软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2017,28(9):2281−2292 [doi: 10.13328/j.cnki.jos.005187]

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

# 一种基于隐私保护下的多方记录链接方法*

韩姝敏, 申德荣, 聂铁铮, 寇 月, 于 戈

(东北大学 计算机科学与工程学院,辽宁 沈阳 110819)

通讯作者: 韩姝敏, E-mail: hanshumin_summer@yeah.net

摘 要: 多方隐私保护下的记录链接(privacy-preserving record linkage,简称 PPRL)是在隐私保护下,从多个数据源中找出代表现实世界中同一实体的过程.该过程除了最终匹配结果被数据源之间共享外,其他信息均未被泄露.随着数据量的日益增大和现实世界数据质量问题的存在(如拼写错误、顺序颠倒等),多方 PPRL 方法的可扩展性和容错性面临挑战.目前,已有的大部分多方 PPRL 方法都是精确匹配方法,不具有容错性.还有少部分多方 PPRL 近似方法具有容错性,但在处理存在质量问题的数据时,由于容错性差和时间代价过大,并不能有效地找出数据源间的共同实体.因此,提出一种结合布隆过滤、安全合计、动态阈值、检查机制和改进的 Dice 相似度函数的多方 PPRL 近似方法.首先,利用布隆过滤将各数据源中的每条记录信息转换成由 0 和 1 组成的位数组.然后,计算每个对应位置 bit 1 所占的比率,并利用动态阈值和检查机制来判定匹配成功的位置.最后,通过改进的 Dice 相似度函数计算出记录间的相似度,进而判断记录间是否匹配成功.实验结果表明:所提出的方法具有较好的可扩展性,并且在保证查准率的同时,比已有的多方近似 PPRL 方法具有更高的容错性.

关键词: 记录链接;隐私保护;布隆过滤;动态阈值;检查机制;改进的 Dice 相似度函数
中图法分类号: TP311

中文引用格式: 韩姝敏,申德荣,聂铁铮,寇月,于戈.一种基于隐私保护下的多方记录链接方法.软件学报,2017,28(9):2281−2292. http://www.jos.org.cn/1000-9825/5187.htm

英文引用格式: Han SM, Shen DR, Nie TZ, Kou Y, Yu G. Multi-Party privacy-preserving record linkage approach. Ruan Jian Xue Bao/Journal of Software, 2017,28(9):2281−2292 (in Chinese). http://www.jos.org.cn/1000-9825/5187.htm

## Multi-Party Privacy-Preserving Record Linkage Approach

HAN Shu-Min, SHEN De-Rong, NIE Tie-Zheng, KOU Yue, YU Ge

(School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)

**Abstract**: Multi-party privacy-preserving record linkage is the process of identifying records that correspond to the same real-world entities across several databases without revealing any sensitive information about these entities. With the increasing amount of data and the real-world data quality issues (such as spelling errors and wrong order), scalability and fault tolerance of PPRL have become the main challenges. At present, most of the existing multi-party PPRL methods apply exact match without fault-tolerant. There are a few other PPRL approximate methods with fault-tolerant, but when dealing with the existing data quality issues, due to the low fault-tolerance and high time cost, they cannot effectively find out the common entities between databases. To tackle this issue, this paper proposes a multi-party PPRL approximate approach combined with bloom filter, secure summation, dynamic threshold, check mechanism, and improved Dice similarity function. First, bloom filter is used to convert each record in the databases to an array of 1 and 0. Then, ratio of bit 1 is calculated for each corresponding position, and dynamic threshold and check mechanism are used to determine matched position.

* 基金项目: 国家自然科学基金(61472070, 61672142); 国家重点基础研究发展计划(973)(2012CB316201)

Foundation item: National Natural Science Foundation of China (61472070, 61672142); National Grand Fundamental Research Program of China (973) (2012CB316201)

收稿时间: 2016-07-11; 修改时间: 2016-09-04; 采用时间: 2016-11-10; jos 在线出版时间: 2017-02-20

CNKI 网络优先出版: 2017-02-20 14:05:59, http://www.cnki.net/kcms/detail/11.2560.TP.20170220.1405.020.html

Finally, the similarity between records is calculated by improved Dice similarity function to judge whether records are matched. Experimental results show the proposed method has good scalability and higher fault tolerance than the existing multi-party PPRL approximate method with good precision.

**Key words**: record linkage; privacy-preserving; Bloom filter; dynamic threshold; check mechanism; improved Dice similarity function

随着科技的不断进步,数据正快速地增长和累积.减少数据冗余、实现数据共享,已成为大数据时代的首要任务.记录链接(record linkage)[1]是指从一个或多个数据源中匹配出代表现实世界中同一实体的记录,记录链接在金融、医疗、政府等领域具有广泛的应用前景.但是,当记录信息涉及到个人隐私或敏感信息时,我们必须要考虑记录信息的隐私保护问题.因此近年来,研究隐私保护下的记录链接(PPRL).PPRL 技术可以保证在记录链接的过程中,只有最终匹配结果被各数据源间共享,其他未匹配的记录信息均未被泄露.例如在分散的医疗体系中,某人的医疗信息可能分布在多个医院,找出同一个人在不同医院的诊断信息有利于更准确的分析病情,但由于涉及到患者隐私,各医院并不希望暴露患者的医疗信息.PPRL 技术既可以找出某位患者在各医院的医疗信息,又能够保证各医院其他患者的医疗信息不被泄露.因此,PPRL 技术不仅具有理论研究价值,而且有着重要和迫切的实际应用价值.

目前,已有很多文献提出了 PPRL 的方法[2−8],其中大多数方法都只适用于两个数据源,对于 3 个及以上的多方 PPRL 方法的研究还很少.这是因为想要找到一个可以合理度量多条记录相似度的方法并不容易.适用于两个数据源的相似度度量方法大多数并不适用于多数据源.然而,现实中的很多应用进行记录链接的数据源往往不只两个.因此,多方 PPRL 的研究具有重要的现实意义.随着数据量的不断增长和现实世界数据质量问题的存在(如拼写错误、顺序颠倒等),多方 PPRL 方法的可扩展性和容错性成为研究的重点.目前,大部分多方 PPRL 方法只适用于精确匹配[9−11],即:只有各记录完全一致才认为匹配成功,并不能处理存在质量问题的数据(data with quality issues,简称 DQI),制约了其实际应用.还有少部分非精确的近似匹配方法可以处理 DQI,但由于应用密码学保护技术导致时间代价较大,不具有可扩展性.文献[2]提出了一种可扩展的基于“与”运算的多方 PPRL 近似方法(BasedAnd_PPRL,简称 BA_PPRL),但在处理 DQI 类型的数据时,该方法会丢失较多真实匹配的记录,导致查全率过低.可见,已有解决方法不能有效地处理 DQI 类型的数据.因此,研究出一种时间代价较低又能有效处理 DQI 类型数据的多方 PPRL 近似方法,是亟待解决的问题.

本文提出一种基于比率的结合布隆过滤(bloom filter)、动态阈值(dynamic threshold)、检查机制(check mechanism)、安全合计(secure summation)和改进的 Dice 相似度函数的多方近似 PPRL 方法(BasedRatio_PPRL,简称 BR_PPRL).BA_PPRL 方法[2]与本文方法相近,并且具有较高的准确率和较低的时间代价,但当处理存在质量问题的数据时,该方法查全率过低,导致其很难具有应用价值.本文提出的方法在保证准确率的前提下,显著提升了查全率且时间代价较低,更具有应用价值.

本文的贡献点是:

(1) 基于比率来标识记录间在某位置的相似度,能够有效地提高方法的容错性;
(2) 基于动态阈值及检查机制来确定真正匹配的位置,能够保证在提高查全率的同时,查准率不受影响;
(3) 利用改进的 Dice 相似度函数,能够有效计算带有质量问题的记录间的相似度;
(4) 通过实验证明了本文方法可以有效地处理 DQI 类型的数据,即:保证查准率的同时,具有更高的查全率和较小的时间代价.因此,本文方法拥有更高的容错性和较好的可扩展性.

本文第 1 节简述与本文方法有关的相关工作.第 2 节说明本文方法用到的背景知识.第 3 节描述本文提出方法的流程及算法,并对其隐私保护程度进行分析.第 4 节给出实验结果,并对其加以分析.最后,在第 5 节对本文进行总结,并指出进一步的工作方向.

## 1 相关工作

目前,对于多方 PPRL 解决方法的研究还处于起步阶段.国外有较少的相关方法被提出;在国内,还未见有相关方法.最早的多方 PPRL 方法[9]是通过将各个数据源的记录编码,然后传入另一方进行对比.但是,这个方法只

适用于精确匹配.2004 年,一种基于安全多方计算的精确匹配方法[10]被提出.2008 年,文献[11]提出通过安全等价链接进行精确匹配的方法.文献[10,11]应用的隐私保护方法时间代价较高.文献[12]提出一种基于$k$-匿名和博弈概念的近似匹配方法,但只能应用于分类属性上.

文献[13]最早提出了利用 Bloom Filter 将记录属性值转换为 0 和 1 组成的位数组,并对记录间做“与”运算,如果运算结果的位数组与各记录的位数组均相同,则匹配成功.该方法虽然时间代价较小,但却只适用于精确匹配.为此,文献[2]在该方法的基础上提出了与安全合计和 Dice 相似函数相结合的近似匹配方法.同样,该方法时间代价较小,且准确率较高.但当数据源中记录存在质量问题时,该方法的查全率会降低,导致其实际应用价值较低.本文提出的方法通过计算位数组的各对应位置 bit 1 所占的比率,并利用动态阈值、检查机制和改进的 Dice 相似函数,使得在处理存在质量问题的数据时,匹配结果的查全率显著提高.

## 2 背景知识

### 2.1 布隆过滤

Bloom Filter[14]是一种空间效率很高的随机数据结构,它可以将集合转换为位数组.下面具体介绍 Bloom Filter 是如何用位数组表示集合的.初始状态时,Bloom Filter 是一个包含 $m$ 位的位数组,每一位都置为 0.为了表达 $S=\{x_1,x_2,\ldots,x_n\}$这样 $n$ 个元素的集合,Bloom Filter 使用 $k$ 个相互独立的哈希函数(hash function),它们分别将集合中的每个元素映射到$\{1,\ldots,m\}$的范围中.对任意一个元素 $x$,第 $i$ 个哈希函数映射的位置 $h_i(x)$就会被置为 $1(1\leq i\leq k)$.

### 2.2 多方安全Bloom Filter匹配

在判断多个位数组是否来自同一实体时,为了保证隐私,各个位数组应尽少暴露自身信息给其他参与方.因此在进行多方 Bloom Filter 匹配时,每个位数组被均分成 $P$(参与方个数)个部分$\{B_1,B_2,\ldots,B_P\}$,$B_1,B_2,\ldots,B_P$ 分别相应传到参与方 $P_1,P_2,\ldots,P_P$ 中.这样,每个参与方只获知位数组中 $m/P$ 个字符,保证了数据的安全.

### 2.3 安全合计

安全合计是安全多方计算中的一种方法,被广泛应用于各种 PPRL 方法中.它的基本思想是:各参与方将自己的数据加密后再进行合计,得到合计结果后解密即为真实的合计结果.该过程各方除了获得最终的真实合计结果外,对其他方的真实数据均一无所知.

### 2.4 Dice相似函数

在计算多个位数组之间相似度时,可以利用基于集合的 Dice 相似度计算函数.假如计算 $B_1,B_2,\ldots,B_P$ 这 $P$ 个位数组之间的相似度,见公式(1):

$$Dice\_sim(B_1,\ldots,B_P)=\frac{P\times|B_1\cap B_2\ldots\cap B_P|}{\sum_{i=1}^{P}|B_i|} \tag{1}$$

其中,$|B_i|$代表第 $i$ 个位数组中 1 的个数.

## 3 多方 BR_PPRL 近似方法

本文提出的多方 BR_PPRL 近似方法由 3 个模块构成:数据准备及生成模块、记录近似匹配模块和相似度计算模块,如图 1 所示,详见算法 1.

算法 **1**. 多方 BR_PPRL 近似匹配算法.

输入:$D_i$:来自参与方 $P_i$ 的数据源;$i,j,z,1\leq i,j\leq P,1\leq z\leq m$;$T_D$:动态阈值;$T_A$:全局阈值;

输出:匹配记录集合 $M$.

1: *Blocking*($D_i$); //先对每个数据源进行分块处理

2: $B_i$=*genetateBloomfilters*($D_i$,$A$); //对每个数据源中的记录应用 Bloom Filter

3: $divide(B_i)=\{B_{i,1},\ldots,B_{i,j},\ldots,B_{i,P}\}$; //将每个 0,1 字符串分为 $P$ 个部分

4: send $B_{i,j}$ to $P_j$; //将字符串的第 $j$ 部分传给参与方 $j$

5: $R_{i,z}=generateRatios(B_{i,z})$; //计算每个竖列的比率

6: **IF** $R_{i,z}\geq T_D$ **THEN**:

7: **IF** $R_{i,z}=1$ **THEN**:

8: $Y_{i,z}=R_{i,z}$;

9: **ELSE IF** $checkMecharism(S'_{i,z})$=Ture

10: $V_{Rz}=R_{i,z}$;

11: **ELSE** $Y_{i,z}=0$;

12: **ELSE** $Y_{i,z}=0$;

13: $R_a=secureSummation(Y_{i,z},V_{Rz})$;

14: $Dice\_sim=\dfrac{P\times R_a}{\sum_{i=1}^{P}\sum_{z=1}^{m}|B_{i,m}|}$;

15: **IF** $Dice\_sim\geq T_A$

16: $M=\{B_1,B_2,\ldots,B_P\}$ as a Match;

17: **RETURN** $M$;

18: **ELSE**

19: $\{B_1,B_2,\ldots,B_P\}$ as a Non-match;

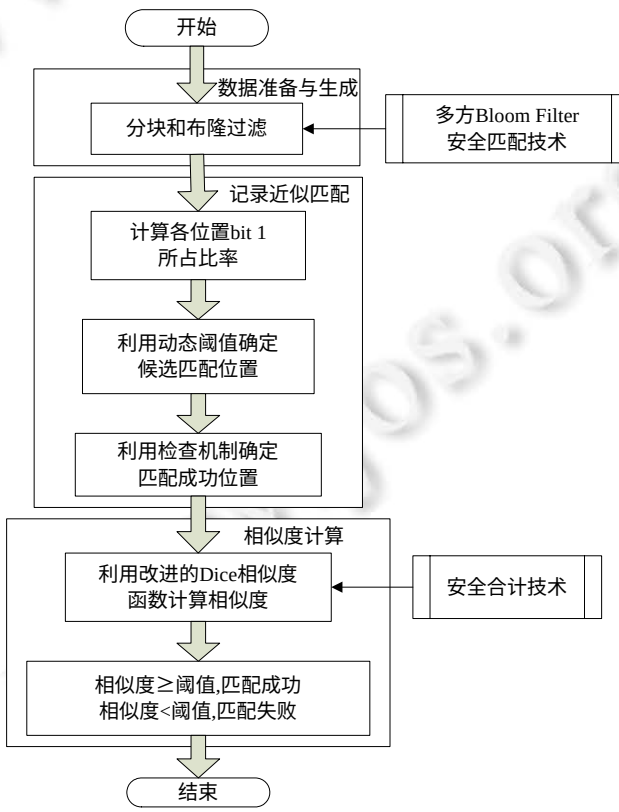


Fig.1 Process of the multi-party BR_PPRL approximate approach

图 1 多方 BR_PPRL 近似方法流程

数据准备及生成模块的主要功能是统一各数据源之间参数、分块,并利用 Bloom Filter 将数据源中各记录的公共属性值转换为由 0,1 组成的位数组.转换得到的位数组在一定程度上代表了记录并保护了记录的隐私,位数组将代表各记录在后续模块中进行匹配.

记录近似匹配模块的主要功能是判断来自各数据源的位数组间匹配成功的位置.首先,计算位数组间各对应位置 bit 1 所占的比率;然后,利用动态阈值选出候选匹配成功位置;最后,通过检查机制得到真正匹配成功的位置.

相似度计算模块的主要功能是计算来自各数据源的位数组间的相似度,进而判断这些位数组是否来自于同一实体.若相似度不小于全局阈值,则匹配成功;反之,匹配失败.

### 3.1 数据准备及生成

在算法开始前,需要先确定输入参数并保证各参与方参数一致,表 1 说明了本文方法用到的参数及其意义;然后,对各数据源 $D_i$ 应用相同的分块方法,减少候选匹配对的数量,从而减少时间代价;接着,将各数据源 $D_i$ 中的 $N_i$ 条记录的属性 $A$ 的值变为 $q$-$gram$;最后,Bloom Filter 应用 $k$ 个哈希函数将其映射成 $N_i$ 个 $m$ 长度的位数组(第 1 行、第 2 行).

为了保证位数组的安全,利用多方安全 Bloom Filter 匹配技术,将长度为 $m$ 的位数组均分成 $P$ 个片段($m/p$),并将第 $j$ 个片段传给 $P_j$($1\le j\le P$),参与方 $P_j$ 收到 $P$–1 个来自其他参与方的第 $j$ 个片段(第 3 行、第 4 行).

**Table 1** Multi-Party BR_PPRL approximate approach parameter table

**表 1** 多方 BR_PPRL 近似方法参数表

| 参数 | 意义 | 参数 | 意义 |
|---|---|---|---|
| $P$ | 参与方的个数 | $D_i$ | 参与方 $P_i$ 的数据源 |
| $N_i$ | 数据源 $D_i$ 的大小 | $i$ | 代表第 $i$ 个参与方($1\le i\le P$) |
| $j$ | 代表第 $j$ 个片段($1\le j\le P$) | $z$ | 代表位数组的第 $z$ 个位置($1\le z\le m$) |
| $A$ | 各数据源间公共属性集合 | $L$ | 各数据源间用于分块的属性集合 |
| $m$ | Bloom Filter 中位数组的固定长度 | $k$ | Bloom Filter 中的 $k$ 个哈希函数 $h_1,h_2,\ldots,h_k$ |
| $q$ | $q$-$gram$ | $T_D$ | 动态阈值 |
| $T_A$ | 全局阈值 | | |

### 3.2 记录近似匹配

本文提出的多方 BR_PPRL 近似方法对位数组间对应的位置进行比率计算,得到 bit 1 所占的比率,然后,利用动态阈值和检查机制判断匹配成功的位置.比率计算、动态阈值和检查机制均为该算法的关键环节,由于相似度结果是依据匹配成功位置得出的,因此,全面而准确地找出匹配成功的位置,对方法的性能具有关键性影响.尤其当数据存在质量问题时,该方法的性能更优于已有方法.

#### 3.2.1 比率计算

BA_PPRL[2]是目前最优的多方近似方法,在该方法中,为了得到匹配成功的位置,各位数组将对应位置字符做“与”运算,$B_{1,z}\cap B_{2,z}\cap\ldots\cap B_{P,z}$ 表示 $P$ 个位数组在位置 $z$ 做“与”运算($1\le z\le m$).只有各对应位置字符均为 1 时,“与”运算结果才为 1,代表该位置匹配成功 $S_z$.然而,当其中某条记录存在质量问题时,可能使得一些原本为 1 的位置变为 0,进而导致一些真实应该匹配成功的位置被丢失,最终造成真实匹配结果的丢失,导致查全率过低.

为支持存在质量问题的数据的多方近似匹配,本文的 BR_PPRL 方法通过计算各位数组对应位置 bit 1 所占比率 $R_z$ 代替“与”运算(第 5 行),即 $B_{1,z},B_{2,z},\ldots,B_{P,z}$ 代表 $P$ 个位数组在位置 $z$ 的值($1\le z\le m$),$O_{i,z}$ 代表 $P$ 个位数组在位置 $z$ 值为 1 的个数($1\le i\le P$),因此, $R_z=\dfrac{O_{i,z}}{P}$. 比率可以更准确地代表各位数组在某个位置的相似度,因此,通过比率而不是“与”运算结果判断某位置是否匹配成功,可以提高方法的容错率,减少真实匹配结果的丢失,提高查全率.

#### 3.2.2 动态阈值

为了利用比率判断某位置是否匹配成功,本文提出动态阈值 $T_D$ 的概念,$T_D$ 用来筛选候选匹配位置.如果某

位置比率为 1,则直接作为匹配成功位置 $S_z$;否则,若大于等于 $T_D$,则作为候选匹配位置 $S'_z$ (第 6 行~第 12 行).在给出 $T_D$ 前,先计算当参与方个数为 $P$、容错率为 $X$ 时,即,允许每条记录发生错误的概率为 $X$,每竖列对应位置中 $a$ 个位置出现错误的概率 $P_e$,$P_e$ 见公式(2).

$$P_e = C_P^a X^a (1-X)^{P-a} \tag{2}$$

由二项分布定理得知:当 $a=PX$ 时,$P_e$ 最大.$a$ 取整 $a'=[PX]$.由于每个竖列出现 $a'$个错误的概率最大,那么每个竖列最多允许 $a'$个错误发生.所以,动态阈值 $T_D$ 见公式(3)所示.

$$T_D = \frac{P-a'}{P} \tag{3}$$

动态阈值由 $P$ 和 $X$ 决定,当 $P$ 一定时,$X$ 越大,$T_D$ 越小,允许发生错误的位置越多;反之,$X$ 越小,$T_D$ 越大,允许发生错误的位置越少.当 $X$=0 时,$T_D$=1,不允许任何错误发生,等同于 BA_PPRL 方法.

3.2.3 检查机制

候选匹配位置 $S'_z$ 代表该位置有可能是真实匹配,但又存在质量问题的位置.为了保证准确率不受影响,需要对候选匹配位置做进一步判断.本文提出一种检测机制,可以测出某条记录是否是存在质量问题但又匹配成功的记录,详见算法 2.

算法 **2**. 检查机制算法.

输入:来自于 $P$ 个参与方的位数组 $B_1,B_2,\ldots,B_P$;候选匹配位置 $S'_z$;

输出:True or False.

1: $\sigma=\sqrt{PX(1-X)}$; //计算$\sigma$

2: $F'=getlocation0(S'_z)$; //找出候选位置中存在质量问题的记录

3: $F=|F'|$; //集合 $F'$的个数即存在质量问题记录的个数

4: **IF** $F<a'+3\sigma$ **THEN**:

5: $B_a=getBF0(S'_z)$; //找出存在质量问题的记录

6: $B_b=getclosedBF1(F')$; //找出与其最近的正常记录

7: $Dice\_sim(B_a,B_b)=\frac{2\times|B_a\cap B_b|}{|B_a|+|B_b|}$;

8: **IF** $Dice\_sim>T_A$ **THEN**:

9: RETURN **True**;

10: **ELSE**

11: RETURN **False**;

12: **ELSE**

13: RETURN **False**;

检查机制的总体思想:

1. 利用切比雪夫定理判断存在质量问题的记录数是否在合理范围内(第 1 行~第 4 行),即,找出所有候选位置中存在质量问题的记录即位置为 0 所对应的记录:若存在质量问题记录的个数 $F<a'+3\sigma$,则继续检查;否则,检查不通过.其中,$\sigma$为公式(2)的标准差,见公式(4).

$$\sigma=\sqrt{PX(1-X)} \tag{4}$$

由切比雪夫定理得到公式(5):

$$F<a'+3\sigma \tag{5}$$

由于 $F<a'+3\sigma$的概率为 89%,所以当存在质量问题记录的个数 $F>a'+3\sigma$时,直接判断记录间不匹配;

2. 通过上述检查的候选匹配位置继续判断存在质量问题的记录与距其最近的正常记录是否匹配成功(第 5 行~第 11 行).将位置为 0 的位数组与距离最近的位置为 1 的位数组传入一个绝对安全的第 $P$+1 方,同样利用 Dice 相似度函数计算两者相似度,若大于全局阈值则检测通过,$S_{Rz}$ 代表真实匹配但又存在质量问题的位置;否

则不通过.

### 3.3 相似度计算

传统的 Dice 相似度函数见公式(1).本文方法提出改进的 Dice 相似度函数,见公式(6).

$$Imp\_Dice\_sim(B_1,...,B_P)=\frac{P\times(|S_z|+V_{R_z})}{\sum_{i=1}^{P}|B_i|} \tag{6}$$

其中,$B_1,B_2,\ldots,B_P$ 代表来自于 $P$ 个参与方的 $P$ 个位数组,$|B_i|$代表 $B_i$ 中 bit 1 的个数,$|S_z|$代表比率为 1 的位置个数,$V_{R_z}$ 代表位置 $S_{R_z}$ 的比率值.

由此,可以推导出公式(7):

$$Imp\_Dice\_sim=Dice\_sim+\frac{P\times V_{R_z}}{\sum_{i=1}^{P}|B_i|} \tag{7}$$

- 当不存在位置 $S_{R_z}$ 时,*Imp_Dice_sim=Dice_sim*;
- 当存在位置 $S_{R_z}$ 时,*Imp_Dice_sim* 在 *Dice_sim* 上加入了存在质量问题的真实匹配位置.

因此,本文提出的方法对于存在质量问题但又真实匹配的记录依然可以计算出较高相似度,显著减少了真实匹配记录的丢失.所以,本文方法在处理带有质量问题的数据时,显著提高了查全率.

应用多方 Bloom Filter 安全匹配技术的改进的 Dice 相似度函数见公式(8).

$$Dice\_sim=\frac{P\times\left(\sum_{j=1}^{P}|S_{j,z}|+\sum_{j=1}^{P}V_{j,R_z}\right)}{\sum_{i=1}^{P}\sum_{j=1}^{P}|B_{i,j}|} \tag{8}$$

其中,加和运算均利用安全合计保证数据安全,若相似度不小于阈值,则匹配成功;否则,匹配失败(第 13 行~第 19 行).$|B_{i,j}|$代表第 $i$ 个位数组片段 $j$ 中 bit 1 的个数.

### 3.4 多方**BR_PPRL**近似算法示例

下面举个例子来说明上述过程.如图 2 所示,有五方参与本文方法.

| $P_1$ | | $P_2$ | | $P_3$ | | $P_4$ | | $P_5$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | $B_1$ |
| 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | $B_2$ |
| 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | $B_3$ |
| 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | $B_4$ |
| 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | $B_5$ |
| 1 | 0 | 1 | 0 | 1 | 1 | 4/5 | 0 | 3/5 | 0 | *Ratio* |

Fig.2 An example of BR_PPRL approach with five party

图 2 五方 BR_PPRL 方法示例

首先,从 5 个数据源中各提取出一条记录,并用 Bloom Filter 将记录中属性 $A$ 的值转换为 5 个位数组{$B_1,B_2,B_3,B_4,B_5$};接下来,将 5 个位数组均分为 5 个片段,并把相应片段传到对应参与方中,然后计算各对应位置中 bit 1 所占的比率.当 $P$=5,$X$=1/5 时,动态阈值为 $T_D$=4/5.因此,1 作为匹配成功位置继续保留下来,4/5 不小于动态阈值,作为候选匹配位置进入检查机制进行判断:首先找出带有质量问题的记录 $B_1$,个数小于 $a'+3\sigma$.继续检查.由于 $B_1$ 与距其最近且为 1 的 $B_2$ 相似度为 0.89,大于全局阈值 0.8,所以 4/5 作为带有质量问题的匹配成功位置.3/5 小于动态阈值则未匹配成功变为 0.由公式(8)可得,Dice 相似度结果为 0.89,大于全局阈值 $T_A$,所以这 5 条记录匹配成功,均来自同一实体.方法 BA_PPRL 计算得到的相似结果为 0.74,小于全局阈值 $T_A$,导致真实匹配记录的丢失.

$$Dice\_sim=\frac{P\times\left(\sum_{j=1}^{P}|S_{j,z}|+\sum_{j=1}^{P}V_{j,R_z}\right)}{\sum_{i=1}^{P}\sum_{j=1}^{P}|B_{i,j}|}=\frac{5\times(1+1+1+1+4/5)}{4+5+6+6+6}=0.89.$$

### 3.5 隐私保护分析

我们假设所有参与方都遵循 Honest-But-Curious 攻击方法,即:参与方都尽力通过已获得的信息推测其他方的信息,但不会串通、恶意攻击等.本文方法应用安全多方 Bloom Filter 匹配技术,使得每个参与方只能看见其他方 $m/P$ bits 的片段.因此,$P$ 越大,则安全性越高.

在检查机制中由于在绝对安全的第 $P$+1 方进行匹配,所以过程中并没有信息泄露.

## 4 实验与分析

### 4.1 准备工作

本文用 Python(2.7.11 版本)实现文中提出的方法.处理器:Intel(R)Core(TM) i5-4590;主频:3.3GHz,8 核;内存:8GB;操作系统:Windows 7,64 位.

本文实验采用的数据集为北卡罗来纳州选民登记名单(North Carolina voter registration list,简称 NCVR),该数据集中的记录都是公共可获得的真实个人信息.数据集可以从 ftp://alt.ncsbe.gov/data/上下载.

对本文方法,我们从运行时间(runtime)、查准率(precision)、查全率(recall)和 $F$ 指数这 4 个方面进行评估[15].运行时间用来评估本文方法的可扩展性,查准率、查全率和 $F$ 指数用来评估本文方法的匹配质量.查准率代表本文方法找出匹配成功的记录中,真实匹配记录所占的比率,查准率越高,代表方法的结果越准确.查全率是指本文方法找出的真实匹配的记录占所有真实匹配记录的比率,查全率越高,代表方法找到的真实匹配记录越全.$F$ 指数则用来综合评估方法的结果.

本文分别以数据源大小、参与方个数、容错率和扰乱比例为自变量,评估以上 4 个指标的变化情况.参与方个数的变化范围为[3,5,7,10];数据源大小是从 NCVR 中分别提取 5 000,10 000,50 000,100 000,500 000 和 1 000 000 条记录分给每个参与方,并保证不同大小的数据源中均有一半的记录是各参与方共有的;容错率变化范围为[1/5,2/5,3/5,4/5,1];扰乱比例(disturbed ratio,简称 DR)变化范围为[1/4,1/2,3/4,1],扰乱比例为占各数据源共有记录的比例.

本文方法的优势在于:处理存在质量问题的现实数据时,比之前方法具有更高的查全率且;在检查机制的保护下,查准率也不会受到太大影响.因此,为了评估不同程度质量问题对本文方法匹配结果的影响,利用文献[16]中的工具,可以生成不同程度质量问题的数据集.本文生成 3 个扰乱数据集,扰乱程度分别为每条记录至多一个拼写错误(Mod-1)、两个拼写错误(Mod-2)和 3 个拼写错误(Mod-3).拼写错误包括增加、删除和替换字符.

本文实验的参数设置设定 $m$=1000,$k$=30,$q$=2,$T_A$=0.8.本文应用基于语音分块的探测法[17]来提高方法的可扩展性.

### 4.2 实验结果与分析

#### 4.2.1 可扩展性评估

首先评估本文方法的运行时间随数据源大小增加的变化情况.其中,参与方个数 $P$=5,容错率 $X$=1/5,扰乱比例占共有记录中的 1/2.图 3 表示本文方法比 BA_PPRL 方法运行时间稍长,但仍具有较好的可扩展性.这是因为本文提出的比率计算和检查机制导致运行时间变长.

下面评估本文方法运行时间随着参与方个数增加的变化情况.数据集大小均为$|D_i|$=100000,容错率 $X$=1/5,扰乱比例占共有记录中的 1/2.图 4 表示随着参与方个数的增加,运行时间变短.这是因为随着 $P$ 的增加,1/$P$ 变小,即,分到每个参与方的片段变短,所以运行时间变短.

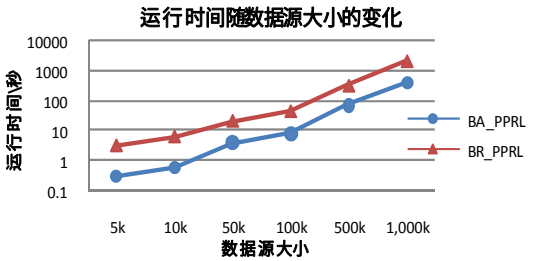


Fig.3 Runtime with database sizes

图 3 运行时间与数据源大小关系图

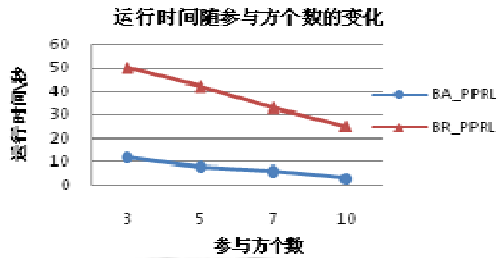


Fig.4 Runtime with different $P$

图 4 运行时间与参与方个数关系图

4.2.2 方法性能评估

本节分别从查全率、查准率和 $F$ 指数这 3 方面对方法 BR_PPRL 的性能进行评估,并与方法 BA_PPRL 进行对比.数据集分别选取 3 个不同程度扰乱数据集 Mod-1,Mod-2 和 Mod-3.

首先评估在 Mod-1 数据集中,两种方法的 3 个评价指标随参与方个数增加的变化情况.其中,$|D_i|$=5000,容错率 $X$=1/5,扰乱比例占共有记录中的 1/2.如图 5 所示,两种方法的查全率均随 $P$ 的增加而减小.这是因为当数据存在质量问题时,$P$ 越大,丢失的真实匹配记录越多.BA_PPRL 方法在处理存在质量问题的数据时,查全率下降很快.当 $P$=3 时,查全率最高,$Recall$=0.5;当 $P$=10 时,查全率几乎为 0;本文提出方法的查全率始终较高,当 $P$=10 时,查全率依然在 0.5 以上.两种方法的查准率均随 $P$ 的增加而减小,且两者的查准率较为相近.通过观察 $F$ 指数的变化可以看出:在处理扰乱数据集时,本文方法的性能明显优于方法 BR_PPRL.

接下来评估在不同程度扰乱数据集上,本文提出方法的 3 个评价指标随参与方个数增加的变化情况.其中,$|D_i|$=5000,容错率 $X$=1/5,扰乱比例为共有记录中的 1/2.如图 6 所示:随着扰乱程度的增加,本文方法的查全率、查准率和 $F$ 指数都在下降.这是由于扰乱程度增加后,一些真实匹配的记录会更容易被丢失.

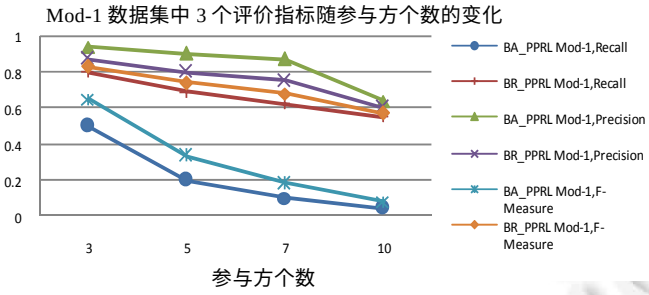


Fig.5 Three evaluation indexs with different $P$ in mod-1 databases

图 5 在 Mod-1 数据集中,3 个评价指标与参与方个数关系图

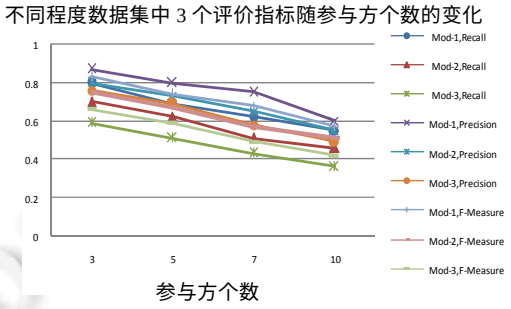


Fig.6 Three evaluation index with different $P$ in different mod databases

图 6 在不同程度扰乱数据集中,3 个评价指标与参与方个数关系图

然后评估在 Mod-1 数据集中,两种方法的 3 个评价指标随容错率增加的变化情况.其中,$P$=5,$|D_i|$=5000,扰乱比例为共有记录中的 1/2.如图 7 所示:在方法 BA_PPRL 中,3 个评估指标并不受容错率变化的影响,因此一直保持恒定值,查全率为 0.2,查准率为 0.9,$F$ 指数为 0.33.本文方法随着容错率的增大,查全率不断增大,查准率不断下降.这是因为容错率越大,动态阈值越小,找到的匹配对数越多,但相应的准确率会下降.在检查机制的保护下,容错率为 1 时,查准率依然在 0.4 以上.通过 $F$ 指数的变化可以看出,本文方法能够更好地处理存在质量问题的数据集.

接下来评估在不同程度扰乱数据集上,本文提出方法的 3 个评价指标随容错率增加的变化情况.其中,$P$=5,$|D_i|$=5000,扰乱比例为共有记录中的 1/2.如图 8 所示:随着扰乱程度的增加,本文方法的查全率、查准率和 $F$

指数都在下降.同样是由于扰乱程度增加后,一些真实匹配的记录会更容易被丢失.

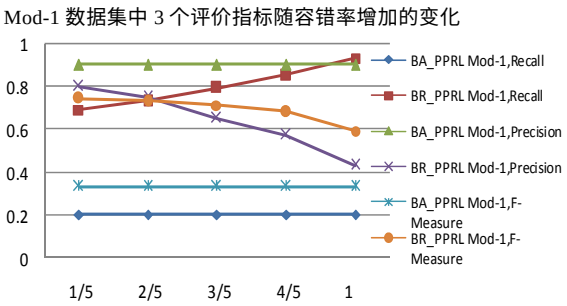


Fig.7 Three evaluation index with different $X$ in mod-1 databases

图 7 在 Mod-1 数据集中,3 个评价指标与容错率大小关系图

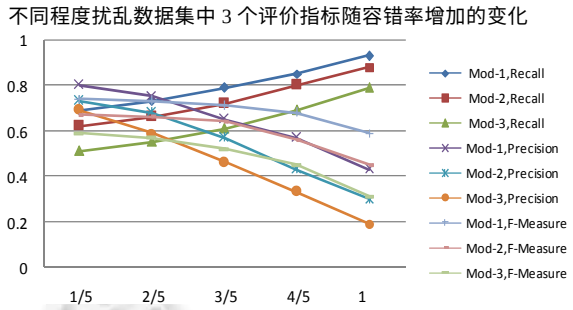


Fig.8 Three evaluation index with different $X$ in different mod databases

图 8 在不同程度扰乱数据集中,3 个评价指标与容错率大小关系图

接下来评估在精确数据集上,两种方法的查准率随容错率增加的变化情况.其中,$P$=5,$|D_i|$=5000,$DR$=0.如图 9 所示,方法 BA_PPRL 的查准率并不会随容错率的变化而变化.本文方法的查准率随容错率的增大而变小,但当 $X$=1 时,本文方法的查准率依然在 0.7 以上.由于数据集为精确的,所以方法 BA_PPRL 能准确地找全各参与方间的共有记录.随着容错率的增加,本文方法在找全各参与方间共有记录的同时还会误找一些记录认为匹配成功.因此,通过对比方法 BA_PPRL 与本文方法在精确数据集中查准率随容错率的变化情况,可以推断出本文方法的失误率.

继续评估在 Mod-1 数据集中,两种方法的 3 个评价指标随扰乱比例增加的变化情况.其中 $P$=5,$|D_i|$=5000,$X$=1/5.如图 10 所示,3 个评估指标均随扰乱比例的增加而下降.方法 BA_PPRL 的查全率大幅度下降,当 $DR$≥3/4,其查全率几乎为 0.本文方法的查全率即使在 $DR$=1 时,仍然在 0.4 以上.$F$ 指数可以说明本文方法在处理存在质量问题数据集时的有效性.

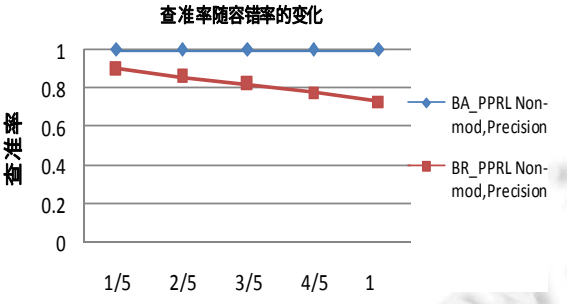


Fig.9 Precision with different $X$ in non-mod databases

图 9 在精确数据集中查准率与容错率大小关系图

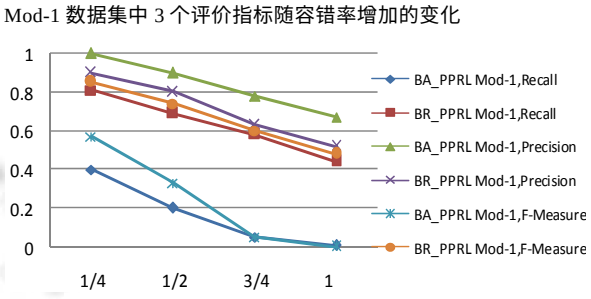


Fig.10 Three evaluation index with different $DR$ in mod-1 databases

图 10 在 Mod-1 数据集中,3 个评价指标与扰乱比例大小关系图

接下来评估在不同程度扰乱数据集上,本文提出方法的 3 个评价指标随扰乱比例增加的变化情况.其中,$P$=5,$|D_i|$=5000,$X$=1/5.如图 11 所示:随着扰乱比例的增加,本文方法的查全率、查准率和 $F$ 指数都在下降.但 Mod-3 的 $F$ 指数依然保持较好效果.

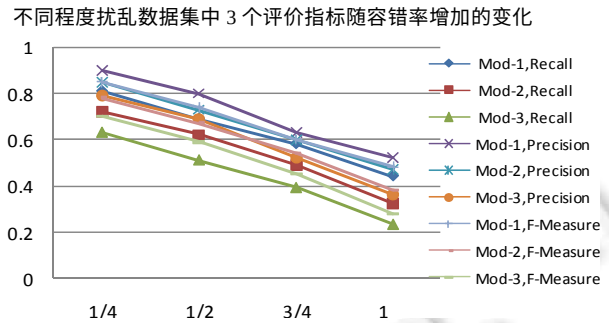


Fig.11 Three evaluation index with different *DR* in different mod databases

图 11 在不同程度扰乱数据集中,3 个评价指标与扰乱比例大小关系图

4.2.3 实验总结

通过实验可以得知:本文提出的方法在处理带有质量问题的数据时具有较高的查全率;且在有效的检查机制的保护下,查准率并未受到太大影响.在运行时间上,本文方法比 BA_PPRL 方法要长,但对比一些应用密码学保护技术的方法,时间代价仍然较低,具有较好的可扩展性.因此,本文提出的方法虽然牺牲了少部分时间代价,但在保证准确率的前提下,显著提高了查全率.

## 5 结束语

多方 PPRL 方法的研究是实现大数据共享的必然要求,具有广泛的应用价值.但目前的方法在处理带有质量问题的现实数据时,存在查全率低和时间代价大的特点.因此,本文提出了一种结合布隆过滤、安全合计、动态阈值、检查机制和改进的 Dice 相似度函数的多方近似 PPRL 方法.动态阈值和检查机制可以有效地选出存在质量问题的真实匹配的位置,并利用改进的 Dice 相似度函数计算相似度.本文方法可以有效减少由于数据质量问题带来的真实匹配记录的丢失,进而提高查全率.检查机制则可以保证查准率依然较高.本文方法具有良好的扩展性,并且在保证准确率的前提下显著提高了查全率,具有广泛而深远的现实意义.在未来的工作中,将进一步研究更优的多方隐私分块方法来提高本文方法的可扩展性.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是对东北大学计算机科学技术系申德荣教授、聂铁铮副教授及各位给予帮助的同学表示感谢.

**References**:

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

[2] Vatsalan D, Christen P. Scalable privacy-preserving record linkage for multiple databases. In: Proc. of the 23th Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2014. 1795−1798. [doi: 10.1145/2661829.2661875]

[3] Al-Lawati A, Lee D, McDaniel P. Blocking-Aware private record linkage. In: Proc. of the Int'l Conf. on IQIS. 2005. 59−68. [doi: 10.1145/1077501.1077513]

[4] Bonomi L, Xiong L, Chen R, Fung BCM. Frequent grams based embedding for privacy preserving record linkage. In: Proc. of the 21th Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2012. 1597−1601. [doi: 10.1145/2396761.2398480]

[5] Bonomi L, Xiong L, Chen R, Fung BCM. Privacy preserving record linkage via grams projections. Computer Science, 2012.

[6] Clifton C, Kantarcioglu M, Doan A, Schadow G, Vaidya J, Elmagarmid A, Suciu D. Privacy preserving data integration and sharing. In: Proc. of the 9th SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. ACM Press, 2004. 19−26. [doi: 10.1145/1008694.1008698]

[7] Inan A, Kantarcioglu M, Ghinita G, Bertino E. Private record matching using differential privacy. In: Proc. of the 13th Int'l Conf. on Extending Database Technology. ACM Press, 2010. 123−134. [doi: 10.1145/1739041.1739059]

[8] Kuzu M, Kantarcioglu M, Inan A, Bertino E, Durham E, Malin B. Efficient privacy-aware record integration. In: Proc. of the Int'l Conf. on Extending Database Technology. ACM Press, 2013. 167−178. [doi: 10.1145/2452376.2452398]

[9] Quantin C, Bouzelat H, Allaert FAA, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow-up: Quality assessment of an anonymous record linkage Procedure. The Int'l Journal of Medical Informatics, 1998,49(1):117−122. [doi: 10.1016/S1386-5056(98) 00019-7]

[10] O'Keefe CM, Yung M, Gu L, Baxter R. Privacy-Preserving data linkage protocols. In: Proc. of the Workshop on Privacy in the Electronic Society. ACM Press, 2004. 94−102. [doi: 10.1145/1029179.1029203]

[11] Kantarcioglu M, Jiang W, Malin B. A privacy-preserving framework for integrating person-specific databases. In: Proc. of the Int'l Conf. on Privacy in Statistical Databases. 2008. 298−314. [doi: 10.1007/978-3-540-87471-3_25]

[12] Mohammed N, Fung BCM, Debbabi M. Anonymity meets game theory: Secure data integration with malicious participants. The Int'l Journal on Very Large Data Bases, 2011,20(4):567−588. [doi: 10.1007/s00778-010-0214-6]

[13] Lai P, Yiu S, Chow K, Chong C, Hui L. An efficient Bloom filter based solution for multi-party private matching. In: Proc. of the Conf. on SAM. 2006.

[14] Schnell R, Bachteler T, Reiher J. Privacy preserving record linkage using Bloom filters. MIBM, 2009,9(1):41. [doi: 10.1186/1472-6947-9-41]

[15] Vatsalan D, Christen P, O'Keefe CM, Verykios VS. An evaluation framework for privacy-preserving record linkage. Journal of Privacy and Confidentiality, 2014,6(1):35−75.

[16] Christen P, Vatsalan D. Flexible and extensible generation and corruption of personal data. In: Proc. of the 23th Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2013. 1165−1168. [doi: 10.1145/2505515.2507815]

[17] Christen P. Data Matching-Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer-Verlag, 2012. [doi: 10.1007/978-3-642-31164-2]

韩姝敏(1991－),女,辽宁盖县人,CCF 学生会员,主要研究领域为实体识别,隐私保护.

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.

聂铁铮(1980－),男,博士,副教授,CCF 专业会员,主要研究领域为数据质量,数据集成.

寇月(1980－),女,博士,副教授,CCF 专业会员,主要研究领域为实体搜索,数据挖掘.

于戈(1962－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为数据库,大数据管理.