# 劣质数据库上阈值相似连接结果大小估计

张 岩 杨 龙 王宏志
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

**摘 要** 劣质数据普遍存在于现代数据管理系统中,严重影响了数据的质量,从而降低了数据的实用性以及数据的价值,这为数据管理带来了新的挑战.当前,已经有不少管理劣质数据的数据模型被提出,实体关系数据模型是其中一种,其中每条元组表示一个现实世界中的实体.该模型允许劣质数据的存在,给出了衡量数据质量的方法,并且可根据用户对结果质量的需求给出达到一定质量的查询结果.鉴于该模型的特点,传统的查询代价估计方法不再适用,需要新的代价估计技术.文中研究如何估计连接操作结果的大小,提出了在应用局部敏感 Hash 算法对属性值聚类的基础上,再进行采样估计的方法,并且在聚类过程中考虑数据质量对查询结果的影响.与传统随机采样方法对比,实验结果表明文中估计方法有更好的准确性.

**关键词** 代价估计;采样估计;劣质数据;数据质量;阈值
**中图法分类号** TP311 **DOI 号**: 10.3724/SP.J.1016.2012.02159

## Similarity Join Size Estimation with Threshold for Dirty Database

ZHANG Yan YANG Long WANG Hong-Zhi
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

**Abstract** Dirty data exists with large probability in modern data management systems, which affects the quality of the data, and determines data utility and data value. This brings new challenges for data management. Currently, many dirty data management models have been proposed, and one of them is entity-based relational database model in which one tuple represents a real-world entity. This model allows the existence of dirty data, and proposes the evaluation of data quality. It also can generate query results satisfying the quality requirements provided by users. With the features of the model, traditional query cost estimation models are not suitable for this model. Therefore, new cost estimation methods need to be developed. This paper focuses on the estimation of the result size of join operator and proposes a sampling-based algorithm based on the Locality Sensitive Hashing (LSH) to cluster similar objects. Compared with the traditional random sampling method, experimental results show that our method gives more accurate estimations.

**Keywords** result size estimation; sampled-data estimation; dirty data; data quality; threshold

## 1 引 言

数据质量[1]问题已经在很多研究领域引起了人们的广泛关注,如统计学领域、管理学领域以及计算机科学领域等.数据质量的好坏直接决定了数据的实用性和数据的价值,劣质数据是导致数据质量问题的主要原因.

收稿日期:2012-06-30;最终修改稿收到日期:2012-08-24.本课题得到国家"九七三"重点基础研究发展规划项目基金(2012CB316200)、国家自然科学基金(61003046,61033015,61133002)、RSE-NSFC 交流项目(61111130189)、教育部博士点基金(20102302120054)以及中央高校基本科研业务费转向资金(HIT.NSRIF.2013064)资助.**张 岩**,男,1965 年生,博士研究生,副教授,主要研究方向为数据质量、数据管理等.E-mail: zhangy@hit.edu.cn.**杨 龙**,男,1987 年生,硕士研究生,主要研究方向为数据质量、查询优化.**王宏志**(通信作者),男,1978 年生,博士,副教授,主要研究方向为 XML 数据管理、数据质量等.E-mail: wangzh@hit.edu.cn.

所谓劣质数据主要是指错误的、不确定的、不一致的以及重复的数据.研究表明,大部分数据库中都含有劣质数据.例如,有调查报告[2]指出超过65%的零售商库存记录数据库中的数据是不精确的.劣质数据给企业带来的后果是严重的.有错误、重复、不精确或者不一致的数据会导致错误的业务决策、不好的客户关系以及无效的营销策略等,从而给企业带来巨大损失.文献[3]指出每年美国各企业因为劣质数据所带来的损失高达6000亿美元.因此,需要有新的技术来处理劣质数据,降低劣质数据的危害.

目前,处理劣质数据的方法主要集中于数据清洗[4]和数据修复[5].数据清洗是指检测并移除数据库中的错误数据,从而提高数据质量.数据修复是指检测数据库中错误或不一致的数据,通过给定的规则(如条件函数依赖规则等)或者专家反馈来修复数据.然而,无论是数据清洗还是数据修复都存在一定的局限性.第一,多数情况下,数据清洗无法彻底除去劣质数据,而且如果清洗过度可能会导致信息的丢失;第二,数据修复过程中,不可能给出所有的规则,而且多条规则的修复结果可能会产生冲突.因此,数据清洗和数据修复均不能彻底除去数据库中的劣质数据,也就不能有效地解决劣质数据引起的问题.

基于上述原因,一些研究人员开始考虑如何在不经过数据清洗和数据修复的劣质数据上直接处理查询,从中得到正确、一致的查询结果或者满足质量要求的查询结果.当前,已经有部分工作开始了对劣质数据库(含有劣质数据的数据库)上查询处理技术的研究[6-8].但是这些工作中,大部分都是针对特定的查询或者特定方法展开的.因此,为了更好地处理劣质数据,需要统一的模型来管理劣质数据.

当前,已经有部分劣质数据模型被提出,其中应用最为广泛的是概率数据模型[9].概率数据模型可以表示非精确的数据,但这只是劣质数据的一种,而且从数据操作角度来看,这种模型无法描述数据操作对操作结果质量的影响,更重要的是,应用概率数据模型的数据库在执行查询的过程中,会产生所有的可能世界实例,这将导致数据规模的指数增长,从而降低查询处理效率.

本文考虑基于实体的关系数据模型[10],在这种模型中,每条元组表示一个实体,每个属性值是不确定的,可能含有多个值,对每个值用一个清洁度来表示该值的质量.采取这种表示方式的原因在于,在实际应用中,对同一个现实世界实体的不同表示方法是劣质数据的一个重要来源.尤其在数据融合[11](信息集成)过程中,将会导致出现不一致的、不确定的或者重复的数据.在实体关系数据模型中,表示同一个实体的数据合并为一条记录,为不确定或者不一致的属性值赋予一个概率值,表示该值的清洁度(或质量,本文中,清洁度即表示质量,不加以区别).例1描述了这种表示方法.

**例1.** 表1是一个劣质数据片段.通过实体识别算法[12-13],可以判断元组1、3和6表示同一个实体.因此,可以合并这3条元组,得到一条实体元组(如表2).在合并过程中不删除任何数据,因为,根据这个数据片段只能判断"Name"取值更可能为"Wal-Mart",但无法完全排除"Mal-Mart".因此,我们保留所有可能取值,即属性取值是不确定的.为了描述每个可能取值的质量,赋予每个值一个概率值,表示该值的清洁度.如"Wal-Mart"在3条元组中出现了2次,则清洁度为$2/3\approx0.67$,类似的,可以计算出每个可能值的清洁度,从而得到表2中的实体元组.

**表1 劣质数据片段**

| ID | Name | City | Zipcode | Phn | Reprsnt |
|---|---|---|---|---|---|
| 1 | Wal-Mart | Beijing | 90015 | 80103389 | Sham |
| 2 | Carrefour | Harbin | 20016 | 80374832 | Morgan |
| 3 | Wal-Mart | BJ | 90015 | 010-80103389 | Sham |
| 4 | Wal-mart | Harbin | 20040 | 70937485 | Sham |
| 5 | Carrefour | Beijing | 90015 | 83950321 | Morgan |
| 6 | Mal-Mart | Beijing | 90015 | 80103389 | Sham |

**表2 实体元组**

| ID | Name | City | Zipcode | Phn | Reprsnt |
|---|---|---|---|---|---|
| 1 | (Wal-Mart,0.67),<br>(Mal-Mart,0.33) | (Beijing,0.67),<br>(BJ,0.33) | (90015,1.0) | (80103389,0.67),<br>(010-80103389,0.33) | (Sham,1.0) |

实体关系数据模型通过实体识别和清洁度巧妙地表示了劣质数据.该模型利用了实体识别的结果,每条元组表示一个实体.因此,该模型中的所有查询操作的对象均是实体;相反地,概率数据模型中的查询操作针对的是所有的可能世界实例,而在所有可能世界实例中会有多个实例表示的是同一个实体.现实应用中,很多情况下,用户查询希望得到的是实体结果,而不是同一个实体的多种不同表示.例如,表2表示的是各地区超市基本信息,如果在该表中查找企业法人代表(Reprsnt)为"Sham"的超市,则

概率数据模型会产生并返回 8 个可能世界实例(如表 3 所示),但这 8 个实例实际表示的是同一个超市,而实体关系数据模型只会返回一个超市,显然这更符合查询的期望.因此,与概率数据模型相比,该模型在查询处理过程中不会产生所有的可能世界实例,也就不会导致数据规模的指数增长,从而提高了查询处理的效率.

**表 3　查询结果**

| ID | Name | City | Zipcode | Phn | Reprsnt |
|---|---|---|---|---|---|
| 1 | Wal-Mart | Beijing | 90015 | 80103389 | Sham |
| 2 | Wal-Mart | Beijing | 90015 | 010-80103389 | Sham |
| 3 | Wal-Mart | BJ | 90015 | 80103389 | Sham |
| 4 | Wal-Mart | BJ | 90015 | 010-80103389 | Sham |
| 5 | Mal-Mart | Beijing | 90015 | 80103389 | Sham |
| 6 | Mal-Mart | Beijing | 90015 | 010-80103389 | Sham |
| 7 | Mal-Mart | BJ | 90015 | 80103389 | Sham |
| 8 | Mal-Mart | BJ | 90015 | 010-80103389 | Sham |

考虑到实体关系数据模型的特点,每个属性值是不确定的,可能还有多个取值,而且都有对应的清洁度,现有的查询优化方法不再适用.因此,我们需要新的查询优化技术.作为查询优化的基础,如何准确估计一个数据操作的结果大小是十分重要的.考虑到连接操作是数据库中的重要操作之一,本文着重研究如何估计实体关系数据库中相似连接操作的结果大小.

当前已经有很多工作研究连接(主要是等值连接)结果大小的估计[14-16],主要是应用采样进行估计,如自适应采样(Adaptive sampling)[14]、双重采样(Bifocal sampling)[15]、交叉采样(Cross sampling)[16]等.其中部分方法可以应用于相似连接,但是无法保证采样结果适用于不同的相似度要求.例如,文献[17]中指出,对 DBLP 数据集合进行自连接时,如果相似度阈值设为 0.9,则连接结果的选择率只有 0.00001%.这种情况下,上述采样方法很难抽取到可以连接的元组对,从而无法给出准确估计.针对相似连接估计算法[17-18]也已经有部分工作,但这些算法相对于实体关系数据库均不太适用.主要原因有两点.第一,现有算法中相似连接的属性值是确定的,而实体关系数据库中属性可能包含多个值,是不确定的;第二,实体关系数据库中引入了清洁度的概念,而现有算法中均没有这一概念.

基于上述分析,本文提出了新的相似连接结果大小估计算法.考虑到如果先对相似元组加以聚类,然后在聚类集中采样可以提高采样的质量.因为即使相似度阈值较高,在聚类集中也能以很大概率采样得到满足相似度要求的元组对.因此,本文首先应用局部敏感 Hash 算法(Locality Sensitive Hashing,LSH)对不确定属性值进行聚类,然后在聚类集中采样估计相似连接的结果大小.鉴于实体关系数据库的特点,本文提出了基于清洁度的局部敏感 Hash 聚类算法,在聚类过程中,充分考虑了实体关系数据模型中属性取值的不确定性以及不同取值的清洁度的影响.本文的主要贡献如下:

(1)提出了一种适用于实体关系数据库的基于清洁度的局部敏感 Hash 聚类算法.该算法可以把相似的不确定属性值聚集在一起.

(2)提出了一种基于相似聚类的采样估计算法,能够更加准确地估计相似连接的结果大小.

(3)通过实验证明,本文的估计算法可以得到比现有方法更高的估计准确率.

本文第 2 节介绍实体关系数据模型以及相似连接的相关概念;第 3 节给出基于采样的相似连接估计算法的框架;第 4 节提出基于清洁度的局部敏感 Hash 聚类算法;第 5 节给出实验结果和分析;第 6 节对本文进行总结.

## 2　相关定义

### 2.1　实体关系数据模型

第 1 节中已经简单介绍了实体关系数据模型,本节中我们通过几个概念定义该模型.

首先给出不确定属性值的概念,正如第 1 节介绍的,在实体关系数据模型中,属性值是不确定的,包含多个可能值.定义 1 给出了不确定属性值的定义,它不仅包含所有的可能值,也包含其对应的清洁度,而且一个不确定属性值的所有可能值的清洁度之和为 1.然后,我们给出了实体的定义,实体是实体关系数据模型中的基本存储单位,是多个不确定的属性值的集合.

**定义 1.**　不确定属性值.一个不确定的属性值是一个集合 $A=\{(v,p)\mid v$ 是某个可能的取值,$p$ 是该可能值对应的清洁度$\}$.

**定义 2.**　实体.一个实体是一条元组,可以表示为一个二元组 $E=(K,A)$,其中 $A$ 是不确定属性值的集合,$K$ 是主键集合,它可以唯一确定一个实体(如实体 ID 号).

表 2 给出了一条实体元组,包含主键 ID 和 5 个不确定属性值,每个不确定属性值都包含一个或多个可能取值,而且多个可能取值对应的清洁度之和为 1.

有了这些概念,我们可以给出实体关系数据库上的相关操作的定义.

### 2.2 阈值相似连接

本文主要考虑实体关系数据库中的相似连接操作,估计其结果的大小.

相似连接是在两个元组集合中选择出满足相似性下界的元组对.当前已经有很多衡量相似性的方法,应用较广泛的有:编辑距离(Edit distance)、Hamming 距离(Hamming distance)、杰卡德相似度(Jaccard similarity)以及余弦相似度(Cosine similarity)等.本文考虑的是基于编辑距离的相似连接.对于给定的两个字符串 $s$ 和 $t$,$s$ 和 $t$ 的编辑距离 $ed(s,t)$是指从 $s$ 变为 $t$ 所需要的最少的编辑操作(插入、删除和替换)次数.

考虑到实体关系数据库中,属性值都是不确定的,可能含有多个取值.本文给出相似连接定义如下.

**定义 3.** 相似连接.给定两个不确定属性值集合 $R$ 和 $S$ 以及编辑距离阈值 $k$,$R$ 和 $S$ 的相似连接是指选出所有的属性值对$(r,s)$,其中 $r\in R$,$s\in S$,而且 $r$ 和 $s$ 满足至少存在一组可能取值 $v_{r_i}$ 和 $v_{s_j}$,它们的编辑距离 $ed(v_{r_i},v_{s_j})\leq k$,即$\{(r,s)\mid r\in R,s\in S,\exists v_{r_i}\in r,\exists v_{s_j}\in s,\text{s.t. } ed(v_{r_i},v_{s_j})\leq k\}$,其中 $v_{r_i}$ 表示不确定属性值 $r$ 的第 $i$ 个可能取值.

**例 2.** 表 4 给出了一组不确定属性值集合 $R$ 和 $S$.若编辑距离阈值 $k$ 设为 3,则相似连接结果为 $\{(r_1,s_1),(r_2,s_2)\}$.因为$ed(\text{Wal-Mart},\text{Wal-mart})=1\leq 3$,$ed(\text{John Strauss},\text{John Strauss})=0\leq 3$.表 5 给出了连接结果.

**表 4 两个集合 $R$ 和 $S$**

| 集合 $R$ | | 集合 $S$ | |
|---|---|---|---|
| $r_1$ | (Wal-Mart,0.67),(Mal-Mart,0.33) | $s_1$ | (Wal-Mart,1.0) |
| $r_2$ | (John Smith,0.8),(John Strauss,0.2) | $s_2$ | (John Strauss,0.6),(Johann Strauss,0.4) |

**表 5 集合 $R$ 和 $S$ 的相似连接中间结果**

| *ID* | 集合 $R$ | 集合 $S$ | 清洁度 |
|---|---|---|---|
| 1 | (Wal-Mart,0.67),(Mal-Mart,0.33) | (Wal-Mart,1.0) | 1.0 |
| 2 | (John Smith,0.8),(John Strauss,0.2) | (John Strauss,0.6),(Johann Strauss,0.4) | 0.2 |

由于实体关系数据库中,所有可能值都有对应的清洁度,所以,所有相似连接结果也都会有一个清洁度,表示该结果的质量(反映了结果的价值),如表 5 所示.实际应用中,多数情况下,我们只对清洁度比较高的结果感兴趣,对于较低清洁度的结果可以忽略.在本文中,只考虑达到一定清洁度要求(不低于某个阈值 $\theta$)的连接结果,本文称这样的连接为阈值相似连接,其定义如下.

**定义 4.** 阈值相似连接.给定两个不确定属性值集合 $R$ 和 $S$,清洁度阈值 $\theta$ 和编辑距离阈值 $k$,$R$ 和 $S$ 的阈值相似连接是指选出所有属性值对$(r,s)$,$r$ 和 $s$ 连接结果的清洁度不低于 $\theta$.即阈值相似连接结果集合可表示为$\left\{(r,s)\mid \sum_{ed(v_{r_i},v_{s_j})\leq k} p_{r_i}*p_{s_j}\geq\theta\right\}$,其中 $p_{r_i}$ 代表不确定属性值 $r$ 的第 $i$ 个可能取值对应的清洁度.

例如,如果清洁度阈值 $\theta$ 设为 0.3,那么表 4 中的两个集合 $R$ 和 $S$ 的阈值相似连接结果为 $\{(r_1,s_1)\}$.$(r_2,s_2)$不再属于连接结果,因为对于 $(r_2,s_2)$有 $\sum_{ed(v_{r_i},v_{s_j})\leq k} p_{r_i}*p_{s_j}=0.2*0.6+0.2*0.4=0.2<0.3$,即该结果不满足清洁度阈值要求.

本文主要考虑阈值相似连接,估计其连接结果的大小,下一节给出估计算法的基本框架.

## 3 估计算法框架

本节给出估计算法的框架描述.

阈值相似连接的特点是当相似度阈值较小(即编辑距离阈值较大)而且清洁度阈值也较小时,连接结果大小会接近于 $n^2$($n$ 是连接集合的大小),反之,连接结果集比较小.

当前已有的连接结果大小的估计方法主要是基于采样的方法.这些方法中部分可以应用于本文所考虑的阈值相似连接估计,但在应用到这个问题时有一个显著缺陷:只适用于阈值比较小,连接结果集合较大的情况.当阈值较大,连接结果集合较小时,就无法保证估计效果.

针对一般采样的缺陷,文献[17]提出了一种基于局部敏感 Hash(Locality Sensitive Hashing,LSH)的采样方法用来估计相似自连接结果大小.文献[17]主要解决向量(或集合)相似连接问题,其主要思想是应用局部敏感 Hash 算法把相似的向量(或集合)Hashing 到相同的桶中,然后分两种情况进行采样.一是从相同的桶中进行采样估计结果大小;二是从不同的桶中进行采样估计结果大小,两者结合可得到较准确的估计值.在这种方法中,即使取较大的相似度阈值,从同一个桶中的采样满足阈值的可能性也比较大,因为同一个桶中的对象都是比较相似的,从而可以保证估计的效果.

虽然这种方法可以提高相似度阈值较高时采样方法的估计准确度,但是该方法只适用于属性值为

单个向量(或集合)时的相似连接估计.因此,对于本文考虑的实体关系数据库中阈值相似连接不太适用.这主要体现在三个方面.一是在实体关系数据库中,每个属性值有多个可能取值,在聚类过程中,要同时考虑多个取值的影响;二是实体关系数据模型中增加了数据的清洁度,那么即使有相同的可能取值,但是对应的清洁度不同,也会产生不同的连接结果;三是文献[17]中的方法针对的是相似自连接估计,而且其采样方法可能会产生重复的采样.

考虑到基于聚类采样的方法可以解决相似度较高时,一般采样方法的缺陷,本文的估计算法正是以此为基本框架.为了解决其不足,本文提出了基于清洁度的局部敏感 Hash 聚类算法.该算法不仅考虑了每个属性值的多个可能取值,而且还考虑了不同取值对应清洁度的影响,例 3 展示了该算法的优点.其次,为了避免重复采样,本文采用无重复随机采样方法.

**例 3.**　考虑两个不确定属性值:$t_1$:{(Robert, 0.9), (Bob, 0.1)} 和 $t_2$:{(Robert, 0.1), (Bob, 0.9)}.如果应用文献[17]中的方法分别考虑所有取值,那么 $t_1$ 和 $t_2$ 将会分别被计数 2 次,而本文算法将会综合多个取值,对 $t_1$ 和 $t_2$ 只计数 1 次.其次,文献[18]中的方法无法反映不同取值清洁度的影响,则无法区别 $t_1$ 和 $t_2$.而实际上,如果给定 $t_3$:{(Robert,1)},编辑距离阈值为 2,清洁度阈值为 0.3,则仅($t_1$,$t_3$)满足连接条件,($t_2$,$t_3$)并不满足.所以,$t_1$ 和 $t_2$ 仅取值相同,而实际代表的属性值并不相同,本文的算法则可以在聚类过程中区别 $t_1$ 和 $t_2$.

算法 1 给出了阈值相似连接结果大小估计算法的基本框架.

**算法 1.**　JoinEstimation.

步骤 1. 数据预处理

输入:不确定属性值集合 $R$ 和 $S$

输出:集合 $R$ 和 $S$ 的聚类结果集合

//$P$、$C$ 分别表示相似对集合、聚类集合

1. $P_R$←LSH_Quality($R$) /*返回 $R$ 中所有的相似对*/
2. $P_S$←LSH_Quality($S$)
3. $C_R$←*Clustering*($P_R$) /*返回 $R$ 的聚类结果集合*/
4. $C_S$←*Clustering*($P_S$)

步骤 2. 采样估计

输入:不确定属性值集合 $R$ 和 $S$ 的聚类结果,相似度阈值 $\tau$,清洁度阈值 $\theta$

输出:集合 $R$ 和 $S$ 的阈值相似连接结果大小估计值 $N$

// $S_R$、$S_S$ 表示采样集合

1. $S_R$←*Sampling*($C_R$)
2. $S_S$←*Sampling*($C_S$)
3. $N=0$
4. FOR $i$ FROM 1 TO $|S_R|$ /*$S_R$ 中每一个采样集合*/
5. FOR $j$ FROM 1 TO $|S_S|$ /*$S_S$ 中每一个采样集合*/
6. $n=0$
7. FOR EACH $pair(r,s)$,其中 $r\in S_{Ri}$,$s\in S_{Sj}$
8. IF $\sum_{sim(v_{r_i},v_{s_j})\geq\tau} p_{r_i} * p_{s_j} \geq\theta$ THEN
9. $n++$
10. END FOR
11. $N+=|C_{Ri}|*|C_{Sj}|*n/(|S_{Ri}|*|S_{Sj}|)$
12. END FOR
13. END FOR
14. RETURN $N$

框架分为两个部分:首先对数据集进行预处理,把相似的不确定属性值聚集在一起;然后再从聚类集中采样估计连接结果大小.

采样函数 *Sampling*( )采用随机采样,从每个聚类中随机采样组成样本集合,在该样本集合上做阈值相似连接,估计连接结果大小.即,如果 $R$ 中第 $i$ 个聚类 $C_{Ri}$ 的采样集合 $S_{Ri}$ 与 $S$ 中第 $j$ 个聚类 $C_{Sj}$ 的采样集合 $S_{Sj}$ 连接结果大小为 $n$,则 $C_{Ri}$ 和 $C_{Sj}$ 连接结果大小估计值为 $|C_{Ri}|*|C_{Sj}|*n/(|S_{Ri}|*|S_{Sj}|)$.分别估计不同聚类连接结果大小,相加即可得到整个集合连接结果大小的估计值.

对于聚类函数 *Clustering*( ),本文以局部敏感 Hash 算法为基础提出了基于清洁度的局部敏感 Hash 聚类算法,下一节我们给出该算法的详细介绍.

## 4　基于清洁度的局部敏感 Hash 聚类

局部敏感 Hash 聚类算法是本文估计方法的核心,本节讨论该聚类算法,通过该算法计算出所有的相似对,然后应用该相似对结果进行聚类.首先我们简单介绍局部敏感 Hash 算法,然后讨论如何在实体关系数据库中应用该算法,同时考虑所有的可能取值以及清洁度的影响,最后给出聚类算法.

### 4.1　局部敏感 Hash 算法

局部敏感 Hash 算法最早由 Indyk 和 Motwani 提出[19],用以解决高维空间的 $k$-NN($k$ Nearest Neighbor)问题.算法的主要思想是:对于给定的高维向量集合,选择一族局部敏感 Hash 函数,如果两个向量夹角越小(越相似),那么它们对应的 Hash 值就越有可能相近.从而可以根据向量的 Hash 值来判断向量是否相似.文献[20]对局部敏感 Hash 函数做了详细描述,并给出定义如下.

**定义 5.**　局部敏感 Hash 函数.给定一族 Hash

函数 $H$，如果对于任意两个向量 $\boldsymbol{u},\boldsymbol{v}\in R^m$，$h\in H$ 都满足

$$P(h_r(\boldsymbol{u})=h_r(\boldsymbol{v}))=sim(\boldsymbol{u},\boldsymbol{v}) \quad (1)$$

则称其为一族局部敏感 Hash 函数，其中 $sim(\boldsymbol{u},\boldsymbol{v})$ 为相似性衡量函数.

对于给定的 $m$ 维向量集合，文献[20]给出了一种选择敏感 Hash 函数的方法：随机选择一个 $m$ 维向量 $\boldsymbol{v}$，定义敏感 Hash 函数如下：

$$h_r(\boldsymbol{u})=\begin{cases}1, & \boldsymbol{v}\cdot\boldsymbol{u}\geq 0\\ 0, & \boldsymbol{v}\cdot\boldsymbol{u}<0\end{cases}.$$

然后对于给定两个向量 $\boldsymbol{v}$ 和 $\boldsymbol{u}$，有如下公式：

$$P(h_r(\boldsymbol{u})=h_r(\boldsymbol{v}))=1-\frac{\theta(\boldsymbol{u},\boldsymbol{v})}{\pi} \quad (2)$$

如果取相似度函数为余弦相似度，那么夹角越小，余弦值越大，两个向量就越相似. 因此，根据等式(1)和(2)，该 Hash 函数符合定义 5. Goemans 和 Williamson 在文献[21]给出了等式(1)的证明.

为了提高 Hash 函数的准确率，在实际应用中，一般选取多个随机向量，然后分别应用 Hash 函数得到 0/1 值，从而可以把一个高维向量转化为一个 0/1 串. 即如果选取 $d(d\ll m)$ 个随机向量，那么每个 $m$ 维向量就转化为长度为 $d$ 的 0/1 串. 根据文献[22]，有如下公式：

$$P(h_r(\boldsymbol{u})=h_r(\boldsymbol{v}))=1-\frac{hamming\_distance(\boldsymbol{u},\boldsymbol{v})}{d} \quad (3)$$

其中 $hamming\_distance(\boldsymbol{u},\boldsymbol{v})$ 表示 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 的 Hamming 距离. 两个 0/1 串的 Hamming 距离等于两者相同位置不同取值的个数. 如 1011101 和 1101111 的 Hamming 距离为 3. 根据等式(1)和(3)可得

$$sim(\boldsymbol{u},\boldsymbol{v})=1-\frac{hamming\_distance(\boldsymbol{u},\boldsymbol{v})}{d} \quad (4)$$

因此，根据式(4)，可以把判断两个向量是否相似的问题转化为计算两个向量对应 0/1 串的海明距离. 如果海明距离小于某个阈值 $l$，则两个向量相似，反之，则不相似. 而计算两个 0/1 串的海明距离是快速高效的，因此，通过局部敏感 Hash 算法，可以快速计算出所有的相似向量对.

在下一节中，将给出如何在考虑多个可能值和其对应清洁度的同时，应用局部敏感 Hash 算法计算所有不确定属性值相似对.

## 4.2 基于清洁度的 Hash 算法

4.1 节已经介绍了局部敏感 Hash 算法，但该算法是针对高维向量的，如果要在本文的问题中应用该算法，需要将每个不确定属性值转换成一个向量，同时还要考虑不同取值清洁度的影响. 本节主要讨论如何在考虑不同取值的清洁度的影响下，把一个不确定属性转换成一个向量，并给出基于此向量的局部敏感 Hash 算法.

### 4.2.1 不确定属性值的向量化

本文在计算相似连接时，应用编辑距离来衡量相似度. 现有基于编辑距离的相似连接方法都是基于 $q$-gram 做的，$q$-gram 是字符串中长度为 $q$ 的连续子串，$q$-gram 集合是一个字符串中所有 $q$-gram 的集合. 显然如果两字符串之间的编辑距离越小，那么两个字符串共有的 $q$-gram 一定越多. 例如：给定两个字符串 $s_1$ 和 $s_2$，如果 $ed(s_1,s_2)\leq k$，那么 $s_1$ 和 $s_2$ 至少公共的 $q$-gram 数为

$$\max(|s_1|,|s_2|)+1-(k+1)q \quad (5)$$

基于以上分析，实体关系数据库中的所有可能取值都有其对应的 $q$-gram 集合，每个不确定属性值也有其对应的 $q$-gram 集合(即该不确定属性值的所有可能取值的对应 $q$-gram 集合的并集). 表 6 给出了一个实例. 需要注意的是：在这些集合中，我们允许元素重复出现，如表 6 中集合 $S$ 含有两个“ob”.

**表 6 不确定属性值的 2-gram 集合**

| 属性值 | 对应 2-gram 集合 |
|---|---|
| Robert | $S_1$:{Ro,ob,be,er,rt} |
| Bob | $S_2$:{Bo,ob} |
| {(Robert,0.6),(Bob,0.4)} | $S=S_1\cup S_2$:{Ro,ob,be,er,rt,Bo,ob} |

由式(5)可得，如果两个字符串 $s_1$ 和 $s_2$ 相似($ed(s_1,s_2)\leq k$)，则至少含有 $\max(|s_1|,|s_2|)+1-(k+1)q$ 个公共 $q$-gram，表示为 $L(s_1,s_2)$. 再考虑到实体关系数据模型中相似连接(定义 3)的特点：如果元组对$(r,s)$越相似，则它们对应可能取值的相似对$(r_i,s_j)$就越多. 由此，可以得到如下性质.

**性质 1.** 如果元组对$(r,s)$对应可能取值的相似对数量为 $m$，那么 $r$ 和 $s$ 对应的 $q$-gram 集合的公共 $q$-gram 数至少为 $mL(r_i,s_j)$. 从而元组对$(r,s)$越相似，对应的 $q$-gram 集合的公共 $q$-gram 数越多，对应的高维向量也应该越相似.

根据性质 1，假如对每个不确定属性值所有的 $q$-gram 进行计数，那么每个属性值便可以转换成一个高维向量，其中每一维对应一个不同的 $q$-gram，向量的维度等于不确定属性值集合的所有不同的 $q$-gram 的种类数，本文称该过程为不确定属性值的向量化. 需要注意的是，在实际向量化过程中，计数所有 $q$-gram 的代价较大，所以向量的维数会小于不同的 $q$-gram 的种类数. 图 1 展示了一个向量化过程，其中不确定属性值为{(Robert,0.6),(Bob,0.4)}，其向量化结果为(3,0,2,2,0).

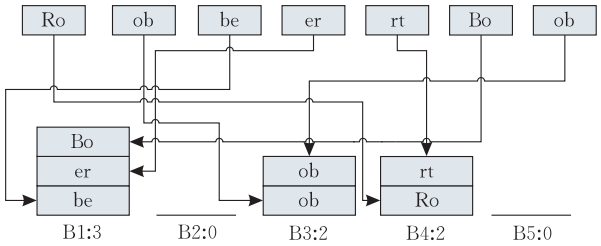


图 1　不确定属性值的向量化

但是上述向量化过程中没有考虑清洁度的影响，例如对于两个不确定属性值：{(Robert，0.9)，(Bob，0.1)}和{(Robert，0.1)，(Bob，0.9)}，上述方法向量化结果均为(3，0，2，2，0)，无法区别两个值.为了处理这个情况，考虑阈值相似连接(定义 4)的特点，我们可以得到另一个性质.

**性质 2.**　如果元组对$(r,s)$对应清洁度为 $p$，由于 $p=\sum_{ed(v_{r_i},v_{s_j})\leq k} p_{r_i} * p_{s_j}$，所以 $p$ 越大，对应的 $p_{r_i}$ 和 $p_{s_j}$ 也越大，即元组对$(r,s)$对应清洁度越大，其公共 $q$-gram 所在的对应值的清洁度也会越大.

根据性质 2，我们可以对不确定属性值的向量化进行改进，加入清洁度的影响，使得到的向量能更准确地代表其对应的属性值.只需要在向量化的过程中，对每一个 $q$-gram 乘以其所在取值的清洁度再进行计数，这样便可以在向量化过程中反映出不同取值的清洁度.如图 2 所示，{(Robert，0.9)，(Bob，0.1)}向量化值为(1.9，0，1，1.8，0)，而类似的，{(Robert，0.1)，(Bob，0.9)}向量化值为(1.1，0，1，0.2，0).可以看出，该向量化方法能更准确地反映不确定属性值.

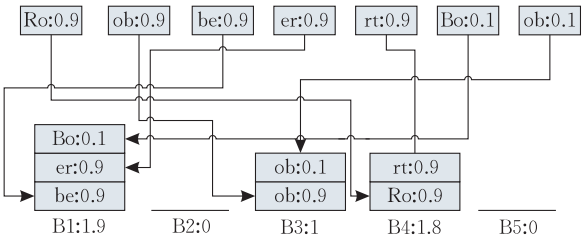


图 2　不确定属性值的基于清洁度的向量化

### 4.2.2　详细算法

由 4.2.1 节中的方法，可以把所有不确定属性值向量化.在此基础上，由 4.1 节中的介绍，我们给出详细算法，计算出所有相似对.算法 2 描述了这一过程.

**算法 2.**　LSH_Quality.

输入：不确定属性值集合 $S$ 和 Hamming 距离阈值 $l$

输出：不确定属性值集合 $S$ 中所有的相似对 $P_S$

```
1.  随机 d 个 m 维向量{v_1,…,v_d}
2.  FOR EACH s∈S DO
3.      v←vectorize(s) /*把 s 向量化结果为 v*/
4.      FOR EACH v_i∈{v_1,…,v_d}
5.          u_i←h_{v_i}(u) /*局部敏感 Hash 函数*/
6.      END FOR /*向量 v 转化为长度为 d 的 0/1 串*/
7.  END FOR
8.  FOR EACH pair(u_i,u_j) do
9.      IF hamming_distance(u_i,u_j)<l THEN
10.         P_S←(u_i,u_j)
11. END FOR
12. RETURN P_S
```

函数 $vectorize(s)$用来将 $s$ 向量化，返回一个 $m$ 维向量.该过程是对 $q$-gram 集合中不同 $q$-gram 计数的过程，假定属性值对应的字符串长度均在一定范围内，则函数可在常数时间内完成，时间复杂度为 $O(1)$.函数 $hamming_distance(\boldsymbol{u}_i,\boldsymbol{u}_j)$计算 $\boldsymbol{u}_i$ 和 $\boldsymbol{u}_j$ 的海明距离，由于每个不确定属性值均转换成长度为 $d$ 的 0/1，而且 $d\ll m$，所以该函数时间复杂度也是 $O(1)$.因此算法 2 的时间复杂度为 $O(n)+O(n^2)$，即 $O(n^2)$.

## 4.3　聚类算法

通过第 4.2 节中的算法，可以得到所有相似对集合.本节给出聚类算法，应用相似对集合把所有相似的不确定属性值聚集在一起.从而可以在每个聚类集上进行采样，估计阈值相似连接的结果大小.

考虑到已经得到了所有的相似对，如果把每个不确定属性值看作一个顶点，每个相似对看作顶点之间的边，即如果两个属性值相似，那么其对应的顶点之间有一条边.经过上述变换，可以把不确定属性值集合转化为一个图，该图反映了集合中所有不确定属性值之间的相似关系，从而把相似属性值聚类问题转化为图聚类(图聚集)或者社区发现问题.

当前已经有很多工作研究图聚类问题[23]和社区发现问题[24-25]，本文中，我们采用了社区发现中应用最为广泛的 CNM 算法进行聚类.该算法是基于模块度划分聚类的.模块度被用来衡量一个划分的好坏.通常一个好的划分会使聚类内部的边较多，而聚类之间的边较少.CNM 算法中定义模块度如下：

$$Q=\sum_i (e_{ij}-a_i^2),$$

其中，$e_{ij}$ 表示为连接聚类 $i$ 和聚类 $j$ 的点的边数所占的比例，$a_i$ 表示聚类 $i$ 内部的边所占的比例.

该算法的主要思想是：初始化聚类，不断合并聚类，使其沿着模块度增大最多或者减少最小的方向合并，直到最终合并为一个聚类.该过程结束后会得

到一个可分解的聚类结构树状图，对该聚类结构树状图的不同断开方式对应不同聚类划分，则其中对应最大模块度的聚类结构树状图断开方式就是最好的图的划分，即最好聚类结果. 文献[25]给出了该算法的详细描述.

通过 CNM 算法得到聚类结果后，我们就可以应用第 3 节中的方法进行采样并估计阈值相似连接结果的大小.

## 5 实 验

为了验证本文算法的估计效果，我们在Windows7操作系统上用 C++实现了全部算法. 实验运行的硬件环境为因特尔酷睿 2 处理器，主频为 2.93 GHz，内存 2 GB. 软件开发环境为 Code::Blocks 10.05.

### 5.1 实验数据以及评估方法

为了充分验证本文方法的估计效果，我们的实验数据集合来自实体识别的结果.

数据集合：首先我们采用数据生成系统生成了 2 个含有 50 K 个不同实体的数据集合，其中所有的属性值均是长度不超过 32 的随机字符串. 通过实体识别并合并表示同一个实体的数据，得到符合本文数据模型的两个实体数据集合. 为了方便实验，我们选取包含 $p$ 个可能值的不确定属性值作为实验数据集合 $R$ 和 $S$，其中 $p\in\{1,2,3,4,5\}$.

评估方法：本文采用平均相对误差来评估本文估计方法的效果. 相对误差可用如下公式计算：

$$\frac{|\hat{C}-C|}{\hat{C}},$$

其中 $\hat{C}$ 为阈值相似连接结果大小的准确值，$C$ 为本文算法的估计值.

### 5.2 实验结果及分析

在本文实验中，我们对比了本文的方法(LSH-S)和一般随机采样方法(RS). 表 7 给出了实验中所涉及的参数默认值. 实验基于编辑距离来衡量相似性，相似性计算公式如下：

$$sim(s,t)=1-\frac{ed(s,t)}{\max(|s|,|t|)}.$$

**表 7 参数默认取值**

| 参数 | 默认值 |
|---|---|
| 相似度阈值 $\tau$ | 0.5 |
| 编辑距离阈值 $k$ | 2 |
| 清洁度阈值 $\theta$ | 0.3 |
| 向量化维度 | 729 |
| Hash 函数数量 $d$ | 50 |
| Hamming 距离阈值 $l$ | 12 |

#### 5.2.1 算法效率实验

表 8 给出了实际阈值相似连接时间与本文采样估计时间的对比. 需要注意的是：在实验过程中，两者只采用了简单的优化过滤操作，但这并不影响估计的准确度. 由表 8 中数据可以看出采样估计时间远远小于实际连接操作执行时间，均不超过 1 s，相对实际连接操作执行时间可以忽略.

**表 8 实际操作执行时间和采样估计时间对比**

| 数据集 | 采样估计时间/s | 实际操作时间/s | 两者比值/% |
|---|---|---|---|
| 5K | 0.091 | 3.964 | 2.29 |
| 8K | 0.172 | 10.794 | 1.59 |
| 10K | 0.228 | 17.075 | 1.33 |
| 20K | 0.819 | 70.723 | 1.16 |

#### 5.2.2 数据集合大小的影响

本实验中，我们考虑数据集合大小对算法估计效果的影响. 我们从实验数据集合 $R$ 和 $S$ 中选取适当数量的不确定属性值组成不同大小的数据集合，包含不确定属性值的数量分别为 5K，8K，10K，20K，50K. 由图 3 可以看出，在采样比例不变的情况下，两种算法均在数据集合较大时，估计值更加准确. 但相比之下，本文算法更加稳定，受数据集合大小影响较小.

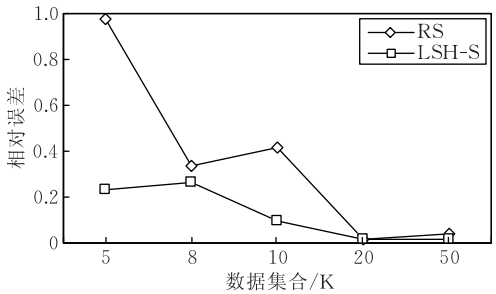


图 3 数据集合大小的影响

#### 5.2.3 相似度阈值大小的影响

相似度阈值是相似连接中一个重要的参数. 不同的相似度阈值会产生不同大小的结果集，阈值越小，连接结果越多，极端情况下，当阈值趋近于 0 时，连接结果趋近于 $n^2$. 为了测试相似度阈值的影响，我们分别取相似度为 0.1～0.9 进行实验. 图 4 给出

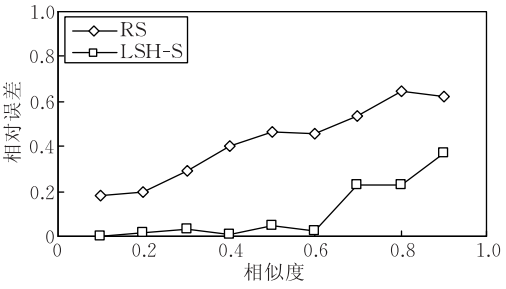


图 4 相似度阈值的影响

了实验结果.两种算法都在阈值较小时更加精确,随着阈值的变大,估计误差逐渐变大,但是很明显本文算法相对更加精确.

5.2.4　清洁度阈值大小的影响

清洁度是本文所考虑的实体关系数据模型特有的参数,反映了连接结果的质量.质量越高,说明结果的价值越大,反之,价值越小,小于一定阈值可忽略不计.我们变化清洁度阈值从 0.1～0.9 来测试它对估计算法的影响.从图 5 中可以看出,随机采样算法随着阈值增大,误差逐渐变大,而本文算法则受清洁度变化的影响较小.这主要是因为本文算法在局部敏感 Hash 过程中,采用的是基于清洁度的 Hash 算法,充分考虑了清洁度在阈值相似连接过程中的影响.

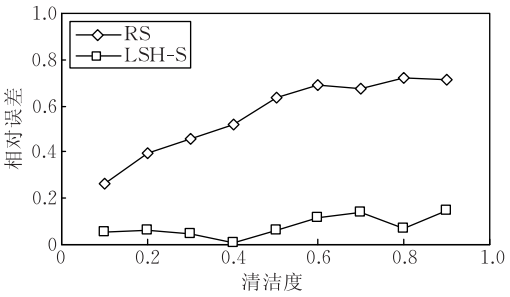


图 5　清洁度阈值的影响

## 6　总　结

实体关系数据模型是一种有效的管理劣质数据的模型.中间结果大小估计是查询优化过程中重要的方法.考虑到传统估计方法已不再适用于实体关系数据库,需要新的技术来估计中间结果的大小.本文针对阈值相似连接操作,给出了新的基于采样的估计方法.该方法首先通过应用局部敏感 Hash 算法对相似对象进行聚类,再在聚类结果中进行采样估计.实验结果表明,与随机采样方法相比,本文方法能给出更加精确的估计结果.未来工作包括在更多真实数据上进行进一步的实验以及将本文的方法应用到实体识别当中去.

## 参　考　文　献

[1] Batini C, Scannapieca M. Data Quality: Concepts, Methodologies and Techniques. New York: Springer-Verlag New York Inc, 2006

[2] Raman A, DeHoratius N, Ton Z. Execution: The missing link in retail operations. California Management Review, 2001, 43(2): 136-152

[3] English L. Information quality management: The next frontier. Information Management Magazine, 2000, 6(4): 17-24

[4] Rahm E, Do H H. Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 2000, 23(4): 3-13

[5] Fan W, Li J, Ma S, Tang N, Yu W. Interaction between record matching and data repairing//Proceedings of the 30th ACM Special Interest Group on Management of Data. Athens, Greece, 2011: 469-480

[6] Khalefa M E, Mokbel M F, Levandoski J J. Skyline query processing for incomplete data//Proceedings of the IEEE 24th International Conference on Data Engineering. Cancun, Mexico, 2008: 556-565

[7] Andritsos P, Fuxman A, Miller R. Clean answers over dirty databases: A probabilistic approach//Proceedings of the 22nd International Conference on Data Engineering. Atlanta, USA, 2006: 30-42

[8] Razniewski S, Nutt W C. Completeness of queries over incomplete databases//Proceedings of the Very Large Databases Endowment. Westin, USA, 2011: 749-760

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

[10] Zhang Y, Yang L, Wang H. Range query estimation for dirty data management system//Proceedings of the 13th International Conference on Web-Age Information Management. Harbin, China, 2012: 152-164

[11] Xin Luna Dong, Alon Halevy, Cong Yu. Data integration with uncertainty. The International Journal on Very Large Data Bases, 2009, 18(2): 469-500

[12] Whang S, Menestrina D, Koutrika G et al. Entity resolution with iterative blocking//Proceedings of the 35th SIGMOD International Conference on Management of Data. New York, USA, 2009: 219-232

[13] Benjelloun O, Garcia-Molina H, Menestrina D, Whang S E, Su Q, Widom J. Swoosh: A generic approach to entity resolution. The International Journal on Very Large Data Bases, 2008, 18(1): 255-276

[14] Lipton R J, Naughton J F, Schneider D A. Practical selectivity estimation through adaptive sampling//Proceedings of the 1990 SIGMOD International Conference on Management of Data. Atlantic, USA, 1990, 19(2): 1-11

[15] Ganguly S, Gibbons P B, Matias Y, Silberschatz A. Bifocal sampling for skew-resistant join size estimation//Proceedings of the 1996 SIGMOD International Conference on Management of Data. Montreal, Canada, 1996: 271-281

[16] Haas P, Naughton J, Seshadri S et al. Fixed-precision estimation of join selectivity//Proceedings of the 16th ACM SIGMOD Symposium on Principles of Database Systems. Washington, USA, 1993: 190-201

[17] Lee H, Ng R, Shim K. Similarity join size estimation using locality sensitive hashing//Proceedings of the Very Large Databases Endowment, Seattle, USA, 2011: 338-349

[18] Lee H, Ng R, Shim K. Power-law based estimation of set similarity join size//Proceedings of the Very Large Databases Endowment. Lyon, France, 2009: 658-669

[19] Indyk P, Motwani R. Approximate nearest neighbors: Towards removing the curse of dimensionality//Proceedings of the 30th Annual ACM Symposium on Theory of Computing. New York, USA, 1998: 604-613

[20] Charikar Moses. Similarity estimation techniques from rounding algorithms//Proceedings of the 34th Annual ACM Symposium on Theory of Computing. Québec, Canada, 2002: 380-388

[21] Goemans M, Williamson D. Improved approximation algorithms for maximum cut and satisfiability problems using semi definite programming. Journal of the Association for Computing Machinery, 1995, 42(6): 1115-1145

[22] Ravichandran D, Pantel P, Hovy E. Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering//Proceedings of the Association for Computational Linguistics. Ann Arbor, USA, 2005, 43(1): 622-629

[23] Schaeffer S. Graph clustering. Computer Science Review, 2007, 1(1): 27-64

[24] Clauset A, Newman M, Moore C. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826

[25] Clauset A, Newman M, Moore C. Finding community structure in very large networks. Physical Review E, 2004, 70(6): 066111-1-066111-6

**ZHANG Yan**, born in 1965, Ph. D. candidate, associate professor. His research interests include data quality and data management, etc.

**YANG Long**, born in 1987, M. S. candidate. His research interests include data quality and query optimization.

**WANG Hong-Zhi**, born in 1978, Ph. D., associate professor. His research interests include XML data management, data quality, etc.

**Background**

In recent years, with the growth of the information, dirty data such as erroneous, duplicate, uncertain or inconsistent exists in many database systems. Dirty data greatly reduces the quality of the data and brings serious losses to the enterprises and communities. Therefore, new techniques are in demand to process dirty data to reduce its harm.

Existing work on processing dirty data is mainly data cleaning and data repairing. However, both data cleaning and data repairing have some limitations. First of all, they cannot clean or repair the dirty data exhaustively. Secondly, excessive data cleaning may lead to the loss of information. The last but not least, these two operations are both time-consuming and will lead to the inefficiency of systems. Therefore, researchers propose techniques to perform queries on dirty data directly without data cleaning and obtain query results with clean degree from the dirty data. However, existing techniques are only suitable for some special queries. In order to manage dirty data effectively and efficiently, a uniform data model is in demand. The most widely used model is the probabilistic data model. This model represents uncertain data effectively, but cannot describe the affect of query operations on the quality of the results. More importantly, it will generate all possible world instances during query processing, which results in the exponential growth of data size and affect the efficiency of the system.

To overcome the drawbacks of current methods, we propose a novel model for dirty data, entity-based relational database model. This model avoids the exponential growth of data size in the probabilistic data model. We also define the traditional query operations for dirty data on the new data model, which support queries with the requirement of data quality. Without the ability of processing dirty data, traditional database implementation techniques are not suitable for this model. Therefore, new techniques are in demand. In this paper, we focus on the implementation of query and propose new similarity join size estimation for the entity-based relational database.

This work is supported in part by the National Basic Research Program (973 Program) of China under Grant No. 2012CB316200; National Natural Science Foundation of China under Grant Nos. 61003046, 61133002, 61033015, and 61111130189, Doctoral Fund of Ministry of Education of China under Grant No. 20102302120054 and the Fundamental Research Funds for the Central Universities (Grant No. HIT. NSRIF. 2013064).

Our group focuses on the research of database for more than 20 years. Many papers have been published in conferences and transactions, such as SIGMOD, VLDB, ICDE, KDD, INFOCOM, TKDE and VLDB Journal. Our papers have been cited by other researchers over 3000 times.