文章编号:1674 - 2869(2017)04 - 0403 - 06

基于模糊综合评判和长度过滤的SNM改进算法

郭文龙,董建怀
福建江夏学院电子信息科学学院,福建 福州 350108

摘 要:为了提高数据库的数据质量,需要对相似重复记录进行清洗,基本邻近排序算法是目前常用的清洗算法之一.针对判重过程中属性权值计算主观性过强的问题,提出通过多用户综合评判确定属性权值的方法,该方法能更客观地评判属性的重要性程度.在此基础上,结合属性权值计算两条记录的长度比例,排除不可能构成相似重复的记录,减少了比较次数,提高了检测效率.实验结果表明改进算法在查全率、查准率及时间效率等方面均有所提高.

关键词:相似重复记录;模糊综合评判;属性;长度过滤;SNM;算法
中图分类号:TP311 **文献标识码**:A **doi**:10.3969/j.issn.1674-2869.2017.04.015

Improved SNM Algorithm Based on Fuzzy Comprehensive Evaluation and Length Filtering

GUO Wenlong, *DONG Jianhuai*
College of Electronics and Information Science, Fujian Jiangxia University, Fuzhou 350108, China

Abstract: To improve the quality of data, the approximately duplicated records need to be cleaned. The basic sorted-neighborhood method (SNM) is one of the commonly used cleaning algorithms. Aimed at the problem of excessive subjectivity of attribute weight calculation in detection algorithm, the article proposes a method based on the fuzzy comprehensive evaluation of multiuser to determine the attribute weight, which can be more objective to judge the importance level of the attribute. The proposed algorithm calculates the length ratio of the two records with attribute weight, then uses the length ratio to exclude records that are impossible to be approximately duplicated, reduces comparison times, and improves the detection efficiency. The experiment results show that the recall, precision and time efficiency are enhanced.

Keywords: approximately duplicated records; fuzzy comprehensive evaluation; attribute; length filtering; SNM; algorithm

当前,全球各行各业均组建了可以管理和推广自身业务的管理信息系统,如何有效管理和利用各类数据资源,是科学研究和决策支持的前提.随着信息化水平的不断提高,全球的各类数据库中存储的数据都呈现井喷式的增长.诸如银行、证券公司、通信公司等数据库的存储量均在百万以上,且可以预见随业务扩张的趋势将继续推动数据增长;而政府的人口基础数据库的数据量更是以亿计.在这些数据量庞大的数据库中存在着诸多重复数据,如何清理相似重复数据便成了亟需解决的问题.

目前常用的相似重复记录清洗算法是基本邻

收稿日期:2017-04-08
基金项目:福建省自然科学基金(2015J01653);福建江夏学院青年科研人才培育基金(JXZ2014011)
作者简介:郭文龙,硕士,副教授.E-mail:wlg1688@sina.com
引文格式:郭文龙,董建怀.基于模糊综合评判和长度过滤的SNM改进算法[J].武汉工程大学学报,2017,39(4):403-408.
GUO W L, DONG J H. The improved SNM algorithm based on the fuzzy comprehensive evaluation and length filtering[J]. Journal of Wuhan Institute of Technology, 2017, 39(4): 403-408.

近排序算法(basic sorted-neighborhood method, SNM)[1-3]. 该算法基于排序和合并的思路,通过关键字对数据记录进行排序,再在一定的窗口内比较临近的数据记录是否构成相似重复记录,如果构成重复记录则合并. 由于算法简单且易于实现,所以得到了大量的应用[4].

相似重复记录的识别通常由两个步骤实现,第一步是逐一对记录属性进行匹配,第二步综合考虑所有属性的相似度并计算两条记录的相似度,最后通过事先设定的阈值判定两条记录是否构成相似重复记录. 而在综合考虑所有属性的相似度并计算两条记录的相似度前必须为记录的所有属性设置权值,逐一利用属性权值乘以属性匹配度,方可客观地判断两条记录是否相似重复,所以属性权值的设置便成了判重的关键.

殷秀叶[5]提出了属性加权和同义属性的概念对相似重复记录进行判定. 文献[6]通过计算记录字段间的相似度,组成特征向量;然后采用改进的量子粒子群优化算法优化反向传播(back propagation,BP)神经网络进行学习,建立最优相似重复记录检测模型. 李鑫等提出了一种分组模糊聚类的特征优选方法,首先进行分组记录的属性处理,然后采用一种相似度比较计算方法进行组内相似重复记录的检测[7]. 周典瑞等采用综合加权法和基于字符串长度过滤法对数据集进行相似重复检测[8]. 然而这些方法的属性权值均是人为主观设置,未能体现不同属性的重要性程度. 针对属性权值确定主观性过强的问题,肖满生等提出基于数据分组和模糊综合评判的相似重复记录识别方法,该方法能较客观反映属性的重要性程度,取得较高的识别精度[9-10]. 文献[11]提出一种基于长度过滤的SNM改进算法,首先在窗口内根据两条记录的长度比例将构成相似重复记录可能性极小的数据排除在外,减少了记录比较的次数,提高了检测效率,为了降低因属性缺失等因素对判重的影响,进一步提出添加属性有效性因子并通过设定可变权值的方法取得了一定的效果. 刘雅思等提出动态容错法,解决了因属性缺失而误判的问题,提高了准确率[12]. 然而文献[11-12]中的权值仍然凭借单用户主观设定.

笔者结合文献[10-12]的思想提出了基于模糊综合评判和长度过滤的相似重复记录检测方法,主要思路为:采用模糊综合评判方法确定属性权值,提高判重精度;在检测相似重复记录前,采用长度过滤并充分考虑属性缺失的影响,排除不可能构成相似重复的记录,减少记录的比较次数,从而提高检测效率;基于SNM算法思想对数据记录集进行检测.

1 相关算法描述

1.1 模糊综合评判法

模糊综合评判法是基于模糊数学的评价法,也就是用模糊数学对现实世界中多种相互制约和关联的事物做出定量的评价[13]. 该方法可以将定性评价转为定量评价,可以较好解决非确定性的难以量化的评价问题[14].

在相似重复记录的属性权重计算中,可以通过多个用户分别对属性进行等级评价,然后取这些评价的平均值,通过模糊综合评判可以更客观的反映属性的重要性程度.

1.2 SNM算法

该算法由Hernandez M等人提出,其主要思路是先利用关键字或关键字的组合对所有数据进行排序,然后设定一个固定长度的窗口,在窗口内进行相似重复记录检测,随后滑动窗口,重复查重的过程[1-3,15].

SNM算法的主要过程如下:

1)排序. 根据属性的重要性程度,选择重要性程度高的一个关键字或若干个关键字的组合作为排序关键字,对数据记录集进行排序. 经排序后,潜在的相似重复记录将被调整到相邻或邻近的一个范围内.

2)归并. 设置固定长度的窗口,将第一条记录依次和窗口内的其他记录匹配,如果构成相似重复记录则合并处理. 当前窗口处理完毕,将窗口内第一条记录移除,然后将当前窗口内最后一条记录的下一条相邻记录移入窗口,循环执行匹配的过程.

由于该方法限定记录的匹配在窗口内进行,所以极大提高了判重效率.

1.3 长度过滤算法

针对大数据集,相似重复记录所占比例较小,如果在相似重复记录识别前首先将不可能构成相似重复的记录排除在外,显然可以提高检测效率. 长度过滤算法是根据两个字符串的长度比例找出每条记录的可能相似重复数据集范围的方法[11]. 然而,实际数据库中的记录常出现属性缺失或属性采用简写方式输入等情况,如在有些数据库中地址属性不是必填项,可能出现缺失也可能采用简写输入,而地址字符串长度在记录总字符串长

度中的占比较高，此时再单纯进行记录长度过滤显然存在偏差.

针对上述问题，结合模糊综合评判计算出的属性权重，本文提出改进的长度过滤方法. 设 C 表示待检测的数据记录集，n 表示数据量，R_i 表示第 i 条记录，R_j 表示第 j 条记录，则 $C=\{R_1,R_2,\cdots\cdots,R_n\}$. 设记录有 m 个属性，经模糊综合评判后计算出来的第 k 个属性权值为 $\boldsymbol{W}_k$ $(0\leqslant k\leqslant m)$. 设 $\mathrm{Len}(R_{ik})$ 表示第 i 条记录的第 k 个属性值长度，u 表示预先设置的长度比例阈值，则当两条记录的长度比例低于 u 时认为这两条记录不构成相似重复记录，即当 $\sum_{k=1}^{m}\boldsymbol{W}_k\cdot\mathrm{Len}(R_{ik})/\mathrm{Len}(R_{jk})<u$ 时，应将记录 R_i 排除在记录 R_j 的相似重复记录集外.

2　模糊综合评判法计算权值

记录属性反映实体的某个特征，但每个属性在实体中的重要程度不同，属性权值可以有效衡量属性的重要程度. 对于相同的属性集，每个用户由于主观性不同，如果单独设置属性权值必然会不一致. 在相似重复记录清洗中，采用单一用户设置的属性权值来判重显然欠合理，采用多用户模糊综合评判来设置属性权值可以很好地解决这个问题.

基于上述思想，记录属性权值的计算方法如下：

1）设数据集的记录属性有 m 个，按照其重要性将属性等级设置为 1~m，其中 1 表示该属性重要性最低、m 表示该属性重要性最高.

2）组织 s 个不同用户分别对记录集的属性进行等级评价，结果如表 1 所示.

表 1　等级评价

Tab. 1　Grade evaluation

用户 user	属性 1 attribute 1	属性 2 attribute 2	…	属性 m attribute m
用户 1	G_{11}	G_{12}	…	G_{1m}
用户 2	G_{21}	G_{22}	…	G_{2m}
…	…	…	…	…
用户 s	G_{s1}	G_{s2}	…	G_{sm}

3）计算属性等级：

$$G_j=(\sum_{j=1}^{s}G_{ij})/s,\quad 1\leqslant j\leqslant m.$$

其中，G_{ij} 表示第 i 个用户对第 j 个属性的等级评价.

4）计算属性权值：

$$W_i=G_i/(\sum_{j=1}^{m}G_j),\quad 1\leqslant i\leqslant m.$$

3　基于模糊综合评判和长度过滤的改进算法

基于SNM算法通过实验设定满足要求的窗口大小N，窗口内的记录利用记录属性长度过滤掉不可能构成相似重复的记录，减少数据记录的比较次数. 设 $\boldsymbol{W}\_C$ 表示窗口内待检测数据记录集，R_i 表示第 i 条记录，记录有 m 个属性，长度阈值为 u，权值向量为 $\boldsymbol{W}$，增设长度过滤标志数组 flag[N]，初始化状态 flag 数组元素全部置 0. 判重前窗口内第一条记录和窗口内其他记录进行长度过滤，如果窗口内某一记录 R_i 和窗口内第一条记录 R_1 长度比小于长度阈值 u，则 flag[i] 置 1. 在窗口内依据长度过滤非相似重复记录算法如下：

```
输入：W_C，N，W，u；
输出：flag[N]
for(i=0，i<N，i++)
    flag[i]=0；
for(i=1；i<N；i++)
{
   if( Σ_{k=1}^{m} Wk*Len(R_ik)/Len(R_1k) <u)
       flag[i]=1；
}
```

经过滤后，和窗口内第一条记录可能构成相似重复的数据集变小了，接下来则只要在窗口内可能构成相似重复的数据集中进行判重即可. 一个窗口判重结束，向下滑动窗口，循环执行这一过程.

基于如上分析，记录清洗算法描述如下：

输入：数据记录集 C，属性权重向量 $\boldsymbol{W}$，滑动窗口大小 N，记录相似度阈值 t，长度比例阈值 u

输出：相似重复记录集

```
Input(C，W， N，t，u)
for(i=0，i<L，i++)  //L表示数据集大小
{
     数据集格式化处理；
     数据清洗预处理；
}
```

根据模糊综合评判的属性权值，按权值高的字段或字段组合对数据集进行排序；

```
for(i=0，i<L-N+1，i++)  //根据设定的窗口大小，总共需滑动L-N+1次窗口
{
```

```
//窗口内检测相似重复记录
for(j=i+1,j<N+i,j++)
{
if( ∑_{k=1}^{m} Wk*Len(R_jk)/Len(R_ik) >=u)
    //满足条件则进行判重,否则过滤掉
{
根据属性权重向量进行判重处理;
Output(R_i,R_j);
}
}
}
```

4 实验部分

4.1 实验环境

实验硬件配置：intel(R)Core(TM)i3-3110M CPU@2.40 GHz,内存4.0 GB,硬盘空间500 GB.

软件环境：操作系统WIN7,数据库软件SQL Serve2005,算法在VC++6.0中编译运行.

4.2 评价方案

衡量清洗算法优劣标准的通常做法是计算查重的查全率和查准率,设C表示数据集中实际的重复记录,T表示检测出来的重复记录,F表示检测出来的错误的重复记录,则查全率R和查准率P的定义如下：

$$R=\frac{T-F}{C}. \tag{1}$$

$$P=\frac{T-F}{T}. \tag{2}$$

除了查全率和查准率外,运行速度也是算法优劣评判的指标之一.本文算法主要在文献[11]的基础上改进,故其性能通过在相同的数据集上分别对比文献[11]算法的查全率、查准率及运行时间来分析.

4.3 实验过程

实验数据来自某地人口基础数据库,共包含76.3万条记录和31个属性.实验中分别随机提取2万条、4万条及6万条数据量,基于人工和软件结合方法将数据集依次处理成包含407条、823条及1 478条的相似重复记录集,实验后检测出来的重复记录数及正确的相似重复记录数由人工方式统计.

在实验中共组织100个用户对记录属性进行等级评判,相似度阈值设为0.9,长度阈值设为0.8.在同样的数据集合上,分别利用文献[11]算法和本文算法进行判重,通过分别计算两种算法的查全率、查准率及运行时间并进行对比.

4.4 实验结果分析

在相同的数据集上分别对文献[11]算法及本文算法进行实验,为了描述方便,将文献[11]算法称为原算法,而将本文算法称为改进算法.根据以上所述方法,分别统计两种算法的查全率、查准率及运行时间,并整合绘制成图来表示,其结果如图1~图3所示.

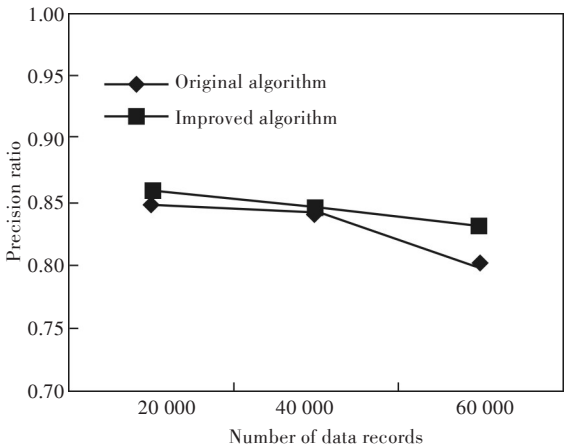

图1 查准率比较

Fig. 1 Comparison of precision ratio

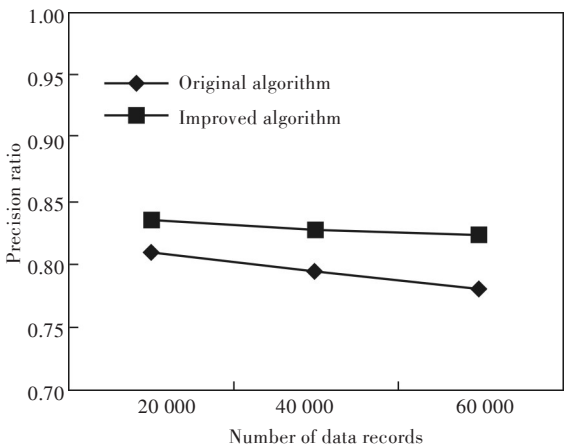

图2 查全率比较

Fig. 2 Comparison of recall ratio

从图1和图2中可以看出,改进算法不管是查准率还是查全率均比原算法有所提高.文献[11]首先根据单用户的主观意识设置属性权值,然后基于SNM算法在窗口内对数据记录进行长度过滤.而改进算法基于多用户对记录集的属性进行模糊综合评判,进而计算出属性权值,此方法必然能更客观地反映出记录属性的重要程度.此外,改进算法在长度过滤时再次利用到模糊综合评判法计算出来的属性权值,首先依次计算出两条记录每个属性的长度,然后分别利用两条记录各自的属性长度依次乘以前面计算出来的对应的属性权值,再把计算出来的两个结果进行相除得出一个值,

根据其值和事先设定的记录相似度阈值进行比较，如果超过阈值则表示两条记录重复，否则两条记录不构成相似重复.改进算法在长度过滤时充分考虑到属性值缺失的情况，如果记录的某个属性值缺失则该属性长度为0，这更能客观反应两条记录的相似重复情况.综上，改进算法的查全率及查准率必然有所提高.

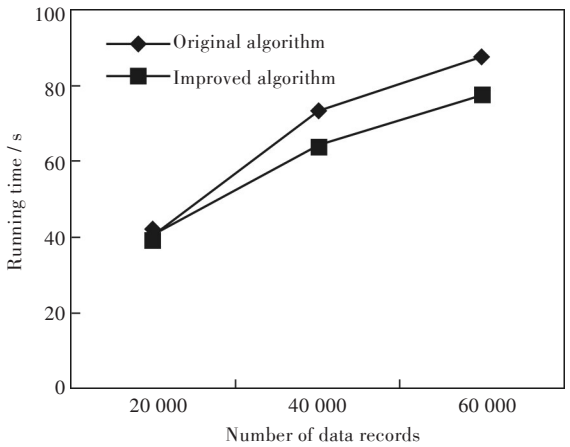

图3　运行时间比较

Fig. 3　Comparison of running time

从图3中可以看出，对于同样的数据量，改进算法在运行时间上均比原算法有所减少. 判重算法中均涉及到排序的操作，而排序操作所耗费的时间一致，两种算法均采用了长度过滤的方法，花费的时间也一致. 但是原算法采用添加属性有效性因子并在算法运行过程中根据实际情况调整属性权值，这必然耗费了时间，所以运行时间更长. 从图3中可知，随着数据量增大，两种算法的时间差越大，这说明改进算法的时间效率更高.

5　结　语

当前，数据呈现井喷式增长，各类数据库中所包含的相似重复记录不断增多，清洗相似重复记录也变得日趋重要. 基于相关文献，本文提出基于模糊综合评判计算属性权值，更客观地反映出属性重要性程度. 在此基础上，充分考虑属性缺失等情况，利用模糊综合评判的属性权值计算记录长度，将不可能构成相似重复的记录过滤掉，进一步减少判重过程中的匹配次数. 实验证明，改进算法一定程度提高了判重的精度和时间效率. 然而，在相似重复记录清洗过程中相似度阈值及长度阈值的设置问题仍是一个值得探讨的问题，两者设置过大或过小都将对查重精度产生影响，这将是今后应继续研究的问题.

参考文献：

[1] HERNANDEZ M, STOLFO S. The merge/purge problem for large databases [C]//Proceedings of the ACM SIGMOD international conference on management of data. California: San Jose, 1995: 127-138.

[2] HERNANDEZ M, STOLFO S. Real-world data is dirty: data cleansing and the merge/purge problem [J]. Data Mining and Knowledge Discovery, 1998, 2(1): 9-37.

[3] 叶焕倬，吴迪. 相似重复记录清理方法研究综述[J]. 现代图书情报技术，2010，26(9)：56-66.
YE H Z, WU D. A survey of approximately duplicate data cleaning method [J]. New Technology of Library and Information Service, 2010, 26(9): 56-66.

[4] 陈爽，宋金玉，刁兴春，等. 基于伸缩窗口和等级调整的SNM改进方法[J]. 计算机应用研究，2013，30(9)：2736-2739.
CHEN S, SONG J Y, DIAO X C, et al. Amelioration method of SNM based on flexible window and ranking adjusting [J]. Application Research of Computers, 2013, 30(9): 2736-2739.

[5] 殷秀叶. 大数据环境下的相似重复记录检测方法[J]. 武汉工程大学学报，2014，36(9)：66-69.
YIN X Y. Method for detecting approximately duplicate database records in big data environment[J]. Journal of Wuhan Institute of Technology, 2014, 36(9): 66-69.

[6] 陈芬. 改进量子粒子群算法优化神经网络的数据库重复记录检测[J]. 计算机应用与软件，2014，31(3)：20-21，115.
CHEN F. Database duplicate records detection using neural network optimized by iqpso [J]. Computer Applications and Software, 2014, 31(3): 20-21, 115.

[7] 李鑫，李军，丰继林，等. 面向相似重复记录检测的特征优选方法[J]. 传感器与微系统，2011，30(2)：37-40.
LI X, LI J, FENG J L, et al. An optimal feature selection method for approximately duplicate records detecting [J]. Transducer and Microsystem Technologies, 2011, 30(2): 37-40.

[8] 周典瑞，周莲英. 海量数据的相似重复记录检测算法[J]. 计算机应用，2013，33(8)：2208-2211.
ZHOU D R, ZHOU L Y. Algorithm for detecting approximate duplicate records in massive data [J]. Journal of Computer Applications, 2013, 33(8): 2208-2211.

[9] 周丽娟，肖满生. 基于数据分组匹配的相似重复记录检测[J]. 计算机工程，2010，36(12)：104-106.
ZHOU L J, XIAO M S. Detection of approximately duplicated records based on data grouping matching[J].

Computer Engineering, 2010,36(12):104-106.

[10] 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,2010,36(13):51-53.
XIAO M S, ZHOU H H, WANG H. Identification method of approximately duplicate records based on fuzzy integrated estimation[J]. Computer Engineering, 2010,36(13):51-53.

[11] 郭文龙. 基于长度过滤和有效权值的SNM改进算法[J]. 计算机工程与应用,2014,50(19):123-127.
GUO W L. Improved SNM algorithm based on length filtering and effective weights[J]. Computer Engineering and Applications,2014,50(19):123- 127.

[12] 刘雅思,程力,李晓. 基于长度过滤和动态容错的SNM改进算法[J]. 计算机应用研究,2017,34(1):147-150.
LIU Y S, CHENG L, LI X. Improved SNM algorithm based on length filtering and dynamic fault-tolerance [J]. Application Research of Computers, 2017, 34(1):147-150.

[13] 刘河香. 模糊数学理论及其应用[M]. 北京:科学出版社,2012.

[14] 张胜礼,李永明. 广义模糊集GFScom在模糊综合评判中的应用[J]. 计算机科学,2015,42(7):125-128,161.
ZHANG S L, LI Y M. Application of generalized fuzzy sets GFScom to fuzzy comprehensive evaluation [J]. Computer Science, 2015,42(7):125-128,161.

[15] 余肖生,胡孙枝. 基于SNM改进算法的相似重复记录消除[J]. 重庆理工大学学报(自然科学版),2016,30(4):91-96.
YU X S, HU S Z. Research on eliminating duplicate records based on SNM improved algorithm[J]. Journal of Chongqing University of Technology (Natural Science), 2016,30(4):91-96.

本文编辑:陈小平