数据集成中不一致性数据相似性比较的加权算法

张艳秋　徐六通　王　柏
（北京邮电大学计算机科学与技术学院　北京100876）

A Weight Algorithm for Similarity Comparison of Inconsistency Data in Integrating Data

ZHANG Yan-Qiu　XU Liu-Tong　WANG Bo

Abstract　Reducing inconsistency is the key problem to improve data quality during data integration. In this paper, we first present a weighted algorithm of similarity coefficient which is superior to traditional algorithms if the source data have multiple characteristic items, all of which have to be taken into account, especially during the complex information integration. Secondly, we apply it to the experiment of telecommunication customers integrating, the results of data clustering show it has high feasibility and precision performance.

Keywords　Data integration, Similarity coefficient, Weight integration, Cluster

1　引言

数据集成是将不同存储平台的数据经过清洁转换等整合过程，以同一形式和内容呈现出来，它解决了多个应用系统或同一系统的不同数据存储共享数据的要求。集成中存在的一个主要困难是数据不一致，即同一数据在各个数据源的表现值不相同。这种不一致性严重影响了集成后目标库的数据质量。如果是数据仓库，还可能进一步影响基于数据仓库之上的OLAP和数据挖掘。因此，消除不一致数据是提高集成质量的关键。

数据不一致的根本原因是各数据源的呈现标准不同，主要可概括为[1]：数字大小写、使用缩写、词语省略、词序不同以及标点符号等方面。

目前，有关数据不一致性的理论研究较少，而且缺乏实践应用。通常的做法是利用相似性系数计算的方法求得不一致数据间的相似度，从而进行归并。但目前的相似性比较算法只适合对数据的单个特征项进行比较。如果信息多维化，即每条数据由多个特征项组成（如客户资料由姓名、地址等项组成），不一致数据归并的正确性不高。

本文试给出一种加权组合的改进方法。该方法在目前单特征项的相似系数计算基础上，增加了对距离值的考虑，并利用加权组合的策略针对各相似系数加权求和。最后利用电信的客户数据进行实验，表明改进后的方法具有更广的适用性和正确率。

2　相似性比较的算法

2.1　目前的相似系数计算方法

相似性数据的比较算法目前主要是：LD算法[2,3]，IDWP（Invariant distance form word position）算法[4]，JC（Jaccard's coefficient）算法[1]。

LD算法思想：任意字符串 x,y 的距离由 x,y 之间最小的简单操作数目决定，简单操作指插入、删除或替代一个字符。LD算法比较简单，但不适用于内容相同或相近，仅仅词序不同的情况，因为这样的字符串用LD算法会得到一个较大的距离值。

IDWP算法思想：将字符串分解成更小粒度的单位子串。在两个字符串的各个子串之间进行匹配，从而减小上述不合理LD的距离值。

LD，IDWP算法在求得距离值 d 后，用式(1)计算相似系数，其中 $\mathrm{Max}(x,y)$ 表示 x,y 中较大的字符串长度：

$$a=1-d*(1/\mathrm{Max}(x,y)) \tag{1}$$

JC算法思想：字符串间的相似程度用两个字符串共同含有的字符数与总字符数的比例表示。

2.2　对目前相似系数计算方法的改进

传统的相似数据比较算法（如LD和IDWP算法）在求得距离值后，除以其中较长字符串的长度。现在，我们对这种计算方法进行改进（其中，$f(d,x,y)$ 为关于距离值 d 以及 x,y 的函数）：

$$a=1-d*(1/f(d,x,y)) \tag{2}$$

改进后的方法在原有算法基础上增加了对距离值的考虑，这样可以提高相似系数计算的准确性。例如，如果两数据相差 n 位以上是同一数据的可能性几近于0，那么我们可以设定：

$$f(d,x,y)=\begin{cases}\mathrm{Max}(x,y) & d<n\\ d & d\geqslant n\end{cases} \tag{3}$$

此外，如果多维数据有 k 个特征项 $q_1,q_2,\cdots,q_k,q_1,q_2,\cdots q_k$ 对应的权值分别是 $w_1,w_2,\cdots,w_k(0\leqslant w_1,w_2,\cdots,w_k\leqslant 1;w_1+w_2+\cdots+w_k=1)$，那么我们进行相似系数的加权组合，即：$a=\sum_{i=1}^{k}a_iw_i$，$a_i$ 是利用上述改进算法得到的单个特征项的相似系数。

3　实验结果与分析

我们取国内某电信运营商各专业系统的客户资料作为实验样本，并利用一种Leader算法进行聚类：设定一个门限值 θ，依次计算该数据与各个已有聚类质心的不相似系数 β 的最小值（$\beta=1-a$），如果此最小值在门限 θ 范围内，则将其分配到最小值所对应的聚类中；如果最小值超出门限值 θ，则产生

（下转封四）

张艳秋　硕士。徐六通　副教授。王　柏　教授，博导。