要专家标注大量的训练数据,从而学习出有效的匹配规则(即分类器).还有一部分分类方法的匹配规则是由领域专家定义的,需要较多的领域知识.基于聚类的匹配决定使用 MinCut,Markov Clustering 等聚类算法来处理成对的相似性,得到的聚类结果即为实体解析结果.同一类簇表示同一实体,不同类簇表示不同实体[2,28].本文将 ER 当作分类问题,认为匹配规则已获得,记作 $m(*,*)$,也称为解析函数.如果 $m(r_i,r_j)$返回真,记录 r_i,r_j 匹配;否则,$m(r_i,r_j)$返回假,记录 r_i,r_j 不匹配.

(3) 分块

实体解析是两两比较的运算,因此计算代价为平方级.当待处理的脏数据集规模较大时,计算代价将是巨大的,并且包含大量的无用计算.分块是 ER 中最常用的减小计算代价的技术[16−26],可以在不影响解析质量的前提下,有效地缩小搜索空间.分块技术将描述可能匹配的记录分到同一块内,将不可能匹配的记录分在不同的块

内.分块通过分块键(blocking key,简称 BK)来实现,而 BK 通过记录属性来构建.当利用一个分块键对数据集进行划分后,拥有相同分块键值(blocking key value,简称 BKV)的记录将进入同一块内.同一块内的任意两条记录称为候选匹配记录对或候选对.

定义 2(时间约束的实体解析). 给定一个脏数据集 R,传统的 ER 处理 R 的时间为 T_{ER},给定时间 $t<<T_{ER}$,时间约束的 ER 将输出尽量多的匹配记录对.给定时间 t 内,TC-ER 比传统 ER 输出更多的匹配记录对,如果运行到自然终止,那么两者的解析结果是相同的.显然,当解析函数的准确率确定时,一个 TC-ER 方法的好坏由时效性和解析的召回率共同决定,可以通过实时召回率来评价.

TC-ER 的流程见算法 1.

算法 1. TC-ER 框架.

输入:脏数据集 R,时间预算 t;

输出:解析结果 Λ.

1. 对 R 进行多路分块,得到块集合 B; //参考定义 3
2. 利用 B 中分块信息来估计候选对的相似性; //本文的核心工作所在
3. **repeat**
4. 根据估计的相似性,对部分候选对进行排序,得到候选列表 *list*;
5. 根据部分排序结果 *list*,依次对候选对进行解析; //包括相似性计算和匹配决定
6. 实时地输出解析结果 Λ;
7. **until** (执行时间>t) or (剩余候选对为 0)

观察算法 1 可知,TC-ER 的核心问题在于优化实体解析的顺序,优先解析匹配可能性大的记录对.如图 1 所示,在分块与相似度计算之间增加记录对排序.记录对排序的依据是记录对的匹配可能性,即估计的记录对相似性,因此,TC-ER 的关键在于如何通过较小的代价准确地估计记录对的相似性.

本文将通过分析和挖掘分块信息来估计记录对的相似性:(1) 第 2 节提出两个基本的记录对相似性估计方法,这两个方法分别从分块质量和记录-块的隶属关系的角度来分析块信息,从而估计记录对相似性;(2) 第 3 节提出基于相似性传播的记录对相似性估计方法,将记录对、块及其关联关系表示为二分图,并通过迭代的传播算法来挖掘分块信息,从而改进基本的相似性估计.

定义 3(多路分块). 给定一个脏数据集 $R=\{r\}$和一组分块键的集合 $BK=\{bk_i|0\leqslant i<K\}$,依次利用 $bk_i\in BK$ 对 R 进行分块,得到有交叠的块集合 $B_{mul}=B^1\cup B^2\cup\ldots\cup B^K=\{b\}$,$B^i$ 表示用分块键 bk_i 生成的无交叠的块集合.任意 $r\in R$ 最多可能出现在 K 个块内,也称为 K 路分块,K 为多路分块的路数.

b 表示块,块中的一对记录 $r_i,r_j\in b$ 称为候选对,记作$\langle r_i,r_j\rangle$;$|b|$表示块 b 的规模,$||b||$表示块 b 的势(cardinality),即块内候选对的数目,记作$||b||=|b|(|b|-1)/2$.$|B|$表示块集合 B 的规模,$||B||$表示块集合 B 的势,即集合内候选对的总数目.B_i 表示记录 r_i 所在块的集合.

2 基本的记录对相似性估计

本节提出两种基本的相似性估计方法(basic estimated similarity,简称 BES),BES 通过直观地分析分块信息来计算记录对的相似性.

2.1 基于块质量的记录对相似性估计

给定一个块,这个块包含的记录对越多,那么这个块内的任意一个记录对的匹配可能性越小.将一个块内记录对的匹配可能性的平均值称为块的冗余性.一个块的信息量是确定的,块内的记录对越多,那么每个记录对平均分得的信息量就越小,块的冗余性就越小.将用冗余性(redudancy,简称 rd)评估块的质量,表示为公式(1):

$$rd(b_i)=1/||b_i|| \tag{1}$$

例如,块 b_1 包括一个记录对,块 b_2 包括 3 个记录对,那么 $rd(b_1)=1>rd(b_2)=1/3$.

给定一个记录对$\langle r_i,r_j\rangle$,将它所在块的冗余性进行聚集来估计相似性,如公式(2),记作 RD-ES,其中,K 是多路

分块的路数,通过 K 来规约,保证相似性落在[0,1]范围:

$$es_{RD}(r_i,r_j)=\frac{1}{K}\sum_{b_k\in B_i\cap B_j} rd(b_k) \tag{2}$$

2.2 基于Jaccard系数的记录对相似性估计

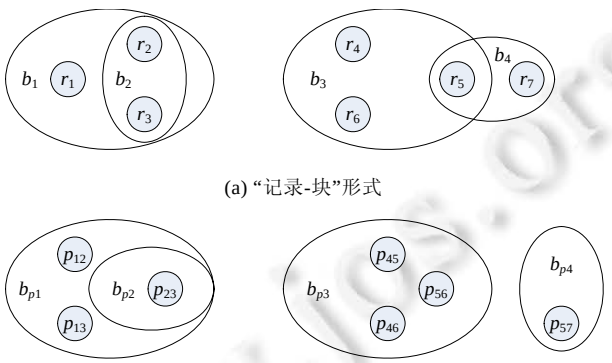
对于一对记录$\langle r_i,r_j\rangle$,如果两者共同出现在越多的块中,那么两者的相似性应该越大;另一方面,如果两者分别出现在越多的不同块中,那么两者的差异性应该越大.通过 Jaccard 系数可以表达上述思想,用一个对记录的共同出现的块的数目除以这对记录各自出现的块的并集的规模,如公式(3),记作 JC-ES:

$$es_{JC}(r_i,r_j)=\frac{|B_i\cap B_j|}{|B_i\cup B_j|} \tag{3}$$

3 基于相似性传播的记录对相似性估计

BES 通过静态地分析分块信息来估计记录对的相似性,没有考虑记录对相似性与块质量的潜在的相互影响.将通过记录对-块之间的相似性传播来改进 BES,称为基于相似性传播的相似性估计(similarity propagation based estimated similarity,简称 SP-ES).显然,SP-ES 是以 RD-ES 或者 JC-ES 为基础的.

例 1:一个脏数据集 $d=\{r_1,r_2,r_3,r_4,r_5,r_6,r_7\}$,经过分块得到块集合 $B=\{b_1,b_2,b_3,b_4\}$,$b_1=\{r_1,r_2,r_3\}$,$b_2=\{r_2,r_3\}$,$b_3=\{r_4,r_5,r_6\}$,$b_4=\{r_5,r_7\}$,如图 2(a)所示;块集合可表示为记录对形式,$B'=\{b_{p1},b_{p2},b_{p3},b_{p4}\}$,$b_{p1}=\{p_{12},p_{13},p_{23}\}$,$b_{p2}=\{p_{23}\}$,$b_{p3}=\{p_{45},p_{46},p_{56}\}$,$b_{p4}=\{p_{57}\}$,如图 2(b)所示.关注两个候选对 p_{12} 和 p_{45},利用两个 BES 方法估计相似性,得到如下结果:(1) RD-ES,$es_{RD}(p_{12})=1/3$,$es_{RD}(p_{45})=1/3$,即 $es_{RD}(p_{12})=es_{RD}(p_{45})$;(2) JC-ES,$es_{JC}(p_{12})=1/2$,$es_{JC}(p_{45})=1/2$,即 $es_{JC}(p_{12})=es_{JC}(p_{45})$.两个 BES 方法都认为 p_{12} 和 p_{45} 的相似性是相等的.然而进一步分析分块情况发现:p_{23} 来自块 b_2 的相似性可以增强块 b_1 的冗余性,进而 p_{12} 从块 b_1 获得更大的相似性;而块 b_3 不存在此类状况.由此可见,应该有 $es(p_{12})>es(p_{45})$.接下来,将通过相似性传播来改进 BES,解决上述问题.



(a) “记录-块”形式

(b) “记录对-块”形式

Fig.2 An example of basic similarity estimations' disadvantage

图 2 基本的记录对相似性估计缺陷示例

3.1 相似性传播的基本思想

本文中,相似性传播的基本思想为:记录对的相似性可以促进所在块的冗余性,块的冗余性可以促进块内记录对的相似性.为了充分地表示记录对与块之间的关联关系,将分块结果表示为“记录对-块”二分图(严格描述见定义 4).相似性传播是一种基于图结构的相似性计算方法[29,30],充分地挖隐藏信息,可以弥补其他相似性的不足.将通过例 2 来直观地了解记录对与块之间的相似性传播.

定义 4(**“记录对-块”二分图,简称“对-块图”**). 给定一个记录对形式的块集合 B,B 对应的候选对集合为 P,构

建一个无向的二分图 $G=(V_p,V_b,E)$,其中,$V_p=\{v_{pi}|0\leqslant i\leqslant m\}$为记录对结点集合,每个记录对结点对应一个记录对;$V_b=\{v_{bj}|0\leqslant j\leqslant n\}$为块结点集合,每个块结点对应一个块;$E\subset V_p\times V_b$ 是边集合,表示记录对与块的隶属关系.可以结合上方向来解读边的含义,给定一条边,从块结点到记录对结点,表示“包含”关系;而从记录对结点到块结点,则表示“出现在”关系.本文为了便于表达,将用 P 同时表示记录对结点集合和记录对集合,用 p 同时表示记录对结点和记录对,用 B 同时表示块结点集合和块集合,用 b 同时表示块结点和块.

例 2:续例 1.将例 1 中 B'表示为二分图 G,如图 3 所示.初始时,用 RD-ES 来估计记录对相似性(es),所有块冗余性(rd)初始化为 0,得到表 1 中第 1 行的初始值.接下来,相似性通过二分图传播.

(1) $P\rightarrow B$ 传播.以图 3(a)为例,记录对结点 p_{12} 和 p_{13} 分别传递各自当前的相似性(都是 1/3)给邻接的块结点 b_{p1},p_{23} 将当前的相似性 4/3 分别传递给邻接的块结点 b_{p1} 和 b_{p2};b_{p1} 获得 3 个相似性(分别为 1/3,1/3,4/3),取均值为 2/3 作为更新的冗余性,同理,b_{p2} 更新的冗余性为 4/3,对 G 所有的连通分量进行相同的操作后,得到表 1 中第 2 行的结果;

(2) $B\rightarrow P$ 传播.以图 3(a)为例,块结点 b_{p1} 将当前的冗余性 2/3 传给邻接的记录对结点 p_{12},p_{13} 和 p_{23},b_{p2} 将当前的冗余性 4/3 传给邻接的结点 p_{23};记录对结点 p_{12} 和 p_{13} 分别获得一个冗余性 2/3,分别作为各自最新的相似性,同理,p_{23} 获得两个冗余性 2/3 和 4/3,求和得到 2 作为 p_{23} 最新的相似性(此处为简单的计算方式,后续将给出严格的计算方式),对 G 所有的连通分量进行相同的操作后,得到表 1 中第 3 行的结果.

经过一次 $P\rightarrow B\rightarrow P$ 传播后,$es(p_{12})=2/3$,$es(p_{45})=1/3$,那么 $es(p_{12})>es(p_{45})$,符合例 1 中提出的预期.可见,相似性传播有助于准确地估计记录对的相似性,弥补基本的相似性估计方法的不足.

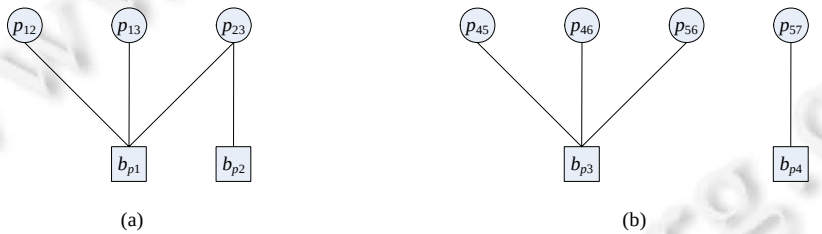


Fig.3 An example of pair-block barpitite graph

图 3 “记录对-块”二分图示例

Table 1 Example of similairity propagation on bipartite graphs

表 1 二分图上相似性传播示例

	$rd(b_{p1})$	$rd(b_{p2})$	$rd(b_{p3})$	$rd(b_{p4})$	$es(p_{12})$	$es(p_{13})$	$es(p_{23})$	$es(p_{45})$	$es(p_{46})$	$es(p_{56})$	$es(p_{57})$
初始化	0	0	0	0	1/3	1/3	4/3	1/3	1/3	1/3	1
$P\rightarrow B$	**2/3**	**4/3**	**1**	**1/3**	1/3	1/3	4/3	1/3	1/3	1/3	1
$B\rightarrow P$	2/3	4/3	1	1/3	**2/3**	**2/3**	**2**	**1/3**	**1/3**	**1/3**	**1**

3.2 基于对-块图的相似性传播

接下来,严格地定义对-块图上相似性传播.首先定义两个单步传播:从记录对结点到块结点($P\rightarrow B$)的传播和从块结点到记录对结点($B\rightarrow P$)的传播.然后,在单步传播的基础上定义迭代的相似性传播.

定义 5($P\rightarrow B$ 传播). 给定一个对-块图 $G=(P,B,E)$,$p_{ij}\in P$,$b_k\in B$,p_{ij} 与 b_k 是邻接结点,p_{ij} 将当前的相似性传递给每一个邻接的块结点,b_k 从每个邻接记录对结点接受相似性,并求平均值作为最新的冗余性,如公式(4)所示:

$$rd(b_k)=\frac{1}{|N(b_k)|}\sum_{p_{ij}\in N(b_k)}es(p_{ij}) \tag{4}$$

其中,$N(b_k)$表示 b_k 的相邻结点的集合.

一次 $P\rightarrow B$ 传播后,所有块结点的冗余性将更新,而记录对结点的相似性不变.

定义 6(*B*→*P* 传播). 给定一个对-块图 $G=(P,B,E)$,$p_{ij}\in P$,$b_k\in B$,p_{ij} 与 b_k 是邻接结点,b_k 将当前的冗余性传递给每一个邻接的记录对结点,p_{ij} 从每个邻接块结点接受冗余性,求和后除以 K 作为最新的冗余性,如公式(5)所示:

$$es(p_{ij})=\frac{1}{K}\sum_{b_k\in N(p_{ij})}rd(b_k) \tag{5}$$

一次 $B\to P$ 传播后,所有记录对结点的相似性将更新,而块结点的冗余性不变.

参数 K 是多路分块的路数,公式(5)中除以 K 是为了将 es 值规约到[0,1]范围内,一个记录对最多可能出现在 K 个块内.es 的规约对于后续迭代计算的收敛性非常重要.

定义 7(*P*↔*B* 传播). 给定一个对-块图 $G=(P,B,E)$,初始的记录对相似性通过 BES 得到,初始的块冗余性为 0.经过一次 $P\to B$ 传播,更新块冗余性;再经过一次 $B\to P$ 传播,更新记录对相似性.如此不断迭代,直到记录对相似性不再发生变化.容易知道,$P\leftrightarrow B$ 传播是不可约、非周期、有限状态的马尔可夫链,因此必定收敛于平稳分布[29].

例 3:对图 3(a)进行一次 $P\to B\to P$ 传播,其中多路分块的路数 $K\geqslant 2$.初始时,各记录对相似性记作 $es_0(p_{12})$,$es_0(p_{13})$和 $es_0(p_{23})$.

(1) $P\to B$ 传播.根据定义 5,计算得到更新的块冗余性:

$$\left.\begin{aligned} rd_1(b_{p1})&=es_0(p_{12})/3+es_0(p_{13})/3+es_0(p_{23})/3,\\ rd_1(b_{p2})&=es_0(p_{23}) \end{aligned}\right\} \tag{6}$$

(2) $B\to P$ 传播.根据定义 6,利用公式(6)的计算结果,计算得到更新的记录对相似性:

$$\left.\begin{aligned} es_1(p_{12})&=es_0(p_{12})/3K+es_0(p_{13})/3K+es_0(p_{23})/3K,\\ es_1(p_{13})&=es_0(p_{12})/3K+es_0(p_{13})/3K+es_0(p_{23})/3K,\\ es_1(p_{23})&=es_0(p_{12})/3K+es_0(p_{13})/3K+4es_0(p_{23})/3K \end{aligned}\right\} \tag{7}$$

可以发现,等式组(7)是记录相似性的递推关系,它隐藏了块结点,将 $P\to B\to P$ 传播转化为 $P\to P$ 传播.

3.3 不动点计算

$P\leftrightarrow B$ 传播的收敛结果计算可以作为一个不动点计算(fixpoint computation)的问题.本文主要关注记录对相似性,因此直接使用记录对相似性的递推关系.将等式组(7)改写成向量与矩阵的运算形式,如公式(8):

$$\begin{pmatrix} es_1(p_{12})\\ es_1(p_{13})\\ es_1(p_{23}) \end{pmatrix}^T=\begin{pmatrix} es_0(p_{12})\\ es_0(p_{13})\\ es_0(p_{23}) \end{pmatrix}^T\begin{pmatrix} 1/3K & 1/3K & 1/3K\\ 1/3K & 1/3K & 1/3K\\ 1/3K & 1/3K & 4/3K \end{pmatrix} \tag{8}$$

其中,向量为行向量.公式(8)中的 3×3 矩阵是一个马尔可夫链的转移矩阵,记作 $\boldsymbol{Q_0}=\{q_{ij}|0\leqslant i,j<3\}$,$q_{ij}$ 表示从状态 i 转移到状态 j 的概率,Q_0 不要求为对称阵.特别地,K 的存在确保 $0\leqslant q_{ij}\leqslant 1$,并进一步保证 $P\leftrightarrow B$ 传播的收敛性(例 3 中 $K\geqslant 2$);如果去掉 K,则 Q_0 不再是转移矩阵,$P\leftrightarrow B$ 传播不一定收敛.计算 $P\leftrightarrow B$ 传播的关键是转移矩阵,接下来讨论如何计算转移矩阵.

定义 8(邻接矩阵). 给定一个对-块图 $G=(P,B,E)$,$P=\{p_i|0\leqslant i<m\}$,$B=\{b_j|0\leqslant j<n\}$,$E\subset P\times B$,G 的对→块邻接矩阵为公式组(9)中的 $\boldsymbol{A_{mn}}$,块→对邻接矩阵则为 $\boldsymbol{A_{nm}}=(\boldsymbol{A_{mn}})^T$.

$$\begin{aligned} \boldsymbol{A_{mn}}&=\{a_{ij}\mid 0\leqslant i<m,0\leqslant j<n\}\\ a_{ij}&=\begin{cases}1, & e_{ij}\in E\\ 0, & e_{ij}\notin E\end{cases} \end{aligned} \tag{9}$$

定义 9(转移矩阵). 给定一个对-块图 G 和对→块邻接矩阵 $\boldsymbol{A_{mn}}$,那么对→块转移矩阵 $\boldsymbol{Q_{mn}}$ 为

$$\boldsymbol{Q_{mn}}=\{q_{ij}\mid q_{ij}=a_{ij}/\textstyle\sum a_{i*},a_{ij}\in\boldsymbol{A_{mn}}\} \tag{10}$$

同理,可以根据块→对邻接矩阵 $\boldsymbol{A_{nm}}$ 计算得到块→对转移矩阵 $\boldsymbol{Q_{nm}}$.

用矩阵来表示 $P\to B\to P$ 传播,当前记录对的相似性向量为 $\boldsymbol{es_x}$,块的冗余性向量为 $\boldsymbol{rd_x}$.那么:

(1) $P\to B$ 传播.块冗余性更新:

$$\boldsymbol{rd}_{x+1}^T=\boldsymbol{es}_x^T\boldsymbol{Q_{mn}} \tag{11}$$

(2) $B\to P$ 传播.记录对相似性更新:

$$\boldsymbol{es}_{x+1}^{T}=\boldsymbol{rd}_{x+1}^{T}\boldsymbol{Q}_{nm}=\boldsymbol{es}_{x}^{T}\boldsymbol{Q}_{mn}\boldsymbol{Q}_{nm} \tag{12}$$

根据公式(12)易知,$P\to B\to P$ 传播的整体转移矩阵为 $\boldsymbol{Q}_{mm}$:

$$\boldsymbol{Q}_{mm}=\boldsymbol{Q}_{mn}\boldsymbol{Q}_{nm} \tag{13}$$

对于记录对相似性,$P\to P$ 传播与 $P\to B\to P$ 传播等价,进而 $P\leftrightarrow P$ 传播与 $P\leftrightarrow B$ 传播等价.

定义 10($P\leftrightarrow P$ 传播). 给定一个对-块图 G 和 $P\to P$ 转移矩阵 $\boldsymbol{Q}_{mm}$,与 $P\leftrightarrow B$ 传播等价的 $P\leftrightarrow P$ 传播中,一次 $P\to P$ 传播可表示为

$$\boldsymbol{es}_{x+1}^{T}=\boldsymbol{es}_{x}^{T}\boldsymbol{Q}_{mm} \tag{14}$$

不动点计算的流程.

(1) 利用 BES 初始化记录对相似性向量为 $\boldsymbol{es}_0$;

(2) 利用公式(14)计算下一轮记录对相似性向量 $\boldsymbol{es}_{x+1}$;

(3) 不断迭代步骤(2),直到残差向量$\Delta(\boldsymbol{es}_{x+1},\boldsymbol{es}_x)$的每一维都小于$\varepsilon$.$\varepsilon$为残差阈值,一般取一个较小常数,例如0.005;

(4) 当迭代结束后,将最终的记录对相似性向量记作 $\boldsymbol{es}_{sp}$,作为基于相似性传播的相似性估计的结果输出.

代价分析:$P\leftrightarrow P$ 传播的一次迭代代价为 $O(m^2)$,迭代次数为 X,那么总代价为 $O(Xm^2)$.对-块图 G 实际是由多个连通分量组成,分量集合记作 $C=\{c_i\}$,$|c_i|<<m$,$P\leftrightarrow P$ 传播只发生在每个连通分量内部,那么实际的代价:

$$O(\sum X_i\,|c_i|^2)<<O(Xm^2).$$

3.4 近似的相似性传播

$P\leftrightarrow P$ 传播的不动点计算的代价较大,对于时间约束的 ER 来说不可接受.为此,希望通过近似方法降低计算代价.下面将通过分析迭代过程中相似性传播的情况,来说明较少的迭代次数可以近似地计算出相似性,并极大地减小计算代价.

例 4:图 4 呈现了一个较大的对-块图,观察记录对 p_3,其初始相似性为 $es_0(p_3)$,第 1 次 $P\to P$ 传播后,p_3 的相似性传递到 p_3 所在块的其他记录对 p_1,p_2 和 p_4,p_5,分别得到 $es_0(p_3)/6$;第 2 次 $P\to P$ 传播后,p_3 的相似性传递到块 b_3 中的记录对 p_6,p_7,分别得到 $es_0(p_3)/36$;第 3 次 $P\to P$ 传播后,p_3 的相似性传递到块 b_4 中的记录对 p_8,p_9,分别得到 $es_0(p_3)/216$.可以发现:随着迭代次数的增加,相似性传递得越来越远,并且传递量发生指数级的衰减,而计算代价成倍增长.从物理意义的角度分析,p_3 与同一块内的记录对关联最强,而与间接关联的记录对的关联强度则随距离增加而极大地减弱.由此得到启发,用较少的 $P\leftrightarrow P$ 传播的迭代来代替不动点计算,即迭代次数 X 取较小数值,如 1 或 2.这样可以近似地计算出记录对相似性,只损失微小的准确性,但可以降低计算代价,从而保证时间约束的 ER 的实时召回率.将通过实验验证这种近似方法的有效性.

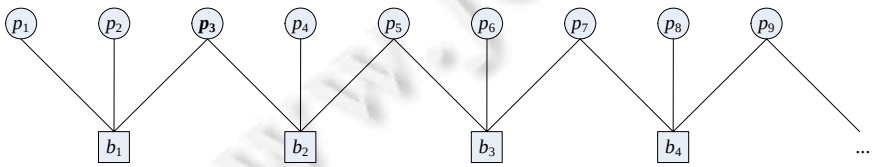


Fig.4 $P\leftrightarrow P$ propagation analysis

图 4 $P\leftrightarrow P$ 传播分析

4 实验评价

4.1 实验设置

实验代码通过 Java 实现,Java 版本为 1.7.运行环境如下:处理器 3.4GHz Intel(R) Core i7-2600,内存 8GB,操

作系统为微软 Windows 10 专业版(64 位).

- 数据集.

实验评价使用两个数据集:一个真实数据集和一个合成数据集.真实的引文数据集 DBLP-Scholar(记作 DBLP)包含 66 879 条引文记录,其中有 5 347 对匹配记录,通过标题、作者、期刊/会议、年份等 4 个属性描述[3].利用 Febrl 数据生成器构建一个合成的个人信息数据集(记作 FB),包含 150K 条个人记录,其中有 81 694 对匹配记录,通过姓名、性别、生日、住址、城市、州和邮编等 8 个属性描述[32,33].

- 评价指标.

本文的研究目标是优化实体解析顺序,认为解析函数已提前确定,准确率与研究目标是正交关系,因此采用实时召回率来评估方法.在第 4.3 节的部分对比测试中,还采用 Top-N 命中率来评价,将在后续详细介绍.

- 解析函数.

本文采用 SVM 来训练分类器作为解析函数 $m(*,*)$[33].

- 方法设置.

对 DBLP-Scholar 数据集进行四路分块,4 个分块键为标题的前 3 个实词、姓+名的前两个字母、期刊/会议的前 3 个实词和年份.对 FB 数据集进行四路分块,4 个分块键为姓+名的前两个字母、生日、城市和邮编.渐进式滑动窗口(progressive sorted neighborhood method,简称 PSNM)方法、渐进式分块(progressive blocking,简称 PB)方法和以记录对排序列表(sorted list of record pairs,简称 SLORP)为线索的方法的排序键[6,7]:DBLP-Scholar 数据集上采用标题的前 3 个实词+前两个作者的姓;FB 数据集上采用姓+名+城市.如果没有特别说明,SP-ES 的迭代次数设置为 1.TC-ER 默认采用基于 RD-ES 的 SP-ES 来估计记录对相似性,记作 TC-ER0;将采用基于 JC-ES 的 SP-ES 的 TC-ER 记作 TC-ER1;将采用 RD-ES 的 TC-ER 记作 TC-ER2;将采用 JC-ES 的 TC-ER 记作 TC-ER3.

4.2 综合测试

综合测试将 TC-ER0 与一个基准方法以及 3 个已有工作进行对比.Papenbrock 等人提出的基于排序列表的 PSNM 和 PB 已经被证明优于 Whang 等人提出的基于“线索”的方法,而基于“线索”的方法中表现最好的为 SLORP[6,7],因此选择 PSNM,PB 和 SLORP 这 3 个方法作为比较对象.基准方法采用 $m(*,*)$直接解析分块后生成的候选对,记作 Baseline.将随机地生成 10 个候选对顺序,分别执行 Baseline 方法,将 10 次的结果取平均值作为 Baseline 的结果.PSNM 将从小到大扩展的窗口多次滑过排序的记录列表,PB 先根据排序的记录列表生成同等规模的小块,然后逐渐拓展分块范围,这两个方法都渐进地生成候选对,从而优先处理匹配可能性大的记录对[7].SLORP 通过排序的记录列表一次性生成排序的记录对列表,并根据记录对顺序来依次进行解析[6].PSNM,PB 和 SLORP 都要依赖于排序的记录列表,其排序键请参考第 4.1 节中的方法设置.

图 5 和图 6 分别呈现了在 DBLP 数据集上和 FB 数据集上 5 个方法的对比情况.

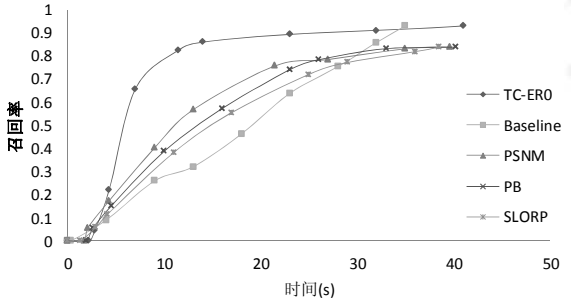


Fig.5 General test on DBLP dataset

图 5 在 DBLP 数据集上的综合测试

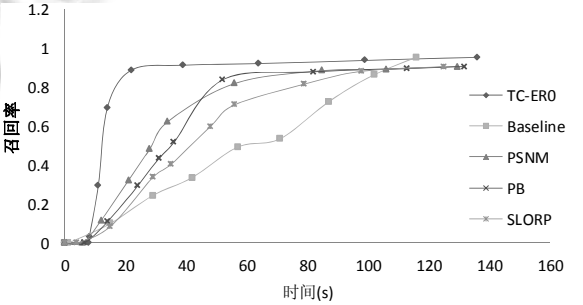


Fig.6 General test on FB dataset

图 6 在 FB 数据集上的综合测试

总体而言,存在 TC-ER0>> PSNM,PB>SLORP>Baseline.TC-ER0 的实时召回率显著地高于其他 4 个方法;

PSNM,PB 和 SLORP 明显优于 Baseline;PSNM,PB 明显优于 SLORP;在前期时,PSNM 总优于 PB,PB 在后期可能有机会超越 PSNM(如在 FB 数据集上).例如:在 DBLP 数据集上,TC-ER0 花费 11.5s 解析出 82.47%的匹配对,PSNM 花费 13s 只能解析出 56.69%的匹配对,PB 花费 16s 只能解析出 57.18%的匹配对,SLORP 花费 17s 只能解析出 55.41%的匹配对,Baseline 花费 18s 甚至只能解析出 46.19%的匹配对;在 FB 数据集上,TC-ER0 花费 22s 解析出 88.9%的匹配对,而 PSNM 花费 34s 解析出 62.47%的匹配对,PB 花费 36s 解析出 51.72%的匹配对,SLORP 花费 35s 解析出 40.41%的匹配对, Baseline 花费 42s 只能解析出 33.67%的匹配对.由此可见,在较少时间预算约束下,TC-ER0 可以解析出更多的匹配对.Baseline 的实时召回率随时间线性增长,因为它随机地解析候选对,解析顺序没有任何优化.PSNM 在迭代中由小到大调整窗口,以此来优化候选对的解析顺序;PB 通过逐渐拓展分块范围来优化解析顺序;SLORP 通过粗糙的方法来估计候选对的相似性来优化解析顺序.因此,它们的实时召回率要比 Baseline 高.然而,PSNM 和 PB 无法将候选对按匹配可能性排序,无法直接定位到最可能匹配的候选对;SLORP 虽然对候选进行了全排序,但其相似性估计十分粗糙,同样无法直接定位到最可能匹配的候选对,甚至不如前两者的表现.这些原因局限了 PSNM,PB 和 SLORP 的实时召回率.TC-ER0 则通过基于相似性估计的候选对排序来全局地优化解析顺序,从而获得最高的实时召回率.

再者,观察最终召回率和运行时间.

(1) PSNM,PB 和 SLORP 的最终召回率要低于 TC-ER0 和 Baseline.前三者只有一个排序键,只产生一个记录排序列表,由此生成的候选对集合对真实的匹配对覆盖较少;而后两个方法,通过多路分块生成候选对集合,可以更好地覆盖真实的匹配对.如果将 PSNM 和 PB 扩展到多个排序键 AC-PSNM 和 AC-PB,可以提高最终召回率,但这样要维护多个排序列表并依次滑动,会大大增加实时的计算代价和总的计算代价,明显降低实时召回率;

(2) 就总的运行时间而言,TC-ER0,PSNM,PB 和 SLORP 都比 Baseline 要长,因为前四者针对解析顺序的预处理操作花费了一定的时间,而 Baseline 没有预处理操作.就预处理时间而言,TC-ER0 要比 PSNM,PB 和 SLORP 稍长一些,但它们的预处理时间占总运行时间的比例都极小.

4.3 分项测试

4.3.1 相似性估计测试

相似性估计测试将比较两个基本的相似性估计方法 RD-ES,JC-ES 和基于前两者的 SP-ES 方法在 TC-ER 中的表现,对应该 ER 方法分别为 TC-ER2(RD-ES),TC-ER3(JC-ES),TC-ER0(基于 RD-ES 的 SP-ES)和 TC-ER1(基于 JC-ES 的 SP-ES).相似性估计方法的好坏将决定记录对排序的有效性,进而影响实时召回率.图 7 和图 8 分别是在 DBLP 数据集上和 FB 数据集上 4 个方法的对比情况,Baseline 作为参考.

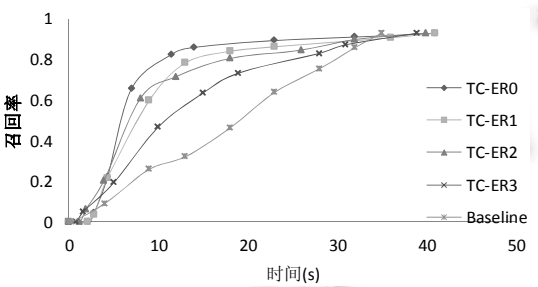


Fig.7 Similarity estimation test on DBLP dataset

图 7 在 DBLP 数据集上的相似性估计测试

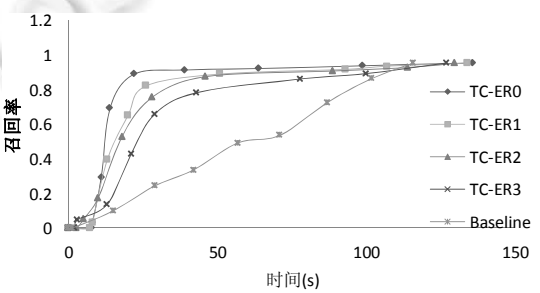


Fig.8 Similarity estimation test on FB dataset

图 8 在 FB 数据集上的相似性估计测试

整体而言,就实时召回率而言,在两个数据集上均有 TC-ER0>TC-ER1,TC-ER2>TC-ER3>Baseline.在 DBLP 数据集上,TC-ER0 显著地优于其他 3 个方法;TC-ER1 与 TC-ER2 不相上下,但都明显优于 TC-ER3.取单点进行对比,TC-ER0 花费 11.5s 解析出 82.47%的匹配对,TC-ER1 花费 13s 解析出 78.18%的匹配对,TC-ER2 花费 12s

解析出 71.54%的匹配对,TC-ER3 花费 15s 只解析出 63.41%的匹配对.在 FB 数据集上,TC-ER0 显著地优于其他 3 种方法;TC-ER1 微弱地优于 TC-ER2,但两者都明显优于 TC-ER3.取单点进行对比,TC-ER0 花费 22s 解析出 88.9%的匹配对,TC-ER1 花费 26s 解析出 82.14%的匹配对,TC-ER2 花费 28s 解析出 75.38%的匹配对,TC-ER3 花费 29s 只解析出 65.94%的匹配对.接下来,观察相似性传播对相似性估计的影响.分别对比 TC-ER0 和 TC-ER2,TC-ER1 和 TC-ER3 可以发现:TC-ER 中,基于相似性传播的相似性估计方法(TC-ER0 和 TC-ER1)要明显优于基本的相似性估计方法(TC-ER2 和 TC-ER3).相似性传播挖掘了记录与块之间隐藏的关联关系,从而有效地改进了两个基本的相似性估计方法.最后,分别对比 TC-ER0 和 TC-ER1,TC-ER2 和 TC-ER3 可以发现:在 TR-ER 中,基于分块质量的相似性估计(对应 TC-ER0 和 TC-ER2)要明显优于基于 Jaccard 系数的相似性估计(对应 TC-ER1 和 TC-ER3),说明分块质量可以更有效地帮助估计记录对的相似性.

4.3.2 SP-ES 的迭代测试

本节测试 TC-ER 中 SP-ES 的迭代次数对相似性估计的影响,验证近似的相似性传播的有效性.当残差阈值 ε=0.005 时,TC-ER0 在 DBLP 数据集上和 FB 数据集上分别需要 7 次和 11 次迭代达到收敛.将从两个角度来进行测试:(1) 在两个数据集上,测试 TC-ER0 分别进行 1,2,4 和 7 次迭代的实时召回率的情况,从直观上了解随着迭代次数的增加,相似性估计和运行时间的变化情况;(2) 在两个数据集上,测试 TC-ER0 随着迭代次数的增加直到自然收敛过程中,Top-N 命中率和启动时间的变化情况.

TC-ER0 默认进行 1 次迭代,现将 TC-ER0 分别进行 2m4 和 7 次迭代,分别记作 TC-ER0-2,TC-ER0-4 和 TC-ER0-7.图 9 和图 10 分别展示了在 DBLP 数据集上和 FB 数据集上这 4 种方法的对比情况,Baseline 作为参考.

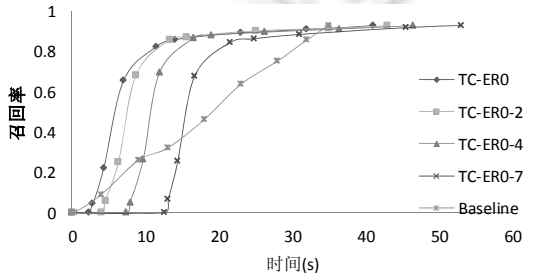


Fig.9 SP-ES's iteration test on DBLP dataset

图 9 在 DBLP 数据集上 SP-ES 的迭代测试

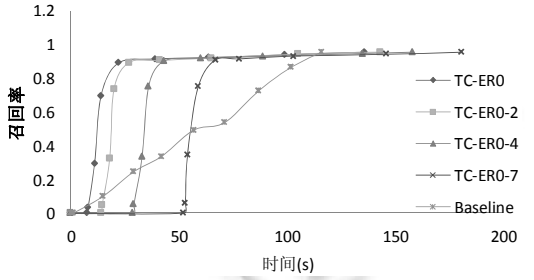


Fig.10 SP-ES's iteration test on FB dataset

图 10 在 FB 数据集上 SP-ES 的迭代测试

整体而言,随着迭代次数的增加,预处理的时间代价接近成倍地增加;然而预处理之后的实时召回率却没有明显的提升,即相似性估计的准确性只是非常微弱地提高.取单点来对比,在 DBLP 数据集上,运行时间为 7s 时,TC-ER0 的实时召回率为 65.63%,而 TC-ER0-2 约为 40%,TC-ER0-4 和 TC-ER0-7 都为 0;在 FB 数据集上,运行时间为 18s 时,TC-ER0 的实时召回率超过 70%,而 TC-ER0-2 为 32.47%,TC-ER0-4 和 TC-ER0-7 都为 0.直观来看,基于 1 次迭代 SP-ES 对于 TC-ER 来说是最佳的选择.

接下来,从 Top-N 命中率和启动时间的角度来分析迭代的效果,Top-N 命中率是指在观测的前 N 次比较中匹配对占的比例,启动时间是指 TC-ER0 从启动运行到开始产生解析结果的时间间隔.将数据集中真实的匹配对数目设为 N,那么,在 DBLP 数据集上 N=5347,在 FB 数据集上 N=81694.图 11、图 12 展示了 TC-ER0 方法在两个数据集上的 Top-N 命中率(对应主轴刻度)和启动时间(对应次轴刻度)随迭代次数增加而变化的情况.可以发现:随着迭代次数的增加,命中率的提高非常小,而启动时间则几乎是线性增长.由此可知,迭代次数的增加不会明显提高相似性估计的准确性,而启动时间大幅增长.结合图 9、图 10 可知:随着迭代次数的增加,实时召回率曲线的趋势变化微弱,大体上是整体向后平移,启动时间占总运行时间比重大幅增加.这导致解析结果输出推迟,影响 TC-ER0 的表现.综上分析,为了兼顾时效性和召回率,应当选择较少的迭代次数,1 次迭代是 TC-ER0 的最好选择.

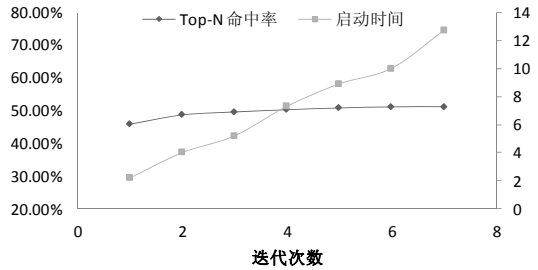


Fig.11 SP-ES's hitting rate & start-up time tests on DBLP dataset

图 11 在 DBLP 数据集上 SP-ES 的 Top-*N* 命中率及启动时间测试

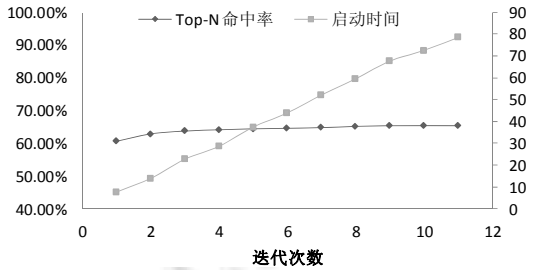


Fig.12 SP-ES's hitting rate & start-up time tests on FB dataset

图 12 在 FB 数据集上 SP-ES 的 Top-*N* 命中率及启动时间测试

5 相关工作

实体解析是数据集成与数据清洗不可或缺的组成部分,也称为实体识别、实体匹配、记录链接等[1−15].传统的实体解析是批处理操作,将整个数据集输入,经过分块、相似性计算和匹配决定后,输出解析结果[1,2].这种整体解析的运行时间通常比较长.随着大数据产业的发展,数据产生的速度和更新的频率与以往相比都有了质的飞跃,而一些数据应用要求(近似)实时的响应,因此时间约束的 ER 成为研究热点[6,7].与本文相关的研究还包括分块技术和基于图的相似性传播.

Whang 等人提出了 Pay-as-you-go 实体解析的概念:在运行时间或计算资源有限的情况下,使得实体解析的输出结果最大化;并定义了"线索"的概念,帮助预测哪些记录的匹配可能性更大,它需要与已有的 ER 方法结合起来使用[6].Papenbrock 等人提出了一组时间约束的 ER 方法,它们都基于排序的记录列表[7].渐进式滑动窗口(progressive sorted neighborhood method,简称 PSNM)通过从小到大扩大窗口来多次滑过列表,渐进地生成候选对;渐进式分块(progressive blocking,简称 PB)先根据记录列表生成同样规模的小块,然后逐渐拓展分块范围,渐进地生成候选对.在此基础上,还对两个方法进行了多属性扩展,同时生成多个排序列表,并交替地对排序列表执行 PSNM 和 PB,从而提高总的召回率,但同时也降低了渐进性.这两类方法都无法将所有候选对按匹配可能性进行全局排序,限制了实时召回率;再者,两类方法都依赖于已知的分块键或排序键,限制了适用范围.

分块是 ER 中最常用的降低时间开销的技术,它可以有效地缩小搜索空间[16−26].分块方法可分为两类:基于分块键的方法和基于排序键的方法.前者定义分块键(blocking key,简称 BK),然后根据每条记录的属性信息生成对应的分块键值(blocking key value,简称 BKV),最后将拥有相同 BKV 的记录分在同一块内,分块方法以此类居多[17−20,23−26].后者也称为滑动窗口方法,首先定义排序键,然后将记录按排序键值排序,最后将一个窗口在记录列表上滑动来生成候选对[21,22].

基于图的相似性传播可以挖掘结构信息来计算数据对象(data object)之间的相似性,这类方法已经应用在了多个领域,如模式匹配[29]、联合式实体解析[4]、推荐系统[30]等.Melnick 等人设计了 SF(similarity flooding)算法来帮助模式匹配,但其应用范围不局限于此[29].将两个关系模式分别构建成模式图,并根据领域知识计算出两个图之间结点的初步的相似性,将这两个图作为 SF 算法的输入.SF 将两个图中的结点建立映射关系,并构建成一个成对的关联图,图中的每个结点对应原模式图中一个映射结点对,例如关联图中的三元组$((x,y),p,(x',y'))$,对应模式图中的两个三元组(x,p,x')和(y,p,y').相似性通过$((x,y),p,(x',y'))$的正向和反向不断迭代地传播,迭代停止时每个结点上的对象对(例如(x,y))获得最终的相似性.利用 SF 最终的相似性可以决定模式匹配的结果.SF 通过不动点计算来获得最终的相似性.如果相似性收敛,那么 SF 自然终止;如果不收敛,则运行到 SF 设定的最大迭代次数时终止.Simrank 是一个更通用的两两(pairwise)相似性计算方法,其基本思想是:如果与两个对象关联的对象是相似的,那么这两个对象也是相似的[30].Simrank 将一个有向的对象图转换成一个有向的对象对图,对象对图与 SF 中的成对关联图类似,也是$((x,y),p,(x',y'))$的形式.在对象对图中,初始时将由同一对象组成的结点的相似

性设为 1,其他结点的相似性为 0.然后相似性在对象对图中沿着有向边不断传递,直到收敛.在传递过程中,一个结点(x,y)将相似性经过衰减少后传给它所有指向的结点;另一个结点(x',y')从指向它的所有结点处获得相似性,取均值作为自己最新的相似性.Simrank 保证收敛性,可以通过不动点计算获得收敛结果.Dong 等人通过相似性传播来解析关联的数据,例如引文数据、电影数据等[4].以引文数据为例,文章、作者及会议之间存在语义关联,如果两个文章记录是匹配的,那么它们的作者记录的匹配可能性将会增加.将关联的数据构建依赖图,其中,边既有单向的,也有双向的.根据文本相似性来计算每个结点的初始相似性,然后,相似性通过有向边来传播.当某个结点的相似性超过阈值,就认为它对应的记录是匹配的,匹配的结点将进入非激活状态.不断迭代,直到所有结点都被解析完.以上 3 种相似性传播都是在对象(记录)之间的传播,而本文的相似性传播则是在记录与块之间进行.

当前,还出现了一些新型的 ER 方法:Ramadan 等人提出了面向查询的 ER 方法[5];Kushagrat 等人针对 ER 中聚类问题,提出了选择策略[8];Lin 等人提出了面向异构记录的 ER 方法[9];多个研究团队提出了基于图的 ER 方法[10−12];多个研究团队提出了基于深度学习的 ER 方法[13,14].

6 结束语

时间约束的实体解析是大数据研究的热点问题,本文研究时间约束的 ER 中记录对相似性估计与排序.在多路分块的基础上,分析块内信息,提出了基于块质量的记录对相似性估计方法;分析块间信息,提出了基于 Jaccard 系数的记录对相似性估计方法.针对两个基本的相似性估计方法,提出了基于相似性图传播的改进.构建记录对和块组成的二分图.在二分图上运行相似性传播,在此过程中记录对的相似性动态变化,直到收敛.提出了基于不动点迭代的收敛结果计算方法,并提出了近似方法来降低计算代价.在一个真实数据集和一个合成数据集上测试提出的方法,证明其有效性,并测试了提出方法的各个方面的特点.在未来的工作中,将研究联合式实体解析在时间约束条件下的解决方案.

References:

[1] Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: A survey. IEEE Trans. on Knowledge and Data Engineering, 2007,19(1):1−16. [doi: 10.1109/TKDE.2007.250581]

[2] Hassanzadeh O, Chiang F, Lee HC, Miller RJ. Framework for evaluating clustering algorithms in duplicate detection. Proc. of the VLDB Endowment, 2009,2(1):1282−1293. [doi: 10.14778/1687627.1687771]

[3] Köpcke H, Thor A, Rahm E. Evaluation of entity resolution approaches on real-world match problems. Proc. of the VLDB Endowment, 2010,3(1-2):484−493. [doi: 10.14778/1920841.1920904]

[4] Dong X, Halevy A, Madhavan J. Reference reconciliation in complex information spaces. In: Proc. of the 2005 ACM SIGMOD Int'l Conf. on Management of Data. ACM, 2005. 85−96.

[5] Ramadan B, Christen P, Liang H, Gayler RW. Dynamic sorted neighborhood indexing for real-time entity resolution. Journal of Data and Information Quality (JDIQ), 2015,6(4):15.

[6] Whang SE, Marmaros D, Garcia-Molina H. Pay-as-You-Go entity resolution. IEEE Trans. on Knowledge and Data Engineering, 2013,25(5):1111−1124.

[7] Papenbrock T, Heise A, Naumann F. Progressive duplicate detection. IEEE Trans. on Knowledge and Data Engineering, 2015,27(5): 1316−1329.

[8] Kushagra S, Saxena H, Ilyas IF, Ben-David S. A semi-supervised framework of clustering selection for de-duplication. In: Proc. of the 35th Int'l Conf. on Data Engineering (ICDE). IEEE, 2019. 208−219.

[9] Nie H, Han X, He B, Sun L, Chen B, Zhang W, Wu S, Kong H. Deep sequence-to-sequence entity matching for heterogeneous entity resolution. In: Proc. of the 28th ACM Int'l Conf. on Information and Knowledge Management. ACM, 2019. 629−638.

[10] Tauer G, Date K, Nagi R, Suditc M. An incremental graph-partitioning algorithm for entity resolution. Information Fusion, 2019,46: 171−183.

[11] Kwashie S, Liu L, Liu J, Liu L, Stumptner M, Yang L. Certus: An effective entity resolution approach with graph differential dependencies (GDDs). Proc. of the VLDB Endowment, 2019,12(6):653−666.

[12] Reas R, Ash S, Barton R, Borthwick A. Superpart: Supervised graph partitioning for record linkage. In: Proc. of the 2018 IEEE Int'l Conf. on Data Mining (ICDM). IEEE, 2018. 387−396.

[13] Ebraheem M, Thirumuruganathan S, Joty S, Ouzzani M, Tang N. Distributed representations of tuples for entity resolution. Proc. of the VLDB Endowment, 2018,11(11):1454−1467.

[14] Mudgal S, Li H, Rekatsinas T, Doan A, Park Y, Krishnan G, Deep R, Arcaute E, Raghavendra V. Deep learning for entity matching: A design space exploration. In: Proc. of the 2018 Int'l Conf. on Management of Data. ACM, 2018. 19−34.

[15] Li JZ, Wang HZ, Gao H. State-of-the-Art of research on big data usability. Ruan Jian Xue Bao/Journal of Software, 2016,27(7): 1605−1625 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5038.htm [doi: 10.13328/j.cnki.jos.005038]

[16] Christen P. A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans. on Knowledge and Data Engineering, 2012,24(9):1537−1555. [doi: 10.1109/TKDE.2011.127]

[17] Shao J, Wang Q, Lin Y. Skyblocking for entity resolution. Information Systems, 2019,85:30−43.

[18] Nascimento DC, Pires CES, Mestre DG. Exploiting block co-occurrence to control block sizes for entity resolution. Knowledge and Information Systems, 2019, 1−42.

[19] Allam A, Skiadopoulos S, Kalnis P. Improved suffix blocking for record linkage and entity resolution. Data & Knowledge Engineering, 2018,117:98−113.

[20] Fisher J, Christen P, Wang Q, Rahm E. A clustering-based framework to control block sizes for entity resolution. In: Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2015. 279−288.

[21] Hernández MA, Stolfo SJ. Real-World data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery, 1998,2(1):9−37.

[22] Draisbach U, Naumann F, Szott S, Wonneberg O. Adaptive windows for duplicate detection. In: Proc. of the 2012 IEEE 28th Int'l Conf. on Data Engineering. IEEE, 2012. 1073−1083.

[23] Papadakis G, Ioannou E, Palpanas T, Niederee C, Nejdl W. A blocking framework for entity resolution in highly heterogeneous information spaces. IEEE Trans. on Knowledge and Data Engineering, 2013,25(12):2665−2682.

[24] Papadakis G, Koutrika G, Palpanas T, Nejdl W. Meta-Blocking: Taking entity resolution to the next level. IEEE Trans. on Knowledge and Data Engineering, 2014,26(8):1946−1960.

[25] Gentile AL, Ristoski P, Eckel S, Ritze D, Paulheim H. Entity matching on Web tables: A table embeddings approach for blocking. In: Proc. of the 22nd Int'l Conf. on Extending Database Technology. 2017. 510−513.

[26] Kenig B, Gal A. Mfiblocks: An effective blocking algorithm forentity resolution. Information Systems, 2013,38(6):908−926.

[27] Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records. In: Getoor L, Senator TE, Domingos PM, Faloutsos C, eds. Proc. of the ACM KDD Workshop on Data Cleaning and Object Consolidation. New York: ACM, 2003. 73−78.

[28] Sun CC, Shen DR, Kou Y, Nie TZ, Yu G. Entity resolution oriented clustering algorithm. Ruan Jian Xue Bao/Journal of Software, 2016,27(9):2303−2319 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5043.htm [doi: 10.13328/j.cnki.jos.005043]

[29] Melnik S, Garcia-Molina H, Rahm E. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In: Proc. of the 18th Int'l Conf. on Data Engineering. IEEE, 2002. 117−128.

[30] Jeh G, Widom J. SimRank: A measure of structural-context similarity. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2002. 538−543.

[31] Motwani R, Raghavan P. Randomized Algorithms. Cambridge: Cambridge University Press, 1995.

[32] Christen P, Pudjijono A. Accurate synthetic generation of realistic personal information. In: Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer-Verlag, 2009. 507−514.

[33] Christen P. Automatic record linkage using seeded nearest neighbour and support vector machine classification. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2008. 151−159.

附中文参考文献:

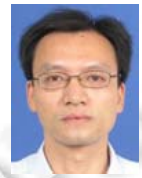
[15] 李建中,王宏志,高宏.大数据可用性的研究进展.软件学报,2016,27(7):1605−1625. http://www.jos.org.cn/1000-9825/5038.htm [doi: 10.13328/j.cnki.jos.005038]

[28] 孙琛琛,申德荣,寇月,聂铁铮,于戈.面向实体识别的聚类算法.软件学报,2016,27(9):2303−2319. http://www.jos.org.cn/1000-9825/5043.htm [doi: 10.13328/j.cnki.jos.005043]

孙琛琛(1987－),男,山西晋中人,博士,讲师,CCF 专业会员,主要研究领域为实体解析,实体消歧,大数据分析与挖掘.

申德荣(1964－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为分布式数据管理,数据集成.

李玉坤(1969－),男,博士,教授,CCF 专业会员,主要研究领域为数据库,信息检索,个人信息管理.



肖迎元(1969－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,个性化推荐系统,大数据挖掘与分析.

马建红(1965－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为软件工程,智能处理,自然语言处理,知识图谱,创新理论与方法.