一种改进的相似重复记录检测算法

郭 文 龙
（福建江夏学院电子信息科学学院 福建 福州 350108）

摘 要 数据仓库中相似重复记录的清洗对于提高数据质量有着重要意义，传统的判重算法有 SNM 算法、MPN 算法及 KNN 算法等。针对 MPN 算法判重精度和时间效率不高等缺陷，提出一种改进的 MPN 算法。根据记录属性的重要性分别设定属性权值，将属性切分为原子，进一步计算原子的权值，通过判定属性相似度进而判定记录相似度，提高查准率和查全率。采用均分大数据集为若干数据子集，并行采用 MPN 算法进行判重，提高判重时间效率。理论和实验分析表明该方法提高了相似重复记录检测的准确率和时间效率。

关键词 相似重复记录 属性 检测 MPN 算法

中图分类号 TP311 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2014.01.079

AN IMPROVED DETECTION ALGORITHM FOR SIMILAR AND DUPLICATED RECORDS

Guo Wenlong
(*College of Electronics and Information Science, Fujian Jiangxia University, Fuzhou 350108, Fujian, China*)

Abstract To clean similar and duplicated records in data warehouse is significant in improving data quality. Traditional duplication discriminant algorithms are the SNM, MPN, KNN and so on. For the defects of MPN algorithm, such as low precision in duplication discriminant and low time efficiency, we propose an enhanced MPN algorithm. It sets attribute weights separately according to the importance of record attributes, and divides attributes into atoms, then it further calculates the weights of atomic. By judging attribute similarity it determines the similarity of records, thus improves the precision ratio and recall ratio. By dividing large data set into data subsets equally, it adopts MPN algorithm in parallel to determine the duplications, and improves the time efficiency of duplication discriminant. Theoretical and experimental analyses show that the method improves the accuracy rate and time efficiency of similar and duplicated records detection.

Keywords Similar and duplicated records Attribute Detect MPN algorithm

0 引 言

以往各单位组建的各类信息管理系统都是各自为政、分头建设的，对于信息的采集也是分别进行的。随着数据库技术和网络技术的发展，各级各类信息系统通过联网进行资源共享，这就必须对数据库进行集成和重构。然而在数据的集成过程中，数据库中出现了大量的重复记录和“脏”数据，如何对这些数据进行清洗，便成了一个重要问题。

所谓重复记录是指一个表中出现两条记录的属性值完全一样的记录。对于重复记录的清除处理方法实现起来较为简单，只要把重复记录直接删除，保留其一就可以了，实际上由于在数据库整合过程中，关系表中设置了主键，所以重复记录是不允许出现的。然而许多情况下，由于格式或拼写的差异导致关系表中出现两条或多条记录表达的是同一个信息，而显示的却是两条记录。如表 1 所示，由于对地址和出生日期的拼写格式不一样，导致两条记录存在。如果两条记录在某些字段上值相等或足够相似，则就称这两条记录互为相似重复记录。数据库中相似重复记录的检测和清洗是目前学术界上研究的热点。

表 1 相似重复记录示例表

序号	姓名	地址	出生日期
1	张三	福建江夏学院电子信息科学学院	1978 - 01 - 03
2	张三	江夏学院电科院	78/01/03

1 已有算法概述

相似重复记录的确定，最简单的检测方法是对所有记录进行遍历，该方法实现算法简单，但时间复杂度为 $O(n^2)$，如果数据量很庞大，则对时间的消耗是很可怕的。目前国内外有很多研究成果，常见的有排序邻居算法 SNM[1,2]、多趟近邻排序算法 MPN[1,2] 和最近邻居算法 KNN[3-5]。排序邻居算法主要有三个缺陷：一是对关键字的依赖较大，如果关键字选取不当，可能造成许多相似重复记录未能检测出；二是滑动窗口大小的选取难以控制，滑动窗口设定大了，势必增加比较的次数和时间，选取过小，又可能造成遗漏；三是许多比较是不必要的，数据库中的

收稿日期：2013 - 06 - 16。福建省教育厅科技项目（JA12335）。**郭文龙**，讲师，主研领域：数据清洗技术，数据库技术。

相似重复记录就整个数据库而言是少量的,大部分数据不是相似重复记录,所以也就没有必要对所有的记录执行该算法。针对 SNM 算法的缺陷,MPN 算法提出了改进。该算法的思想是执行多次的 SNM 算法,每次使用较小的滑动窗口和不同的排序关键字。在比较时采用传递原理进行判定,即记录 A 和记录 B 相似重复,记录 B 和记录 C 相似重复,那么记录 A 和记录 C 相似重复,MPN 算法同 SNM 算法相比时间效率明显提高了,算法相对灵活,能得到较好的重复记录集,并能解决部分纰漏的问题。KNN 算法的不足之处在于当数据量很大时,计算每个待分类数据到已知样本的距离即求该样本的 K 个最近邻居计算量庞大。

由于数据量庞大,采用传统的方法进行重判需要大量的空间,时间效率也不高。针对这一情况,本文在 MPN 算法的基础上,提出改进的 MPN 算法 IMPN(Improved Multi-Pass Sorted Neighborhood)。具体过程是先将数据按某属性或某些属性的组合排序,然后均分成若干组,每一组单独进行相似重复记录检测,然后按其他方案再次排序、分组,重复执行判重操作,直到数据结果集满足要求为止。为了进一步提高时空效率,判重过程中如发现两条记录相似重复,则将其合并,删除原表中的两条记录,插入合并的记录,然后用合并的记录继续执行判重操作。

2 相关定义

为了描述方便,制定如下定义:

定义 1 设数据记录有 n 个属性,根据属性的重要性,设置每个属性的权重为 W,第 i 个属性的权重为 W_i,$1 \leqslant i \leqslant n$,则有权重向量 $W = \{W_1, W_2, \cdots, W_n\}$, $\sum_{i=1}^{n} W_i = 1$。

定义 2 将属性值切分为若干个原子,原子是属性中不可再分的最小单位,如姓名的原子为字,出生日期的原子为出生年、出生月和出生日,地址的原子为每个行政级别的地址元素。设第 i 个属性的原子个数为 k,则原子的权值为 W_i/k。

定义 3 设数据记录集 $R = \{R_1, R_2, \cdots, R_L\}$,$SimR(R_i, R_j)$ 表示记录 R_j 相对记录 R_i 的相似度,$1 \leqslant i \leqslant L, 1 \leqslant j \leqslant L$;$SimA(R_{ip}, R_{jp})$ 表示记录 R_j 的第 p 个属性相对记录 R_i 的第 p 个属性的相似度,$1 \leqslant i \leqslant L, 1 \leqslant j \leqslant L, 0 \leqslant p \leqslant n$;则有:

$$SimA(R_{ip}, R_{jp}) = \frac{\sum_{t=1}^{k} Sim(A_t, R_{jp})}{k} \tag{1}$$

$$SimR(R_i, R_j) = \sum_{p=1}^{n} SimA(R_{ip}, R_{jp}) W_p \tag{2}$$

其中 $Sim(A_t, R_{jp})$ 表示 R_{jp} 中的原子 A_t 在 R_{ip} 中的匹配情况,$Sim(A_t, R_{jp}) \in \{0, 1\}$,$1 \leqslant t \leqslant k$。

定义 4 设置阈值 U,如果 $SimR(R_i, R_j) \geqslant U$,则说明记录 R_i 和 R_j 相似重复,记为 $< R_i, R_j >$。

定义 5 记录具有传递性,如果 $< R_i, R_j >$,$< R_j, R_k >$,则 $< R_i, R_k >$。

定义 6 记录的查准率(Precision Ratio)表示检测出来的正确的相似重复记录在检测出来的相似重复记录中所占的比例,记录的查全率(Recall Ratio)表示检测出来的正确的相似重复记录在数据库中的实际的重复相似记录中所占的比例。设 C 表示数据库中实际的相似重复记录,F 表示检测出来的正确的相似重复记录,T 表示检测出来的相似重复记录,则查准率 R_P 和查全率 R_R 计算公式如下:

$$R_P = F/T \tag{3}$$

$$R_R = F/C \tag{4}$$

3 IMPN 算法

通过对大量数据库的分析,发现记录集相似重复记录的主要情况有以下三种:1. 由于同音字造成的相似重复;2. 由于日期格式不一致造成的相似重复;3. 由于地址描述方法不一样或采用简称造成的相似重复。鉴于此,本文在检测相似重复记录过程中分别对以上三种情况进行处理,以便提高查准率和查全率。对于同音字造成的相似重复,可以设置一个同音字表,当两条记录相似重复时,查找同音字表再行清洗;日期格式不一致的问题,如可以设定标准日期格式为"YYYY-MM-DD",在检测之前统一转化为标准格式再进行匹配,而地址数据则采用分为若干个原子再进行匹配的方法进行检测。

算法的基本思想如下:

1) 将数据集记录的日期格式统一转化为标准格式;

2) 按某些属性或属性的若干个原子进行排序;

3) 将已排序的数据集均分成若干个数据子集;

4) 判重检测;

5) 按其他的属性或属性的若干个原子再次排序,重复 2)—4) 的步骤,直到结果满意为止。

根据以上分析,判重算法描述如下:

输入:已按某属性或属性的原子进行排序的记录子集 R,权重向量 W,闸值 U;

输出:相似重复的记录集

```
Input(R,W,U)
For(m=0,m<F,m++)                    //F 表示多趟检测的次数
{ For(s=0,s<L,s++) rep_flag=0;
                    //设置重复标志,初始状态为 0 表示不重复
  input(VL); //VL 表示将大数据集均分成小数据集的最大记录数
  for(i=0;i<VL;i++)
  for(j=i+1;j<VL;j++)
  for(p=0;p<n;p++)                  //n 表示记录中的属性个数
  SimR(Ri,Rj) += SimA(Rip,Rjp) * Wp;
                    //根据式子(1)和(2)计算两条记录的相似度
  If (SimR(Ri,Rj) > U) R1_rep_flag = R2_rep_flag = j;
                                    //设置相似重复标记
  Output(Ri,Rj);
R = Union(Ri,Rj);
Delete(Ri,Rj);
Insert(R);
VL = VL - 1;
          //当对相似重复记录进行合并后,记录子集的个数少 1
}
```

4 实验分析

实验环境 实验计算机配置:处理器 Intel(R) Core(TM) i3-3110M CPU @ 2.40 GHz 四核,3.30 GB 内存,500 GB 硬盘;操作系统:Windows XP;软件:SQL Server2000 + visual C++ 6.0。

实验数据 来自某区常住人口数据库,共有 76.3 万条

记录,31 个属性。实验分三次进行,分别随机提取包含 1 万、5 万和 20 万条记录的数据量,通过人工和软件相结合的方式将提取的三个数据集分别处理成包含 102、467 和 1 989 条相似重复记录的数据集。即定义 6 中的 C 在三个数据集中的数量分别为 102 条、467 条和 1 989 条,实验检测出来的相似重复记录和正确识别的相似重复记录则由人工方式进行统计。

评价标准　本文是在 MPN 算法的基础上,通过均分数据集,制定属性权重进行判重的,所以实验与 MPN 算法进行对比。评价标准围绕常用的查准率和查全率按照式(3) 和式(4)进行计算加以对比讨论。

图 1 和图 2 的 X 轴坐标表示实验的数据记录数量。对三次实验结果分别计算两种算法的查准率和查全率,图 1 的 Y 轴坐标显示两种算法的查准率,图 2 的 Y 轴坐标显示两种算法的查全率。

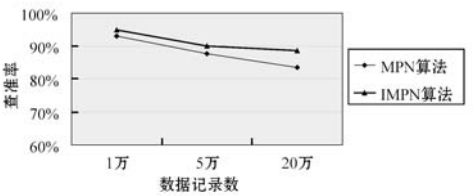

图 1　查准率比较

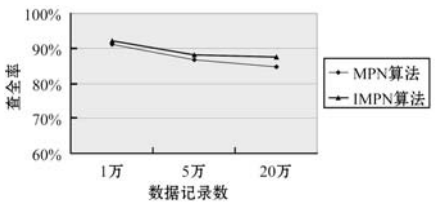

图 2　查全率比较

从图 1 和图 2 可以看出改进后的 IMPN 算法在查准率和查全率两方面均比 MPN 算法来得高,随着数据量的增大,IMPN 算法在查准率和查全率渐趋稳定且始终保持较高的水平。主要原因在于 IMPN 算法采用了计算字段权重且某些字段切分成原子分配权值再行匹配,提高了相似重复记录检测的准确度。

图 3 的 X 轴坐标表示数据记录数,Y 轴坐标表示两种算法的运行时间。由图 3 可以看出在时间效率方面 IMPN 算法比 MPN 算法提高了许多,且随着数据记录数的增多,IMPN 在时间上面的效率体现得愈发明显。

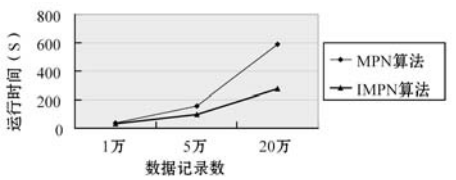

图 3　运行时间比较

算法的运行时间主要包括两个方面,一是排序操作,二是匹配操作。MPN 算法和改进的 IMPN 算法都需要用到多次排序操作,在排序方面所花费的时间基本一致。而在记录匹配方面,MPN 算法是直接在整个数据集上通过移动滑动窗口进行比较匹配的,而改进的 IMPN 算法将数据集均分成若干份,每个数据子集单独执行 MPN 算法,且均分的数据集是并行执行匹配操作的,显然时间效率提高了许多。

5　结　语

本文在 MPN 算法的基础上提出了 IMPN 算法,通过分配属性权值,切分属性为原子,计算属性的相似度,最后再计算记录的相似度。算法对大的数据集采用均分的方法切分为小数据子集,并对每个数据子集并行进行记录的相似重复检测,提高了时间效率。通过多次排序多次检测的方法提高了相似重复记录的查准率和查全率。和传统的 MPN 算法进行比较,实验证明取得了很好的效果。

下一步的工作是进一步研究大数据集的切分方法,进一步改进算法的某些环节,以进一步提高检测效率。

参　考　文　献

[1] Hernandez M A, Stolfo S J. Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem[J]. Data Mining and Knowledge Discovery,1998,2(1):9-37.

[2] Hernandez M, Stolfo S. The Merge/Purge Problem for Large Databases[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data, San Jose, California,1995:127-138.

[3] 陈振洲,李磊,姚正安. 基于 SVM 的特征加权 KNN 算法[J]. 中山大学学报:自然科学版,2005(1):17-20.

[4] 刘慧. 基于 KNN 的中文文本分类算法研究[D]. 西南交通大学,2010.

[5] 刘晓艳,王丽珍. 基于 KNN-Join 和 SNN 相似度的空间异常点检测算法[C]//中国计算机学会数据库专业委员会. 第二十五届中国数据库学术会议论文集(一). 中国计算机学会数据库专业委员会,2008:4.

[6] 李坚,郑宁. 对基于 MPN 数据清洗算法的改进[J]. 计算机应用与软件,2008(2):245-247.

[7] 王常武,韩菁华,张付志. 一种相似重复元数据记录检测方法[J]. 计算机工程,2009(21):85-87.

[8] 孟祥逢,鲁汉榕,郭玲. 基于遗传神经网络的相似重复记录检测方法[J]. 计算机工程与设计,2010(7):1550-1553.

[9] 肖满生,周浩慧,王宏. 基于模糊综合评判的相似重复记录识别方法[J]. 计算机工程,2010(13):51-53.

[10] 曹小峰. 基于相似重复记录检测的特征优选方法研究[J]. 计算机工程与设计,2009(23):5492-5495.

[11] Madeiro S S, Bastos-Filho C J A, Neto F B L, et al. Adaptative Clustering Particle Swarm Optimization[C]//Proceedings of the 23rd IEEE International Symposium on Parallel and Distributed Processing, Rome, Italy,2009:2257-2264.

[12] Aratsu T, Hirata K, Kuboyama T. Approximating Tree Edit Distance Through String Edit Distance for Binary Tree Codes[J]. Lecture Notes in Computer Science, 2009,54(4):93-104.

[13] 申德荣,刘丽楠,寇月,等. 一种面向 Deep Web 数据源的重复记录识别模型[J]. 电子学报,2010(2):275-281.

[14] Lwin T, Nyunt T T S. An Efficient Duplicate Detection System for XML Documents[C]//Proceedings of the 2nd International Conference on Computer Engineering and Applications, Bali Island, Indonesia, 2010:178-182.

[15] 刘伟,曹先彬. 对基于 MPN 的相似重复记录识别算法的改进[J]. 微计算机信息,2005(14):147-149.

[16] 张建中,方正,熊拥军,等. 对基于 SNM 数据清洗算法的优化[J]. 中南大学学报:自然科学版,2010(6):2240-2245.

[17] 刘彪. 空间数据库中基于 MapReduce 的 kNN 算法研究[D]. 大连海事大学,2012.