对基于 MPN数据清洗算法的改进

李 坚 郑 宁
(杭州电子科技大学 浙江 杭州 310012)

摘 要 相似重复记录的清除是数据清洗领域中的一个很重要的方面,它的目的是清除冗余的数据。介绍了该问题的流行算法—多趟近邻排序算法 MPN(Multi-Pass Sorted Neighborhood),该算法能较好地对相似重复记录进行清除,但也有其不足:一是在识别中窗口大小固定,窗口的大小选取对结果影响很大。二是采用传递闭包,容易引起误识别。提出了基于 MPN算法的一种改进算法,试验结果证明改进算法在记忆率和准确率上优于 MPN算法。

关键词 数据清洗 相似重复记录 MPN

IMPROVMENT ON THE ALGORITHM OF DATA CLEANING BASED ON MPN

Li Jian Zheng Ning
(HangzhouDianziUniversity, Hangzhou 310012, Zhejiang, China)

Abstract Cleaning approximately duplicate records is an important task in data cleaning. MPN, a popular algorithm for this task, is introduced and its deficiencies are analyzed. Firstly, window is fixed in detecting approximately duplicate records. Secondly, transfer closure is used but it is easy to make errors. An improved algorithm of data cleaning based on MPN is introduced. The experimental results prove that this improved algorithm is better than MPN in the aspects of recall and precision.

Keywords Data cleaning Approximately duplicate records MPN

0 引 言

现在社会已经进入信息时代,正确的决策成为企业成败的关键。所以很多企业纷纷建立数据仓库,为进一步挖掘数据信息作准备,为企业决策提供信息。数据仓库的数据一般来自于多个相对独立的业务系统。由于录入错误、语义表示不一致、拼写错误等原因,数据仓库中的原始数据往往存在很多问题,这将直接影响决策的正确性。所以必须对原始数据进行清洗。相似重复记录的识别是数据清洗中的一个十分重要的方面。它的目的是识别并消除那些实际上映射同一个实体但在语义表示上存在差异的数据库记录。

本文介绍了目前用于相似重复记录识别的流行算法 SNM和 MPN,分析了它们的缺陷,提出了一种改进的用于相似重复记录识别的算法。该算法主要有两方面的改进,一个是变固定窗口为可变窗口,另一个是根据属性的重要性,给不同的属性不同的权值,根据总相似度决定相似重复记录。

1 已有算法分析

相似重复记录清洗算法主要有以下两个评价标准:

标准 1 记忆率 是识别出的相似重复记录占所有相似重复记录的百分比。

标准 2 准确率 是指在算法识别出的相似重复记录里,那些真正的相似重复记录所占的百分比。

识别相似重复记录最可靠的办法是嵌套循环法,具体的方法就是比较数据库中每对记录,但该算法时间复杂度太大,需要 $N(N-1)/2$次比较,其中 N是数据仓库中记录的总数。排序合并方法是清除数据库中相似重复记录的标准方法。它的基本思想是,先对数据集排序,然后比较相邻记录是否相等。这一方法也为在整个数据库级上检测重复记录提供了思路,目前已有的检测重复记录的方法也大多以此思想为基础。

基本近邻排序算法 SNM的主要思路是首先选出一个关键字,数据库中的记录根据关键字进行排序,使相似的记录尽可能排在一起,然后在排好序的数据集上滑动一个窗口,数据集中每条记录仅与窗口内的记录进行比较。如图 1所示。

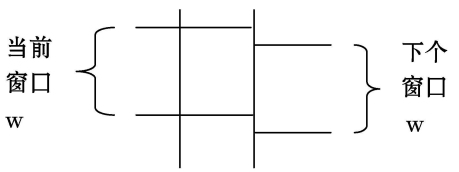

图 1 滑动窗口扫描数据集示意图

窗口的大小为 W条记录,则每条新进入窗口的记录都要与先前进入窗口的 W- 1条记录进行比较,以此来检测重复记录,然后最先进入窗口内的记录滑出窗口,最后一条记录的下一条记录移入窗口,再把此 W条记录作为下一轮比较对象,直到记录集的最后。该算法采用滑动窗口技术,减少了比较次数,也能有基本满意的记忆率和准确率。但是该算法有两个不足:(1)对关键字依赖程度太大,若选取的关键词本身错误严重或不能

收稿日期:2006-04-21。浙江省科技厅重点科研社会发展项目(2006C23060)。
李坚,硕士生,主研领域:数据清洗,数据仓库。

反映现实对象的特征，则真正的重复的记录不能聚集在一定范围内，失去匹配比较的机会。（2）窗口大小 w很难选取，窗口过大，则比较时间增加；窗口过小，则容易漏掉重复记录。

针对 SNM算法的对关键字依赖程度太大这个不足，Hernandez等人提出了多趟近邻排序算法 MPN，该算法的基本思想是独立地执行多趟 SNM算法，每趟创建不同的排序关键字和使用相对较小的滑动窗口。然后采用基于规则的知识库来生成一个等价原理，作为合并记录的判等标准，将每趟扫描识别出的重复记录合并为一组，在合并时假定记录的重复是具有传递性的，即计算其传递闭包。所谓传递闭包，是指若记录 R1与 R2互为重复记录，R2与 R3互为重复记录，则 R1与 R3互为重复记录。通过将每趟扫描识别出的重复记录计算传递闭包的方法，可以得到较完全的重复记录集合，能部分解决漏配问题。该算法也有一定的不足：闭包传递容易引起误识别，例如 R1和 R2相似，R2和 R3相似，但 R1和 R3可能已经差别很大了，但是闭包传递却认为 R1和 R2是重复记录。

本文改进算法主要有两个方面的改进。首先是针对 SNM算法中的窗口大小 w很难选取这个不足，采用可变窗口。固定窗口大小很难选取，较大时进行的比较次数多，而有些比较是没有必要的；w较小时可能漏配。采用可变窗口就是设两个窗口值 min和 max，一个阈值，窗口大小可在 min与 max之间变化，根据相似度与阈值的比较及时调整窗口值，减少遗漏，提高效率。

其次是针对 MPN算法中的不足，不采用闭包，而是首先根据数据库选出部分重要属性，并给属性以权值，权值乘以在该属性上的相似度得出相似度，最后不同相似度相加得出总相似度，根据总相似度与阈值的比较确定两条记录是否是重复记录。

2 改进算法

2.1 相关定义

定义 1 属性标记集 本文采用 n-gram方法，把属性值进行分割所构成的集合。如 college用 3-gram方法可以分割成 col，oll，lle，leg，ege，就是 happy的标记集。

定义 2 标记集的相似度 两个标记集 S1，S2的相似度 Sim(S1，S2)＝S1∩S2/S1∪S2。如 S1，S2完全相同，则 Sim(S1，S2)＝1。

定义 3 记录集 所有需要进行相似度识别的记录组成的集合。

定义 4 相似记录 如果两个记录的总相似度超过一定的阈值，称这两条记录互为相似记录。

定义 5 重要属性 根据具体的记录集，挑选出的对于识别相似重复记录具有正面效果的属性，一般要求属性值具有唯一性。

定义 6 有效权值 为了消除属性缺失对判断记录相似性的影响，提出了有效权值。即对记录的字段进行比较时，只有当两个字段都不为空时，比较才有效。否则，比较无效。

2.2 算法描述

本算法主要分为四步，如图 2所示。

2.2.1 属性选择

选择单个属性，对属性的依赖度太大，直接影响匹配的效率和精度。选择全部属性，一方面算法执行时间过长，另一方面有

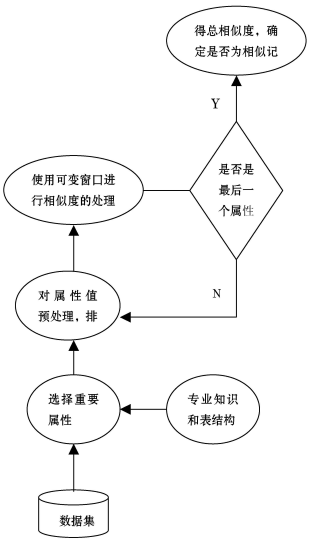

图 2 改进算法基本流程

些属性对匹配的精度产生负面影响。例如表 1中有性别属性，只有男女两个值，那么在性别属性上的匹配就将会出现许多重复记录。所以，应该选择部分重要属性，属性的选择要根据专业知识和表的结构，一般情况下，重要属性的属性值具有唯一性。

表 1 员工基本信息表

员工号	姓名	性别	所属公司
5696	涂利明	男	一公司
5698	宋兆辉	男	二公司
5636	付春捷	女	三公司
5601	赵冠军	男	一公司

2.2.2 属性值的预处理 主要为了解决相似重复记录排序后距离太远的问题。以空格和标点符号为分割符号，将属性值对应的字符串分割成单词，并对单词进行排序，然后再按字符顺序进行排序。如｛115 wenyi road，hangzhou｝预处理后就将为｛115 hangzhou road wenyi｝。

2.2.3 使用可变窗口进行相似度的处理 设置三个参数：W_{min}窗口最小值，W_{max}窗口最大值，LowThreshold最小阈值。窗口初始值 $W=W_{min}$，R_i与窗口内的记录进行匹配。匹配首先使用 3-gram方法分割字符串，计算其相似度 $S_{i,n}=Sim(R_i, R_n)$，用二维数组来存放相似度值。当匹配到 Rw_{min}时，如果 $Sim(R_i, Rw_{min})>$ LowThreshold，则窗口 W将扩大，R_i继续与下一条记录进行匹配。当相似度 $S<$ LowThreshold或窗口 $W>W_{max}$匹配中止。

2.2.4 得总相似度 记录中不同字段对反映记录特征的贡献是不平均的，根据属性的重要性，给予属性 C_i不同的权值 W_i。同时，字段缺失也是造成相似重复记录之间差异的重要原因。如以下两条记录：

1/付春捷/杭州文一路 115号/浙江大学/yezi@163.com/13956896325

2/付春捷/ /浙江大学/yez@163.com/

相对于记录 1，记录 2缺少了两个字段的值。如果按逐个字段比较、按权值累计匹配的方法，这两条记录的相似度将比较低，但事实是它们非常相似。为了消除字段缺失造成的负面影响，引入有效权值。对于进行比较的两条记录 1，2，假定参与比较的字段有 n个，记录 1，2的相似度计算公式为：

$$Tsim(1,2)=\frac{\sum_{i=1}^{n} Valid[i]*W_i*Sim(1,2)}{\sum_{i=1}^{n} Valid[i]*W_i}$$

其中 $W_1+W_2+\cdots+W_n=1$。

只有当两条记录在第 i个属性上对应的值都不为空时, 才进行字段比较, 此时 Valid[i]等于 1, 对应的权值为有效权值, 否则, Valid[i]为 0。设定相似度阈值 MatchThreshold, 如果 Tsim(R_1, R_2) >= MatchThreshold, 则 R_1 和 R_2 为相似重复记录。

3 实验结果

我们把改进后的算法与 MPN算法从两方面进行比较, 一方面是记忆率(R)和准确率(P), 另一方面是算法的运行效率。

我们在配置为 2 4GHz CPU, 256M内存, 安装了 Window 2000操作系统的计算机上, 选用 JAVA作为开发工具, 实现了改进后的算法。我们将其运用到某公交管理系统数据库中, 选取了日票收数据表进行实验。根据已有算法的阈值的取值范围一般为 0 37 ~ 0 68, 因此在实验中取 MatchThreshold= 0 65, LowThreshold= 0 35。在应用中, $W_{min}=10$, $W_{max}=100$。得出结果如下:

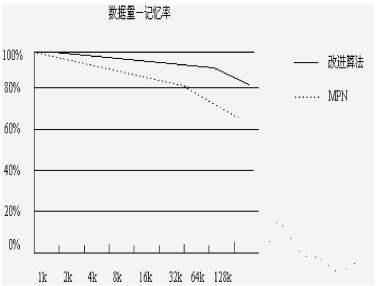

图 3　数据量与记忆率

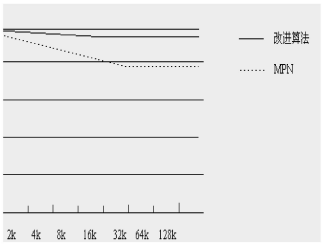

图 4　数据量与准确率

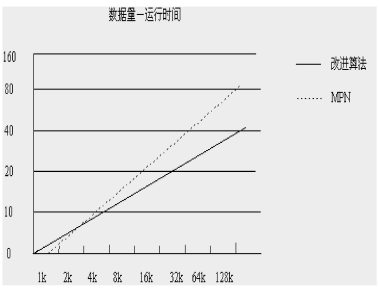

图 5　数据量与运行时间

从以上结果可以看出, 新算法在记忆率和准确率上都比 MPN算法高, 其中准确率几乎不受数据量的影响, 一直保持在 99%以上, 并取得了较好的识别效果。在运行时间上, 此 MPN稍有减少, 同时随着计算机硬件的更新, 这个也将不会是主要问题了。

4 结束语

文章分析当前数据清洗领域中识别相似重复记录的常用算法, 针对他们的不足, 提出了一种改进的算法, 实验表明该算法提高了重复记录判别的准确性, 减少了误识别。参数的确定可以根据具体领域知识, 让用户使用较小的训练数据集得到最合适的参数。

参 考 文 献

[1] Alvaro E Monge, Charles Elkan. An Effieient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records. DMKD, 1997.

[2] Rahm E, Do H H. Data Cleaning: Problems and Current Approaches. IEEE Bulletin on Data Engineering, 2000, 23(4): 3-13.

[3] Mong Li Lee, Wynne Hsu, Vijay Kothari. Cleaning the Spurious Links in Data in IEEE Intelligent Systems, Special issue on Data and Information Cleaning and Preprocessing, Volume 19, No. 2, March/April 2004.

[4] Zhu X, Wu X, Chen Q. Eliminating Class Noise in Large Datasets. Proc. 20 th int'l conf. Machine Learning. AAAI Press, 2003.

[5] Rohit Ananthakrishna, Surajit Chaudhuri, Venkatesh Canti. Eliminating Fuzzy Duplicates in DataWarehouses. VLDB, 2002.

[6] 洪圆, 孙未未, 施伯乐. 一种使用双阈值的数据仓库环境下重复记录消除算法. 2005, 1: 168-170.

[7] 佘春红, 许向阳. 关系数据库中近似重复记录的识别. 计算机应用研究, 2003, 9: 36-39.

(上接第 222页)

(a)原始图

(b)传统算法分割结果

(c)改进算法分割结果

图 2

比较图 1和图 2中(b)和(c)两幅图, 运用传统遗传算法分割结果不是很理想, 运用改进遗传算法搜索阈值, 不但更接近全局最优阈值并且收敛时间相对于传统遗传算法而言减小了, 取得较好的分割效果。

4 结 论

针对标准遗传算法用于寻找最优解时经常陷入局部寻优的问题, 本文提出用改进遗传算法和最大方差法相结合实现图像分割, 其优点是兼顾全局搜索性能和遗传算法的收敛速度。实验结果表明, 该方法用于图像全局最优搜索, 不但能搜索到最优阈值以有效地分割目标外, 还可以大大缩短寻找阈值的时间。

参 考 文 献

[1] 刘峰, 刘贵忠, 等. 遗传算法的 Markov链分析与收敛速度估计. 系统工程学报, 1998, 13(4): 79-85.

[2] 何琳, 王科俊, 等. 最优保留遗传算法及其收敛性分析[J]. 控制与决策, 2001, 15(1): 63-66.

[3] 徐璐. 改进遗传算法(IGA)及其在图像处理中的应用[D]. 2000, 3: 18-21.

[4] 刘文耀. 光电图像处理[M]. 电子工业出版社, 2002, 11: 207-208.