数据仓库中重复记录清理算法研究

钟嘉庆，张义芳，卢志刚
（燕山大学 电气工程学院，河北 秦皇岛 066004）

摘 要：针对重复记录清理中的“排序、识别、合并”算法存在的问题进行了改进。改进后的重复记录清理算法在保证记录匹配率的情况下有效地提高了记录排序的效率；在重复记录识别时，考虑了匹配字段的文字数量、在2个字段中出现的频率、在记录中各字段的重要性(权重)、中文字段的语义和语义重点偏后等5个因素；合并重复记录时采用了聚类和实用算法并用的策略，有效地提高了数据仓库中重复记录清理算法的准确性和健壮性。

关键词：数据清理；重复记录清理；重复记录识别；数据仓库
中图分类号：TP311.13 文献标识码：B

Research of data cleaning algorithm in data warehouse

ZHONG Jia Qing, ZHANG Yi Fang, LU Zhi Gang
(Electrical Engineering Institute, Yanshan University, Qinhuangdao 066004, China)

Abstract: This paper describes some advices for improving the problems in the “scheduling, detecting, merging” algorithm of duplicate elimination. The improved duplicate elimination algorithm has effectively promoted the efficiency of scheduling record on the environment that record matching rate was keeping high. In detecting duplicate records, it takes into account 4 factors. For instance, the number of characters, the frequency of character be found in the 2 fields, the importance (weight) of field in records ,the Chinese semantic and the semantic focus is always in the back location etc; In merging duplicate records, it uses both the cluster algorithm and practical algorithm to do that. It makes the data cleaning algorithm in data warehouse more accurate and healthier.

Key words: data cleaning; duplicate elimination; duplicate detecting; data warehouse

目前，国内外已经有一些对数据清理的研究，由于中文数据之间没有以空格分割，这在识别上带来了一定的难度，因此大部分的研究都只针对英文的数据清理，涉及中文数据清理的研究较少。重复数据清理技术旨在清除冗余的备份数据、确保只有“独有的”数据存储在磁盘上，即容量优化保护技术。重复数据清理技术的关键是只保留唯一的数据实例，有效地解决了“容量膨胀”的效率问题[1]。

从数据清理算法研究内容上讲，重复数据清除算法可分为两类：一类是数据清理的记录间算法，一类是数据清理的记录内算法。目前，研究人员对第一类算法研究得比较多，如：滑动窗口算法[2]、优先队列算法[3]等；对第二类算法的研究一般都是直接引用字符串相似匹配算法[4]，这种算法的缺点是没有考虑到字段不等长、中文字段语义重点偏后等重复记录的特点。

1 重复记录排序算法的改进

重复记录清理的直观方法是将每一个记录与数据库中其余记录逐个进行对比，该方法的识别精度非常高，但是在数据量较大的情况时，其处理时间会让用户难以忍受。邻近排序算法(SNM)[5]是目前常用的一种排序方法，它有效地克服了直观方法的缺点，大大提高了重复记录的匹配效率和重复记录清理的完成效率。但是，SNM算法存在其匹配结果严重依赖于排序关键字的选择和滑动窗口大小 W 的选取很难控制等缺陷。由于在

SNM算法里记录只能与窗口内的纪录进行比较，当W太小时或排序的关键字选择不当时，会造成漏配；而当W太大时又会产生很多没有必要的比较，因此恰当的W无论如何都无法得到。

本节针对SNM算法存在的上述缺陷作了改进，改进后算法的基本思想是使用相对较小的滑动窗口，选择数据库的一个关键字执行SNM算法，存储本次排序后相似记录的序号，然后依次选择数据库中的其他关键字独立地执行SNM算法，并在每次执行完毕后把此次执行结构中新增的相似记录号添加到相似记录存储表中得到所有可能重复记录的序号，然后对对这些可能的重复记录采用直观方法进行清理。

改进后的SNM算法的伪码描述如下：

```
while(还有没用过的关键字)
do{
为记录集TS中的记录选择该趟排序需要的排序关键字；
根据排序关键字对TS中的记录进行排序；
滑动窗口W从TS的第一个记录开始滑动；
while(W没有滑动到TS的尾部)
do{
初始化执行对比的次数n=0；
while(执行的对比次数n<|W|)
do{
新进入滑动窗口的记录与第n+1个进入窗口的记录进行重复记录比较；
if(比较的记录为相似重复记录)
{
把相似重复记录的记录号添加到相似记录存储表；
}
执行n+1；
}
向下滑动窗口；
}
对相似记录存储表中的记录采用直观方法进行比较，记录相似重复记录聚类；
}
```

2 重复记录识别算法的改进

记录排好序后，下一个要解决的问题是如何判断两条记录是否为相似重复记录。识别重复记录首先需要进行字段相似度的计算，然后再根据字段的权重进行加权和计算后得到记录的相似度，最后进行记录相似度和所设定阀值的比较，如果两条记录的相似度小于阀值，则认为这两条记录匹配，否则认为是两个不同的记录。基于相似度的重复记录识别算法[1]是最常用的一种重复记录识别方法，但是恰当阈值的设定仍是一个没有解决的难题。若阈值设定的过小，则容易遗漏某些相似的重复记录，从而降低了算法的匹配率；若阈值设定的过大，则容易将某些不同的记录误判为相似重复记录，从而降低了算法的正确率。此算法对记录的识别仅使用一个单一的阈值过于绝对，且没有考虑文中语句语义偏后的特点，无法满足实际情况的要求。

下面针对基于相似度的重复记录识别算法存在的上述问题对此算法进行了适当改进，给出了一种基于双阈值[6]位置权重[7]的语义重复记录识别算法。本算法的具体描述如下：对记录相似度设定一大一小两个阈值δ_{up}和δ_{low}，当通过位置权重识别法计算出当前两条记录的相似度大于δ_{up}，则直接判定它们是重复记录；若计算出的相似度小于δ_{low}，则可以判定它们是两个不同的记录；而对于相似度在δ_{up}和δ_{low}之间的两条记录，则不能直接确定它们是否重复或不重复，需要通过语义重复识别法[3][8]进行判定；对仍无法判定的记录则需人工进行处理。根据参考文献[9]一般阈值取0.37和0.68，为了提高准确率本文第一次相似度计算取阈值为0.35和0.7。

简单的字段识别法只考虑了字段之间的字符的匹配度，而忽略了匹配字符所在的位置(称为匹配序)。由于大部分中文尤其是特定领域的专业术语的语义重点往往集中在字段的后半部分字符串中，通过调整字段的匹配度和匹配序的权重(记作α和β，满足$\alpha+\beta=1$)，则可以在很大程度上提高字段识别的准确率。具体定义如下：

$$sim(f_1,f_2)=\alpha\times\frac{1}{2}\left(\frac{c}{m}+\frac{c}{n}\right)+\beta\times\min\left(\frac{m}{n},\frac{n}{m}\right)\times\frac{1}{2}\left[\frac{\sum_{i=1}^{t}L_1(i)}{\sum_{t=1}^{m}t}+\frac{\sum_{i=1}^{t}L_2(i)}{\sum_{k=1}^{n}k}\right] \quad (1)$$

其中，f_1和f_2分别为两个中文字段(如果字段中有英文字母，则将连续的英文字母视作一个汉字)，m和n分别为f_1和f_2的字数，c为f_1和f_2的识别字符数量，$L_1(i)$和$L_2(i)$分别为识别字符i在f_1和f_2中的匹配序。匹配序按照从左到右的顺序，从1开始自然数递增的方式计算，而α和β则一般根据黄金分割律来确定，分别取0.6和0.4[10]。例如，f_1=“燕大电气工程学院”、f_2=“燕山大学电气工程学院”，下面通过位置权重识别法判定S_1和S_2是否为重复字段。和的匹配字符为“燕”、“大”、“电”、“气”、“工”、“程”、“学”、“院”，它们在f_1中的匹配序依次为“1、2、3、4、5、6、7、8”，在f_2中的匹配序依次为“1、3、5、6、7、8、9、10”。那么f_1和f_2的语义相

似度为：

$$sim(f_1,f_2)=0.6\times\frac{1}{2}\times\left(\frac{8}{8}+\frac{8}{10}\right)+0.4\times\frac{1}{2}\times\min\left(\frac{10}{8},\frac{8}{10}\right)\times\left(\frac{1+2+3+4+5+6+7+8}{1+2+3+4+5+6+7+8}+\frac{1+3+5+6+7+8+9+10}{1+2+3+4+5+6+7+8+9+10}\right)=0.84 \quad (2)$$

基于语义距离的相似度识别方法体现了字段内部的结构和词语之间语义的相互作用关系,而编辑距离由于同义词词林的应用可以兼顾同义词之间的替换,并体现了组成句子的每个词深层的语义信息。基于语义距离的相似度识别算法的基本思路是：首先，利用参考文献[11]中介绍的骨架依存树思想分析字段的语法结构，得到字段中所有的核心词和直接依存于它们的有效词组成的搭配对(有效词定义为动词、名词和形容词，它是由分词后的词性标注决定的)，然后再进行语义距离(为两个字段有效搭配对的最短距离)的相似度计算，最后根据阈值进行重复识别判断。

设 f_1 和 f_2 为需要识别的两字段，f_1 包含的词为 f_{11}、f_{12}、…、f_{1m}，f_2 包含的词为 f_{21}、f_{22}、…、f_{2n}，则词 $f_{1i}(1\leqslant i\leqslant m)$和 $f_{2j}(1\leqslant j\leqslant n)$之间的相似度可用 $sim(f_1,f_2)$来表示，这样就得到两个字段中任意搭配对的相似度，f_1 和 f_2 两字段之间的语义相似度 $sim(f_1,f_2)$的计算公式如下：

$$sim(f_1,f_2)=\left(\frac{\sum_{i=1}^{m}a_i}{m}+\frac{\sum_{i=1}^{n}b_i}{n}\right)/2 \quad (3)$$

式中：

$$a_i=\min\left(sim(f_{1i},f_{21}),sim(f_{1i},f_{22}),\ldots,sim(f_{1i},f_{2m})\right)$$

$$b_i=\min\left(sim(f_{2i},f_{11}),sim(f_{2i},f_{12}),\ldots,sim(f_{2i},f_{1n})\right)$$

使用双阈值位置权重的语义识别法，虽然在一定程度上增加了用户的工作量，降低了算法的效率，但同时提高了算法的正确性和健壮性；而把位置权重和基于语义距离的相似度识别两种方法结合起来，扬长避短、互为补充，根据这些特征计算字段之间的相似度，可以使本重复识别算法获得很高的准确率。通过以上分析可知，本节对重复识别算法的改进是有效的、值得的。

3 重复记录合并算法的改进

在相似重复记录的识别完成以后，下一步要做的工作就是选择合适的方法合并识别出来的相似重复记录。参考文献[8]、[12]介绍了目前常用的多种重复记录合并方法，它们在合并方面各有利弊，单独使用都无法得到很好的效果，下面对此进行改进。

针对上述缺点，本节采用实用兼人工策略，给出了一种实用和聚类算法结合的合并算法。从一组相似重复记录中选择与其它记录匹配次数最多的一条记录进行保留，如果多个不同的记录具有相同的匹配率，则对这些相似记录进行聚类(会通过屏幕把聚类结果返回给用户)，由用户人工确定要保留的记录，并把其他重复记录从数据库中删除。

针对现有的重复记录清理策略存在的问题，分析了其原因，找出了现有重复记录清理策略里记录排序、重复记录识别、重复记录合并各步骤中所用算法存在的缺陷，给出了各自的改进方案，并通过算法分析和举例说明证明了改进的合理性。改进后的重复记录清理算法可以有效地提高判断质量，减小误判率，缩短了重复记录处理时间，很好地保障了数据仓库数据的质量。

参考文献

[1] LIN De Kang. An Information-theoretic Definition of Similarity[C]//Proc. Of the 15th International Conf. on Machine Learning. San Francisco,CA,USA:Morgan Kaufmann,1998.

[2] MONGE A. E, ELKAN C. An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records. DMKD,1997.

[3] GUTTMAN A. R-trees a dynamic index structure for spatial searching Proc. ACM SIGMOD Int Conf on Management of Data, 1984, 47-57.

[4] 冯玉才，桂浩，李华，等. 数据分析和清理中相关算法研究[J]. 小型微型计算机系统, 2005,26(6):1018-1022.

[5] HEMANDEZ, M A, STOLFO S J. The Merge/Purge Problem for Large Database[C].In SIGMOD Conference,1995:127-138.

[6] 洪圆，孙未未，施伯乐. 一种使用双阀值的数据仓库环境下重复记录消除算法[J]. 计算机工程与应用, 2005, 1:168-170.

[7] 张雪英，闾国年. 基于字面相似度的地理信息分类体系自动转换方法[J].遥感学报，2008，12(3):433-440.

[8] 刘宝艳,林鸿飞,赵晶.基于改进编辑距离和依存文法的汉语句子相似度计算[J].计算机应用与软件，2008，25(7):33-34.

[9] 陈伟. 数据清理关键技术及其软件平台的研究与应用[D]. 南京航空航天大学，2004.

[10]王源,吴小滨,涂从文,等.后控制规范的计算机处理[J].现代图书情报技术，1993，2:4-7.

[11]赵妍妍，秦兵，刘挺，等. 基于多特征融合的句子相似度计算[A]. 全国第八届计算语言学联合学术会议(JSCL-2005)论文集[C]，2006.

[12]DAVIDSON S B，KOSKY A S. Specifying Database Transformations in WOL[J].Data Engineering, 1999 ,22(1):25-31.

（收稿日期：2008-12-19）